Journal of Statistical Physics, Vol. 123, No. 5, June 2006 (© 2006)
DOI: 10.1007/s10955-006-9124-9

A Maximum Entropy Method for Particle Filtering

Gregory L. Eyink[1] and Sangil Kim[2]

Received February 21, 2005; accepted December 15, 2005
Published Online: June 20, 2006

Standard ensemble or particle filtering schemes do not properly represent states of low priori probability when the number of available samples is too small, as is often the case in practical applications. We introduce here a set of parametric resampling methods to solve this problem. Motivated by a general H-theorem for relative entropy, we construct parametric models for the filter distributions as maximum-entropy/minimum-information models consistent with moments of the particle ensemble. When the prior distributions are modeled as mixtures of Gaussians, our method naturally generalizes the ensemble Kalman filter to systems with highly non-Gaussian statistics. We apply the new particle filters presented here to two simple test cases: a one-dimensional diffusion process in a double-well potential and the three-dimensional chaotic dynamical system of Lorenz.

KEY WORDS: Bayesian estimation, filtering, particle methods, maximum-entropy, mixture models, ensemble Kalman filter

1. INTRODUCTION

In many application areas a Markov chain model is appropriate but the process is hidden and the only information available about it comes from a set of incomplete and imperfect measurements. This includes statistical signal processing, econometrics, and data assimilation in the geosciences. In abstract terms, a stochastic evolution equation produces successive transitions $\mathbf{x}_t \to \mathbf{x}_{t+1},\ t \in \mathbb{N}$ between elements of the state space $\mathcal{X}$. An observation equation gives the probabilities of the measured values $\mathbf{y}_t,\ t \in \mathbb{N}$ in the space of possible outcomes $\mathcal{Y}$. From the statistical point of view, the most detailed estimate of the state of the system up to time t is contained in the conditional probability density $P(\mathbf{x}_0, \ldots, \mathbf{x}_t \mid \mathbf{y}_1, \ldots, \mathbf{y}_t)$, given the observations up to that time. Such posterior probabilities can be obtained

[1] [1]Department of Applied Mathematics and Statistics, The Johns Hopkins University, Baltimore, MD 21218 U.S.A.; e-mail: eyink@ams.jhu.edu
[2]Department of Mathematics, University of Arizona, Tucson, AZ 85721, U.S.A.

from the prior distributions in the absence of observations by means of Bayes' theorem. Particularly important in many applications, e.g. meteorological weather forecasting, is the estimation of the *current state* of the system given the past observations, i.e. the probability density $P(\mathbf{x}_t \mid \mathbf{y}_1, \ldots, \mathbf{y}_t)$. To be useful, such an estimate must be obtained *sequentially* or recursively in time, as new measurements become available. Obtaining such probabilities in this way is known as the Bayesian filtering problem or optimal filtering problem.

Ensemble or *particle filtering methods* are a set of efficient and flexible Monte Carlo methods to solve the optimal filtering problem. These methods employ a large number N of random samples or "particles," advanced in time by the stochastic evolution equation, to approximate the probability densities. A resampling at measurement times both generates and destroys particles so as to representatively populate the regions of state space with high posterior probability. Such schemes were first proposed by Metropolis and Ulam,(44) but computing resources available at the time did not lend to their widespread use. However, following the seminal paper of Gordon, Salmon and Smith,(27) particle filtering methods have attracted strong general interest. Several recent books(12,36) and review articles(10) testify to their growing popularity and increasing range of applications. The methods possess several advantages that account for this surge of interest. First, they are straightforward to apply, since they require only the computation of a large number of solutions of the evolution equation of the problem. Second, they can easily be applied when the dynamics are nonlinear and the statistics highly non-Gaussian. Finally, the approximations that are yielded for the system statistics have been proved to converge to the optimal filter results in the limit as $N \to \infty$ and the convergence rate is independent of the dimension of the state space; see refs. 45, 47 or 10 for a review.

However, there are certain important application areas where the number of samples N that are practically available is very restricted, due to the high dimensionality of the state space $\mathcal{X}$. For example, in fields such as oceanography and climatology the computational cost of solving the evolution equation (a General Circulation Model, or GCM) is so high that at most $N = 10$–100 samples may be computed over time-intervals of interest. Nevertheless there is great interest in making proper estimates of the state of such systems given observational data.(25) With so few samples Monte Carlo error estimates are very large and the performance of ensemble/particle methods that are theoretically convergent may in practice be quite poor. A typical failure that occurs when the number of particles is small is that these methods neglect to sample states properly—e.g. states that are very improbable before measurements but very probable after them—because there are no particles present to represent them. In any problem where the number N of samples is so restricted it is clear that one must choose judiciously the members of the small ensemble, for instance, by using some prior knowledge about the statistics of the system.

One way to solve these difficulties with small sample-size is to use *parametric models* to represent the state probabilities. Events of small probability are always represented in such models and, if the model is carefully constructed, at realistic levels. An example of a particle filtering scheme that uses such an approach is the Ensemble Kalman Filter (EnKF), which was proposed by Evensen.(16,17) As in all Kalman filtering schemes, it implements Bayes' theorem using a Gaussian probability density to model the system statistics prior to measurements. It can be expected to work better than convergent particle filtering schemes in certain cases where the number N of samples is small, because the Gaussian model exhibits all states with a certain finite probability, even those far from the mean. This superior performance will be exhibited in some concrete examples presented below. On the other hand, this method does not yield the optimal estimates in the limit $N \to \infty$, unless the state variables are normally distributed. Thus, there is motivation to generalize this approach to better predict large-scale nonlinear systems with highly non-Gaussian statistics.

In this paper, we shall explore such particle filtering schemes, using non-Gaussian parametric models for the purpose of Bayesian updating and resampling. In particular, the simplest generalization of the Ensemble Kalman Filter will be considered, which models prior distributions by *mixture models* of weighted sums of Gaussians.(43) EnKF can be recovered in the special case of a single Gaussian component. Mixture models are a very natural device to accommodate multimodal and skewed distributions, with a modest additional cost in computation compared with EnKF. The use of Gaussian mixture models for optimal nonlinear filtering was proposed already some time ago(2) and such methods have been investigated for application to large-scale geophysical systems in refs. 2 and 5. However, these works used Gaussian mixture-models essentially as density-kernel estimators. Thus, for every ensemble-member (or for some specified fraction of ensemble members) a Gaussian density was determined, and the sum over all of these Gaussian kernels was then taken as the estimator for the probability density in state space. Our approach is rather different and is motivated instead by the notion of "multiple climate regimes," which is an old idea in geophysics (e.g. refs. 53 and 48 for the atmospheric circulation and ref. 62 for the oceanic thermohaline circulation.) The Gaussian components in our mixture models are supposed to represent such "climate regimes" and not individual ensemble members. Of course, this means that to apply our method one must develop appropriate Gaussian mixture models to represent the climate distribution. Fortunately, this problem has already been considered in the literature: see refs. 60 and 28 for a review of advanced statistical techniques that may be applied.

Another very important element of our approach is the use of a *maximum entropy* characterization to select the weights, means, and covariances of the Gaussian components of the mixture. In a certain sense, this scheme provides "minimal models" consistent with the information contained in the particle ensemble and is motivated by a rigorous H-theorem (Section 3.2). The maximum entropy characterization yields a practical optimization algorithm to determine parameters of the model density given moments of the ensemble. In recognition of the important role of this characterization in our proposed new filtering method, we shall refer to the method as the *Maximum Entropy Filter* (MEF). Furthermore, entropy plays other constructive roles in our approach. A maximum-entropy estimate of the post-measurement state provides a simplified "mean-field" approximation to the Bayesian update. This estimate is substantially cheaper to calculate than the full Bayesian estimate and may be a practical alternative when computational requirements for the latter exceed available resources. The entropy itself also serves as a useful measure of the information content of the observations and its rate of degradation over time.(1,9,29,37) The entropy is therefore a potentially very useful side-product of our choice of maximum-entropy distributions as parametric models. A preliminary discussion of this method applied to a particular model system has already appeared.(35)

A brief outline of the contents of the present paper is as follows: In Section 2 we discuss the filtering problem in a general state-space model and its recursive solution. In Section 3 we introduce our new entropy-based particle filters. First we discuss the construction of models for "background" or "reference" distributions in the absence of measurements, using Gaussian mixture models (3.1). Then we discuss the maximum-entropy estimation of probability densities with respect to a chosen reference density (3.2). On the basis of these results, we then elaborate our approach to particle filtering by resampling from maximum-entropy distributions (3.3). A simplified mean-field approach is also introduced to update distributions at measurements, based on a maximum-entropy criterion (3.4). In Section 4 we present results of numerical experiments with these methods applied to a diffusion process in a double-well potential (4.1) and to the chaotic 3-dimensional dynamical system of Lorenz (4.2). Our summary and conclusions are given in Section 5. Finally, in Appendices, we briefly review some standard ensemble/particle filters (Appendix A), and then we present important thermodynamical relations and functions for our maximum-entropy models (Appendix B), strategies for a efficient sampling from the models (Appendix C), and a comparison of the computational costs of the various particle filters considered (Appendix D).

2. THE FILTERING PROBLEM

In this section we shall describe the optimal filtering problem in more technical detail. We first discuss the general state space model set-up of the problem (see refs. 12, 36). Let $\mathbf{x}_t \in \mathcal{X}$ and $\mathbf{y}_t \in \mathcal{Y}$ for $t \in \mathbb{N}$ be two vector-valued stochastic processes, usually called the *signal process* and the *observation process*, respectively. In most of the applications of interest, the states spaces of these processes may be taken to be $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y} = \mathbb{R}^q$, and it will here be assumed that $q < p$. The signal process is a Markov process with initial distribution $\mathcal{P}_0(d\mathbf{x})$ and transition probability $\mathcal{P}_{t+1|t}(d\mathbf{x} \mid \mathbf{x}')$. We shall usually assume that these probability measures have densities $P_0(\mathbf{x})$ and $P_{t+1|t}(\mathbf{x} \mid \mathbf{x}')$ with respect to Lebesgue measure (at least in a generalized sense). A good example to keep in mind is the solution $\mathbf{x}_t$ of the following type of stochastic map

$$\mathbf{x}_{t+1} = \mathbf{f}_t(\mathbf{x}_t, \boldsymbol{\nu}_t), \qquad t = 0, 1, \ldots, T \tag{1}$$

where $\boldsymbol{\nu}_t \in \mathbb{R}^r$ is a random noise vector with known distribution $\Pi_t(d\boldsymbol{\nu})$ and $\mathbf{f}_t : \mathbb{R}^p \times \mathbb{R}^r \to \mathbb{R}^p$. Thus,

$$P_{t+1|t}(\mathbf{x}|\mathbf{x}') = \int \Pi_t(d\boldsymbol{\nu})\delta^p(\mathbf{x} - \mathbf{f}_t(\mathbf{x}', \boldsymbol{\nu})). \tag{2}$$

A special case of (1) of great practical importance is that when the equation is deterministic, i.e. the distribution $\Pi_t(d\boldsymbol{\nu})$ is a delta-measure and $\boldsymbol{\nu}_t$ appears in (1) as a (non-random) parameter. As for the measurement process $\mathbf{y}_t$, it is assumed to be conditionally independent of the signal process and to have marginal distribution $\mathcal{G}_t(\mathbf{y}_t \in A|\mathbf{x}_t = \mathbf{x}) = \int_A d^q\mathbf{y}\ G_t(\mathbf{y}|\mathbf{x})$. A simple example is provided by the following measurement model

$$\mathbf{y}_t = \mathbf{h}_t(\mathbf{x}_t) + \boldsymbol{\epsilon}_t, \qquad t = 1, \ldots, T \tag{3}$$

where $\mathbf{h}_t : \mathbb{R}^p \to \mathbb{R}^q$ are measured functions of the state variable and $\boldsymbol{\epsilon}_t \in \mathbb{R}^q$ are random observation errors, mutually independent and independent of the signal process, with probability density $R_t(\boldsymbol{\epsilon})$. In that case,

$$G_t(\mathbf{y}|\mathbf{x}) = R_t(\mathbf{y} - \mathbf{h}_t(\mathbf{x})). \tag{4}$$

This framework includes the case that measurements are made only at a subset of times $\mathcal{T}_M = \{t_m,\ m = 1, \ldots, M\}$ (or at no times at all) by taking $\boldsymbol{\epsilon}_t$ at all other times $t \notin \mathcal{T}_M$ to be normal with variance tending to infinity.

The optimal filtering problem is to obtain the set of conditional probability densities $P(\mathbf{x}_t|\mathbf{y}_1, \ldots, \mathbf{y}_t)$. These may be obtained by a standard recursive application of Bayes' Theorem. The filter density $P(\mathbf{x}, t)$ obtained by this recursion is discontinuous in time at instants where observations are incorporated. To state the algorithm, we introduce the following notation for the filter densities before and after measurements:

$$P(\mathbf{x}, t^-) = P(\mathbf{x}_t = \mathbf{x}|\mathbf{y}_1, \ldots, \mathbf{y}_{t-1}), \quad P(\mathbf{x}, t^+) = P(\mathbf{x}_t = \mathbf{x}|\mathbf{y}_1, \ldots, \mathbf{y}_t).$$

We also use the convention that $P(\mathbf{x}, 0^+) = P_0(\mathbf{x})$. We shall refer to the lefthand limit $P(\mathbf{x}, t^-)$ as the "prior" or "forecast" density and to the righthand limit $P(\mathbf{x}, t^+)$ as the "posterior" or "analysis" density. Then the sequential filtering algorithm can be implemented through a two-step procedure:

(1) *Prediction:* Advancing the probability density between measurements by means of the forward Kolmogorov equation,

$$P(\mathbf{x}, t^-) = \int d^p\mathbf{x}'\ P_{t|t-1}(\mathbf{x}|\mathbf{x}')P(\mathbf{x}', t-1^+), \tag{5}$$

(2) *Updating:* Conditioning upon measurements by means of Bayes' rule,

$$P(\mathbf{x}, t^+) = \frac{1}{\mathcal{N}_t} G_t(\mathbf{y}_t|\mathbf{x})P(\mathbf{x}, t^-), \tag{6}$$

where $\mathcal{N}_t$ is a normalization factor and $t = 1, \ldots, T$. This simple recursive solution to the optimal filtering problem is the basis of most numerical techniques to approximate the filter densities.

We recall that a useful side-product of this standard filter algorithm is the *likelihood function* $G_{1:T}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T)$ of the observations, or the probability density for this sequence of observations to occur. The normalization factors in Bayes' rule (6) are just the conditional probability densities $\mathcal{N}_t = G_t(\mathbf{y}_t|\mathbf{y}_1, \ldots, \mathbf{y}_{t-1})$ which, taken together, yield

$$G_{1:T}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T) = \prod_{t=1}^{T} \mathcal{N}_t \tag{7}$$

This is called the *innovation form* of the likelihood function.(36) A standard application of this quantity is *maximum-likelihood estimation*.(13,50) For example, suppose that the dynamical Eq. (1) has $\mathbf{f}_t(\mathbf{x}, \boldsymbol{\nu}; \boldsymbol{\theta})$ depending upon a vector parameter $\boldsymbol{\theta}$ with values in some domain Θ. Then—given the observations $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T$—it is natural to estimate the parameter by the value $\boldsymbol{\theta}_*$ which maximizes the likelihood of those data:

$$\boldsymbol{\theta}_* = \underset{\boldsymbol{\theta}\in\Theta}{\text{argsup}}\ \ G_{1:T}(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T; \boldsymbol{\theta}). \tag{8}$$

Thus, the filtering algorithm automatically enables parameter estimation.

3. ENTROPY-BASED PARTICLE FILTERING SCHEMES

We shall now introduce our new ensemble/particle method for approximating filter probability densities. Our aim is to generalize the Ensemble Kalman Filter method (Appendix A.2) to achieve better performance when statistics are highly non-normal. The key idea of the Bayes update in our approach is to use the particle information prior to the measurement to determine a non-Gaussian parametric model of the distribution. Bayes' rule is then applied to this model, altering the probabilities of the various states. Finally, new samples are drawn from the model with the updated parameters to create a particle ensemble that is evolved forward to the next measurement time.

3.1. Mixture Models for the Background Density

The first step in our construction of an appropriate parametric model is to develop a representation for the distribution of the stochastic process in the absence of any measurements. We shall denote this distribution by $Q(\mathbf{x}, t)$ and refer to it as the "background" or "reference" distribution. Its importance is due to the fact that, at long times between measurements and for sufficiently mixing Markov processes, the filter distribution $P(\mathbf{x}, t)$ is expected to relax back to this background distribution. That is, the memory of information gained from observations is expected to fade between measurements. In general, the reference distribution $Q(\mathbf{x}, t)$ is just the evolution of the initial distribution $P_0(\mathbf{x})$ under the forward Kolmogorov equation. If the initial distribution is the invariant distribution $P_*(\mathbf{x})$, then the Markov signal process is stationary and $Q(\mathbf{x}, t) = P_*(\mathbf{x})$ for all times t. In geophysical applications this time-invariant background would be called the "climatology".

Within our general scheme, various approaches may be followed for modeling the reference distribution $Q(\mathbf{x}, t)$. We shall consider here only one possibility, the use of *mixture models*. In this approach, the model of the background is taken to be a weighted combination of a finite number of normal distributions. In other words, the model density is of the form

$$Q_M(\mathbf{x}, t) = \sum_{m=1}^{M} w_m(t) N(\mathbf{x}; \boldsymbol{\mu}_m(t), \mathbf{C}_m(t)) \tag{9}$$

where $N(\mathbf{x}; \boldsymbol{\mu}, \mathbf{C})$ is the multivariate normal density with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$. The positive integer M is called the mixture complexity. See ref. 43 for a comprehensive, current introduction to the literature. A mixture model density can be constructed to converge to the density of the true distribution Q by increasing the mixture complexity M. However, we shall assume here that M is some relatively small positive integer (as in EnKF, where $M = 1$.) Methodologies for consistent estimation of a mixing distribution are discussed in ref. 32 and its references, while schemes to estimate mixture complexity are discussed in more detail in refs. 43, 51, 52. From a geophysical point of view, the mixture components represent "modes"/"patterns"/"regimes" of climate. As discussed in the Introduction, the idea that there may be just a few such long-lived "regimes" through which the climate system cycles in some fashion is an old speculation in the field: see ref. 61 for a recent review. A number of advanced statistical techniques are available to construct mixture models of such multiple-regime climate systems. For example, refs. 28, 60 employ a probabilistic clustering method using cross-validated likelihood.[59] A less objective but more intuitive approach to determine the weights, means and covariances of the mixture components is to perform conditional averaging over a large ensemble of realizations of the dynamics (1), for given values of an "index" that are assumed to characterize the climate "regime". The fraction of time for which those index values occur then determines the weight of the associated component in the mixture model and the conditional mean and covariance determine the corresponding parameters of the component Gaussian. We employ this simple procedure later to construct mixture models for our numerical examples.

3.2. Maximum-Entropy Distributions

We now consider the problem of modeling the filter density $P(\mathbf{x}, t)$, given a model of the background distribution $Q(\mathbf{x}, t)$. Needless to say, the effect in $P(\mathbf{x}, t)$ of conditioning upon observations taken before time t will make it unequal to $Q(\mathbf{x}, t)$. However, at long times between measurements $P(\mathbf{x}, t)$ is expected to converge back toward $Q(\mathbf{x}, t)$. A measure of this is the *relative entropy* or*Kullback-Leibler distance*,[9,39] defined as

$$H(P(t)|Q(t)) = \int d\mathbf{x}\ P(\mathbf{x}, t) \ln \left(\frac{P(\mathbf{x}, t)}{Q(\mathbf{x}, t)} \right). \tag{10}$$

It is known that for an ergodic, Markov process this quantity is a Lyapunov function, that is, a nonnegative, convex function of $P(t)$ which is non-increasing in time and which vanishes only when $P(t) = Q(t)$; see refs. 55, 56 and 9, Section 2.9. When the process is non-deterministic—e.g. a non-degenerate diffusion—then the relative entropy is monotonically decreasing in time. Therefore, to represent the filter distribution we would like to choose a model such that this "distance" of $P(\mathbf{x}, t)$ from $Q(\mathbf{x}, t)$ is as small as possible, consistent with the results of earlier measurements. At the current time t, new measurements of a function $\mathbf{h}_t(\mathbf{x})$ will be taken. We recall that $P(\mathbf{x}, t^-)$ denotes the forecast density, or the filter distribution just before those measurements. The moments in that distribution of the measured variable,

$$\boldsymbol{\eta}_{t^-} = \langle \mathbf{h}_t \rangle_{t^-}, \quad \mathbf{H}_{t^-} = \left\langle \mathbf{h}_t \mathbf{h}_t^\top \right\rangle_{t^-}, \tag{11}$$

represent the *measurement forecast* at the time t, both the mean $\boldsymbol{\eta}_{t^-}$ and the covariance matrix $\mathbf{C}_{t^-}^H = \mathbf{H}_{t^-} - \boldsymbol{\eta}_{t^-} \boldsymbol{\eta}_{t^-}^\top$. Any reasonable model for $P(\mathbf{x}, t^-)$ should be consistent at least with these measurement forecasts. One could demand consistency with still further moment constraints, for example, the first and second moments $\boldsymbol{\mu}_{t^-} = \langle \mathbf{x} \rangle_{t^-}$, $\mathbf{M}_{t^-} = \langle \mathbf{x}\mathbf{x}^\top \rangle_{t^-}$ of the state vector $\mathbf{x}$ itself. These represent the state forecast, both its mean $\boldsymbol{\mu}_{t^-}$ and covariance $\mathbf{C}_{t^-} = \mathbf{M}_{t^-} - \boldsymbol{\mu}_{t^-} \boldsymbol{\mu}_{t^-}^\top$, and are also a very natural set of constraints. In principle, the maximum-entropy approximation could even be systematized by considering sequences of moment-constraints involving polynomials $x_{i_i}, x_{i_1}x_{i_2}, \ldots, x_{i_1}\ldots x_{i_n}$ of increasing degree n. In certain cases this sequence of maximum-entropy approximations to the probability density has been proved to converge to the true density as $n \to \infty$ (e.g. ref. 23). Convergence may hold more generally, but constructing the nth approximant in the sequence involves the determination of $O(p^n)$ parameters and this will be prohibitively difficult when $p \gg 1$. See Appendix D.

We therefore take as our model the *maximum-entropy* (or, equivalently, minimum-information) distribution consistent with the measurement forecast. More precisely, we model $P(\mathbf{x}, t^-)$ with the probability density which minimizes (10) with the moments (11) as constraints. Introducing as Lagrange multipliers a q-vector $\boldsymbol{\lambda}$ and a $q \times q$ symmetric matrix $\boldsymbol{\Lambda}$, one easily finds that the maximum-entropy density belongs to an exponential family[(15,26,64)]

$$P(\mathbf{x}, t; \boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \frac{\exp[\boldsymbol{\lambda}\boldsymbol{\cdot}\mathbf{h}_t(\mathbf{x}) + \frac{1}{2}\boldsymbol{\Lambda}\boldsymbol{:}\mathbf{h}_t(\mathbf{x})\mathbf{h}_t^\top(\mathbf{x})]}{Z_t(\boldsymbol{\lambda}, \boldsymbol{\Lambda})} Q(\mathbf{x}, t). \tag{12}$$

Note that $Z_t(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ is a normalization factor to ensure that (12) integrates to unity. One can use this factor to define the convex, cumulant-generating function $F_t(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \log Z_t(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$. Then the moments $(\boldsymbol{\eta}, \mathbf{H})$ in (11) are obtained by taking derivatives, as follows:

$$\begin{aligned} \eta_i &= \frac{\partial F_t}{\partial \lambda_i}(\boldsymbol{\lambda}, \boldsymbol{\Lambda}), \quad H_{ij} = \frac{\partial F_t}{\partial \Lambda_{ij}}(\boldsymbol{\lambda}, \boldsymbol{\Lambda}), \quad i \neq j, \\ \frac{1}{2}H_{ii} &= \frac{\partial F_t}{\partial \Lambda_{ii}}(\boldsymbol{\lambda}, \boldsymbol{\Lambda}), \quad i = j \end{aligned} \tag{13}$$

In turn, the parameters $(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ corresponding to given $(\boldsymbol{\eta}, \mathbf{H})$ are uniquely determined as the optimizers in the Legendre transform

$$H_t(\boldsymbol{\eta}, \mathbf{H}) = \sup_{\lambda, \Lambda} \left\{ \boldsymbol{\eta}\boldsymbol{\cdot}\boldsymbol{\lambda} + \frac{1}{2}\mathbf{H}\boldsymbol{:}\boldsymbol{\Lambda} - F_t(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) \right\} \tag{14}$$

which gives the relative entropy for the model density (12). See refs. 15, 26, 64.

Technical simplifications in the maximum-entropy formalism occur (just as in Kalman methods) when the model background density $Q_M(\mathbf{x}, t)$ is a mixture of Gaussians (9) and when the measured quantities $\mathbf{h}_t(\mathbf{x})$ in (3) are affine functions of $\mathbf{x}$, i.e. $\mathbf{h}_t(\mathbf{x}) = \boldsymbol{\mathcal{H}}_t\mathbf{x} + \mathbf{d}_t$ for each time $t = 1, \ldots, T$. In that case, the cumulant-generating function $F_t(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \log Z_t(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ can be calculated explicitly, as shown in Appendix B. Another advantage of the mixture model in (9) when the measurement function is affine is that the maximum-entropy densities (12) are also mixture models:

$$P_M(\mathbf{x}, t; \boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \sum_{m=1}^{M} w_m(t; \boldsymbol{\lambda}, \boldsymbol{\Lambda}) N(\mathbf{x}; \boldsymbol{\mu}_m(t; \boldsymbol{\lambda}, \boldsymbol{\Lambda}), \mathbf{C}_m(t; \boldsymbol{\Lambda})), \tag{15}$$

where $w_m(t; \boldsymbol{\lambda}, \boldsymbol{\Lambda})$, $\boldsymbol{\mu}_m(t; \boldsymbol{\lambda}, \boldsymbol{\Lambda})$ and $\mathbf{C}_m(t; \boldsymbol{\Lambda})$ are modified weights, means and covariance matrices, respectively, calculated explicitly in Appendix B. We also discuss in that appendix strategies to deal with measurement of nonlinear (i.e. non-affine) functions, by linearization or by extended state space approaches.

3.3. The Maximum Entropy Filter

We can now outline the basic steps in the Maximum Entropy Filter (MEF) method. Between measurements, the particles $\mathbf{x}^{(n)}(t),\ n = 1, \ldots, N$ evolve independently under (1), just as in the standard particle methods discussed in Appendix A. The main difference with those methods consists in how Bayes' theorem is applied at measurement times. We shall assume that the measurement error $\boldsymbol{\epsilon}_t$ in (3) is an $N(\mathbf{0}, \mathbf{R}_t)$ random q-vector, i.e. normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}_t$. This is the situation most frequently encountered in practice. There are then three main steps in the implementation of Bayes' theorem in the MEF method:

(i) *Matching:* The moments $\boldsymbol{\eta}_{t^-}$, $\mathbf{H}_{t^-}$ in (11) are determined by averaging over the N-particle forecast ensemble:

$$\boldsymbol{\eta}_{t^-} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{h}_t\left(\mathbf{x}_{t^-}^{(n)}\right), \quad \mathbf{H}_{t^-} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{h}_t\left(\mathbf{x}_{t^-}^{(n)}\right)\mathbf{h}_t^{\top}\left(\mathbf{x}_{t^-}^{(n)}\right), \qquad (16)$$

A maximum-entropy density (12) is matched to these forecast statistics, with fitting parameters $(\boldsymbol{\lambda}_{t^-}, \boldsymbol{\Lambda}_{t^-})$ determined from the optimization (14).

(ii) *Updating:* Bayes' theorem is now applied, which, for normal error statistics, yields another maximum-entropy distribution (12) with parameters $(\boldsymbol{\lambda}_{t^+}, \boldsymbol{\Lambda}_{t^+})$ given by

$$\boldsymbol{\lambda}_{t^+} = \boldsymbol{\lambda}_{t^-} + \mathbf{R}_t^{-1}\mathbf{y}_t, \quad \boldsymbol{\Lambda}_{t^+} = \boldsymbol{\Lambda}_{t^-} - \mathbf{R}_t^{-1}, \qquad (17)$$

if $\mathbf{y}_t$ is the outcome of the measurement at time t.

(iii) *Resampling:* A new N-sample ensemble $\mathbf{x}_{t^+}^{(n)}$, $n = 1, \ldots, N$ is created, by sampling from the model posterior $P(\mathbf{x}, t^+)$, the maximum-entropy distribution (12) with updated parameters $(\boldsymbol{\lambda}_{t^+}, \boldsymbol{\Lambda}_{t^+})$.

Let us say a few words on the practical implementation of these three steps.

To carry out the matching in step (i), one must solve the maximization problem (14). When the reference distribution is represented by a mixture model $Q_M(\mathbf{x}, t)$ and the measurement function is affine, then the domain dom(F_t) of the convex function in (14) has a non-empty complement, at points where the matrix $\boldsymbol{\Lambda}$ is too large, and the values of F_t rise to infinity approaching the boundary of the domain from the interior. Therefore, algorithms to carry out the convex optimization must ensure that iterates stay within the feasible region dom(F_t). As shown in Appendix B, inside the domain it is possible to calculate exactly the gradients of F_t in (13), which can be used in minimization by descent algorithms. In our experiments below, we shall use a conjugate gradient (CG) algorithm in the space of $\frac{q(q+3)}{2}$ variables $(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$. Note that the number of variables only depends on the dimension q of the measured vector and not on the dimension p of the state vector, so that computational cost is considerably reduced when $q \ll p$. We employ a feasible Armijo line-search in our CG steps, so that the iterates never go outside of $\mathrm{dom}(F_t)$.[(49)] The calculation of $F_t(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ and its gradients contains an efficient check of feasibility of the current trial vector $(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$, since model realizability can fail if and only if Cholesky factors employed in the calculation fail to exist. See Appendix B.

The update step (ii) is trivial to implement, requiring only matrix additions and multiplications (as long as the inverse measurement error covariance $\mathbf{R}_t^{-1}$ is known). This is another great technical advantage of using maximum-entropy models for the filter distributions.

The resampling step (iii) can also be carried out efficiently using the mixture representation (15). One must simply select among the M components with probabilities $w_m,\ m = 1, \ldots, M$ and then sample from the normal distribution $N(\boldsymbol{\mu}_m, \mathbf{C}_m)$ for the selected m. When the dimension of the state space is small enough, simple standard methods are available for constructing realizations of random vectors $\mathbf{x}$ from the distribution $N(\boldsymbol{\mu}_m, \mathbf{C}_m)$. For example, one may take

$$\mathbf{x} = \boldsymbol{\mu}_m + \mathbf{S}_m \cdot \boldsymbol{\xi} \tag{18}$$

where $\boldsymbol{\xi}$ is a normal random p-vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}$, and $\mathbf{S}_m$ is a matrix square root of the symmetric, positive-definite covariance matrix $\mathbf{C}_m$, satisfying $\mathbf{C}_m = \mathbf{S}_m\mathbf{S}_m^\top$. Computable examples of square roots include the lower-triangular Cholesky factor $\mathbf{L}_m$ and the square root obtained by spectral analysis as $\mathbf{Q}_m = \mathbf{O}_m\mathbf{D}_m^{1/2}$, where $\mathbf{D}_m = \mathrm{diag}(\gamma_m^1, \ldots, \gamma_m^p)$ is the diagonal matrix of eigenvalues of $\mathbf{C}_m$ and $\mathbf{O}_m = [\hat{\mathbf{e}}_m^1, \ldots, \hat{\mathbf{e}}_m^p]$ is the orthogonal matrix whose columns are the orthonormal set of eigenvectors. In that case,

$$\mathbf{x} = \boldsymbol{\mu}_m + \sum_{a=1}^{p} \xi_a \sqrt{\gamma_m^a}\ \hat{\mathbf{e}}_m^a, \tag{19}$$

where $\xi_a,\ a = 1, \ldots, p$ are i.i.d. normal random variables with mean 0 and variance 1. Note that the eigenvectors are just the modes of the "principal orthogonal decomposition" (POD) of the state space $\mathbb{R}^p$ or the "empirical orthogonal functions" (EOF's) corresponding to the covariance $\mathbf{C}_m$ and (19) is the Karhunen-Loève (K-L) representation of Gaussian random vector $\mathbf{x}$; see ref. 41. The methods outlined above are practical when the dimension p of the state space is not too large. For resampling methods in case $p \gg 1$, see Appendix C.

Note that if the model background distribution $Q_M(\mathbf{x}, t)$ consists of a single Gaussian component and if the model prior distribution $P_M(\mathbf{x}, t^-)$ is a maximum-entropy distribution constrained by the full state statistics of second order, $\boldsymbol{\mu}_{t^-} = \langle \mathbf{x} \rangle_{t^-}$, $\mathbf{M}_{t^-} = \langle \mathbf{x}\mathbf{x}^\top \rangle_{t^-}$, then the MEF method is equivalent to the Ensemble Kalman Filter (EnKF)(7,17,66) (also, Appendix A). Thus, our MEF method can be considered a natural generalization of EnKF to problems with highly non-Gaussian statistics. Note that the algorithm yields also a simple formula for the likelihood function, or rather, for the *log-likelihood* $L_{1:T}$ in the innovation form $L_{1:T} = \log G_{1:T} = \sum_{t=1}^{T} \log \mathcal{N}_t$. In fact, it is easy using (12) to calculate the normalization $\mathcal{N}_t$ as

$$\log \mathcal{N}_t = \Delta F_t - \frac{1}{2} \mathbf{y}_t^\top \mathbf{R}_t^{-1} \mathbf{y}_t - \frac{1}{2} \log[(2\pi)^q \mathrm{Det} \mathbf{R}_t], \tag{20}$$

where

$$\Delta F_t = F_t(\boldsymbol{\lambda}_{t^+}, \boldsymbol{\Lambda}_{t^+}) - F_t(\boldsymbol{\lambda}_{t^-}, \boldsymbol{\Lambda}_{t^-})$$

is the jump in the function value of F_t during the measurement at time t.

3.4. A Mean-Field Filter

In certain applications—e.g. numerical weather prediction—the dimension of the measured vector is itself very large, $q \gg 1$. In such cases, the MEF method as discussed above may not be practical. The optimization over $\frac{q(q+3)}{2}$ variables in the matching step (i) of MEF has a computational cost $O(Mq^3)$, from the calculation of the function F_t and its gradients, which grows rapidly with q (see Appendix D). Even EnKF, when implemented by the representer algorithm discussed in Appendix D, requires $O(q^3)$ multiplications in order to invert covariances $\mathbf{C}_{t^-}^H + \mathbf{R}_t$, and this will also not be practical when $p \gg q \gg 1$. An entire literature has grown up around the problem of reducing the cost of Kalman filtering for $q \gg 1$, e.g. by batching measurements(3,31,34) or by other means.(18) To help deal with such cases within our framework, we can formulate an alternative maximum-entropy procedure in which the optimization is over only q variables. In this approach, we still apply Bayes' rule, but in a more approximate manner, to averages over the N particles. Therefore, we call this alternative procedure the MEF method with a "mean-field update," or, more simply, the Mean-Field Filter (MFF). We have previously discussed this mean-field approach applied to hindcasting (smoothing) in ref. 21 and we refer the reader to that work for a more thorough discussion.

Algorithmically, the mean-field method is easily described. Both the matching step (i) and the update step (ii) are changed, as follows:

Matching: We now take as our model of $P(\mathbf{x}, t^-)$, the filter density before the measurement, a maximum-entropy distribution with only the first moments in Eq. (11), i.e. $\boldsymbol{\eta}_{t^-} = \langle \mathbf{h}_t \rangle_{t^-}$, as constraints. This density is a member of the exponential family

$$P(\mathbf{x}, t; \boldsymbol{\lambda}) = \frac{1}{Z_t(\boldsymbol{\lambda})} \exp[\boldsymbol{\lambda} \cdot \mathbf{h}_t(\mathbf{x})] \cdot Q(\mathbf{x}, t) \tag{21}$$

with $Z_t(\boldsymbol{\lambda})$ the normalization factor. The q-vector $\boldsymbol{\lambda}$ is a Lagrange multiplier whose value $\boldsymbol{\lambda}_{t-}$ is that yielding the supremum in

$$H_t(\boldsymbol{\eta}) = \sup_{\boldsymbol{\lambda}} \{\boldsymbol{\eta} \cdot \boldsymbol{\lambda} - F_t(\boldsymbol{\lambda})\} \tag{22}$$

for $\boldsymbol{\eta} = \boldsymbol{\eta}_{t-}$. Note that $F_t(\boldsymbol{\lambda}) = \log Z_t(\boldsymbol{\lambda})$, similar to the definition earlier.

Updating: The update of $\boldsymbol{\eta}_{t-}$ to $\boldsymbol{\eta}_{t+}$ is obtained from the optimization

$$\boldsymbol{\eta}_{t+} = \operatorname*{arginf}_{\boldsymbol{\eta}} \left\{ H_t(\boldsymbol{\eta}|\boldsymbol{\eta}_{t-}) + \frac{1}{2}[\boldsymbol{\eta} - \mathbf{y}_t]^\top \mathbf{R}_t^{-1} [\boldsymbol{\eta} - \mathbf{y}_t] \right\} \tag{23}$$

where

$$H_t(\boldsymbol{\eta}|\boldsymbol{\eta}_{t-}) = H_t(\boldsymbol{\eta}) - H_t(\boldsymbol{\eta}_{t-}) - (\boldsymbol{\eta} - \boldsymbol{\eta}_{t-}) \cdot \boldsymbol{\lambda}_{t-}. \tag{24}$$

The latter is a positive, convex function whose minimum value (zero) is obtained at the unique point $\boldsymbol{\eta} = \boldsymbol{\eta}_{t-}$. Thus, the update $\boldsymbol{\eta}_{t+}$ is a compromise between the minimizers $\boldsymbol{\eta}_{t-}$ and $\mathbf{y}_t$ of the first and second terms in (23).

Resampling: This step is essentially the same as before. Once the value $\boldsymbol{\lambda}_{t+}$ is determined corresponding to $\boldsymbol{\eta}_{t+}$, then the maximum-entropy distribution (21) for $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{t+}$ can be sampled using its representation by a mixture model [Eq. (15) with $\boldsymbol{\Lambda}$ set to zero].

The meaning of the new update procedure is best seen from the significance of the entropy function (24) in large deviations theory.(15) If N independent samples $\mathbf{x}_{t-}^{(n)}$, $n = 1, \ldots, N$ are drawn from the model distribution $P(\mathbf{x}, t; \boldsymbol{\lambda}_{t-})$, the large-deviations result is, roughly speaking, that

$$\text{Prob}\left\{ \frac{1}{N} \sum_{n=1}^{N} \mathbf{h}_t\left(\mathbf{x}_{t-}^{(n)}\right) \approx \boldsymbol{\eta} \right\} \sim \exp[-N \cdot H_t(\boldsymbol{\eta}|\boldsymbol{\eta}_{t-})] \tag{25}$$

as $N \to \infty$, with $H_t(\boldsymbol{\eta}|\boldsymbol{\eta}_{t-})$ as in (24). We can also take an i.i.d. set $\{\boldsymbol{\epsilon}_t^{(n)}, n = 1, \ldots, N\}$ of $N(\mathbf{0}, \mathbf{R}_t)$ random variables, representing observation errors, and define the ensemble of measured values $\mathbf{y}_t^{(n)} = \mathbf{h}_t(\mathbf{x}_{t-}^{(n)}) + \boldsymbol{\epsilon}_t^{(n)}$, $n = 1, \ldots, N$. Then a large deviations result holds also for the joint probability as $N \to \infty$

$$\text{Prob}\left\{ \frac{1}{N} \sum_{n=1}^{N} \mathbf{h}_t\left(\mathbf{x}_{t-}^{(n)}\right) \approx \boldsymbol{\eta}, \frac{1}{N} \sum_{n=1}^{N} \mathbf{y}_t^{(n)} \approx \mathbf{y} \right\} \sim \exp[-N \cdot H_t(\boldsymbol{\eta}, \mathbf{y}|\boldsymbol{\eta}_{t-})] \tag{26}$$

where the joint-entropy $H_t(\boldsymbol{\eta}, \mathbf{y}|\boldsymbol{\eta}_{t-})$ is the function in curly brackets in (23). It follows that the value $\boldsymbol{\eta}_{t+}$ defined in (23) is the most probable value of $\frac{1}{N} \sum_{n=1}^{N} \mathbf{h}_t(\mathbf{x}_{t-}^{(n)})$

for the ensemble conditioned upon $\frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_t^{(n)} = \mathbf{y}$, in the limit as $N \to \infty$. This is still an application of Bayes' rule, but with the above "mean-field condition" on the sum rather than the correct condition that $\mathbf{y}_t^{(n)} = \mathbf{y}$ for all $n = 1, \ldots, N$. There is expected to be not much difference between the mean-field condition and the exact condition when the N-sample average takes on the value $\mathbf{y}$ if and only if every term in the sum is approximately equal to the same value $\mathbf{y}$. For example, it may be that transition between two "regimes" of climate is by some universal pathway. In that case, whenever such a transition is observed, the time-sequence of states will be closely similar and every member of an ensemble of such histories will be nearly the same as the ensemble-mean. The mean-field approximation will then be very good (See ref. 21).

There is also a natural mean-field analogue of the log-likelihood. It follows directly from (26) by a steepest descent result (contraction principle) that

$$\text{Prob}\left\{\frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_t^{(n)} \approx \mathbf{y}\right\} \sim \exp[-N \cdot H_t^Y(\mathbf{y}|\boldsymbol{\eta}_{t^-})] \tag{27}$$

as $N \to \infty$, where

$$H_t^Y(\mathbf{y}|\boldsymbol{\eta}_{t^-}) = \min_{\boldsymbol{\eta}} H_t(\boldsymbol{\eta}, \mathbf{y}|\boldsymbol{\eta}_{t^-}) = \min_{\boldsymbol{\eta}}\left\{H_t(\boldsymbol{\eta}|\boldsymbol{\eta}_{t^-}) + \frac{1}{2}[\boldsymbol{\eta} - \mathbf{y}]^\top \mathbf{R}_t^{-1}[\boldsymbol{\eta} - \mathbf{y}]\right\}$$

which is the same minimization as in the mean-field update (23). From (27) it is reasonable to define

$$\ln \mathcal{N}_t = -H_t^Y(\mathbf{y}_t|\boldsymbol{\eta}_{t^-}) \tag{28}$$

as the mean-field analogue of the log-innovation. Notice that this quantity is always non-positive, is concave in $\mathbf{y}_t$, and $= 0$ if and only $\mathbf{y}_t = \boldsymbol{\eta}_{t-}$. If the dynamics is linear and all statistics are normal, then (28) becomes $\ln \mathcal{N}_t = -[\mathbf{y}_t - \boldsymbol{\eta}_{t-}]^\top(\mathbf{C}_{t-}^Y)^{-1}[\mathbf{y}_t - \boldsymbol{\eta}_{t-}]/2$ with $\mathbf{C}_{t-}^Y = \mathbf{C}_{t-}^H + \mathbf{R}_t$. This is the exact result up to constant terms independent of $\mathbf{y}_t$ (cf. Eq. (37) below). Because the large-deviations result (27) has only logarithmic accuracy, one should expect to miss such constant terms. This does not detract necessarily from the utility of (28) to make maximum-likelihood estimates of parameters for distinct sequences $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T$ of observations.

In a practical implementation of the MFF method, one can avoid the calculation of $\boldsymbol{\eta}_{t^+}$ in (23). Instead, one can calculate $\boldsymbol{\lambda}_{t^+}$, $H_t^Y(\mathbf{y}_t|\boldsymbol{\eta}_{t^-})$ directly by combining (22) and (23) into a single optimization:

$$\begin{aligned}\boldsymbol{\lambda}_{t^+} = \underset{\boldsymbol{\lambda}}{\text{arginf}}\ &\{\boldsymbol{\eta}_t(\boldsymbol{\lambda}) \cdot (\boldsymbol{\lambda} - \boldsymbol{\lambda}_{t^-}) - F_t(\boldsymbol{\lambda}) + F_t(\boldsymbol{\lambda}_{t-}) \\ &+ \frac{[\boldsymbol{\eta}_t(\boldsymbol{\lambda}) - \mathbf{y}_t]^\top \mathbf{R}_t^{-1}[\boldsymbol{\eta}_t(\boldsymbol{\lambda}) - \mathbf{y}_t]}{2}\right\}\end{aligned} \tag{29}$$

with also

$$H_t^Y(\mathbf{y}_t|\boldsymbol{\eta}_{t^-}) = \inf_{\boldsymbol{\lambda}} \Big\{ \boldsymbol{\eta}_t(\boldsymbol{\lambda}) \cdot (\boldsymbol{\lambda} - \boldsymbol{\lambda}_{t^-}) - F_t(\boldsymbol{\lambda}) + F_t(\boldsymbol{\lambda}_{t-}) + \frac{[\boldsymbol{\eta}_t(\boldsymbol{\lambda}) - \mathbf{y}_t]^\top \mathbf{R}_t^{-1} [\boldsymbol{\eta}_t(\boldsymbol{\lambda}) - \mathbf{y}_t]}{2} \Big\} \tag{30}$$

where $\boldsymbol{\eta}_t(\boldsymbol{\lambda}) = \frac{\partial F_t}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda})$. To carry out the optimization in (29) by a descent algorithm, one must be able to calculate $F_t(\boldsymbol{\lambda})$, $\frac{\partial F_t}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda})$, and $\frac{\partial^2 F_t}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}}(\boldsymbol{\lambda})$. In the case that the model $Q_M(\mathbf{x}, t)$ is a finite mixture, these results are given in Appendix B. Although it is necessary in the optimization to use the second-derivative matrix of F_t, which is a $q \times q$ matrix, notice that all that is really needed is the contribution to the $\boldsymbol{\lambda}$-gradient of the function inside the brackets in Eq. (29):

$$\frac{\partial^2 F_t}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}) \left\{ \boldsymbol{\lambda} - \boldsymbol{\lambda}_{t^-} + \mathbf{R}_t^{-1} [\boldsymbol{\eta}_t(\boldsymbol{\lambda}) - \mathbf{y}_t] \right\}. \tag{31}$$

Hence, a descent algorithm may be coded so that storage requirements are only $O(q)$ and not $O(q^2)$. It is important to take advantage of such memory-savings in order to make the algorithm practical when q is very large.

4. NUMERICAL EXPERIMENTS

In this section we shall test the previously discussed particle filtering schemes in application to two simple dynamic models with highly non-Gaussian statistics. The first model is a nonlinear stochastic diffusion process in a double-well potential and the second is the 3-variable chaotic dynamical system of Lorenz.[42] These low-dimensional models have been chosen as test cases so that optimal results from convergent filtering schemes are available for comparison with our approximate (suboptimal) filtering methods. One of these optimal schemes is a standard convergent particle method, which we call the Weight Resampling Filter (WRF), that is reviewed in Appendix A. We shall also compare the results of our new filters with a standard suboptimal method, the Ensemble Kalman Filter (EnKF), also reviewed in Appendix A.

4.1. Double-Well Diffusion

Our first experiments will be for a 1-variable diffusion process which is given as the solution of the (Ito) stochastic differential equation with $\kappa > 0$

$$dx = f(x)\,dt + \kappa\, dW(t), \tag{32}$$

where $W(t)$ is the Wiener process and $f(x) = 4x - 4x^3$. We call this the double-well (DW) diffusion model. The invariant measure of this random process has probability density $P_*(x) \propto \exp(-\frac{2U(x)}{\kappa^2})$ where the potential $U(x) = -2x^2 + x^4$.

This density is bimodal and, in particular, non-Gaussian. Thus, this model has two regimes of "climate", corresponding to the two modes of the stationary distribution. The time series of the process is characterized by random switches between the two wells of the potential with minima located at $x = \pm 1$. An important issue in estimating this process is whether a given method can succeed in tracking a succession of such transitions.

We shall perform so-called "identical twin" experiments on this system with artificial measurements of the state $x(t)$ taken at a discrete sampling interval ΔT from a single realization of the process, which represents "reality". Observational errors will be simulated by adding to each of the sampled values an independent random variable from a normal distribution with mean 0 and variance R. We shall then make an estimate of the process conditioned upon those measurements, using the various particle filtering methods. The algorithms that we discussed in the previous sections were for discrete stochastic maps with measurements taken at each time step. These apply to the above stochastic differential equation when it is discretized for numerical integration. We use the simple Euler-Maruyama scheme[38]

$$\begin{aligned} x(t_{k+1}) &= x(t_k) + f(x(t_k))\Delta t + \kappa N_k \sqrt{\Delta t}, \\ t_{k+1} &= t_k + \Delta t \end{aligned} \tag{33}$$

with $\Delta t = 0.01$, where N_k is a sequence of i.i.d. standard normal random variables. When t_k is an integer multiple of ΔT, then we take a measurement with variance R, and otherwise we take no measurement or, equivalently, a measurement with infinite variance. We shall test our various particle filtering schemes in the experiments below against a convergent optimal filtering scheme using a numerical discretization of the Fokker-Planck equation to evolve the system statistics. For more details, see ref. 21.

In order to apply the MEF method, we must construct a model for the background $Q(x, t)$. Here we shall assume that the initial condition x_0 is drawn from the invariant measure $P_*(x)$, so that the reference density is time-independent and $Q(x) = P_*(x)$. Although we know the invariant measure exactly for this simple model, in order to illustrate the MEF method we need to construct a model by a mixture of Gaussians. Because of the bimodality of the invariant measure, we use a mixture $Q_M(x)$ of complexity $M = 2$. The sign-variable sign(x) can serve as an "index" to distinguish between the two regimes of "climate" in state-space. In order to construct the weights, means, and variances of the mixture components, we thus compute a single realization for a long time and gather probabilities $w_\pm$ for the complementary events $\{\text{sign}(x) = \pm 1\}$, and means and variances $\mu_\pm, C_\pm$ conditioned on these two events. Then we take $w_\pm$ to be the weights, and $\mu_\pm, C_\pm$ the component parameters of our Gaussian mixture model. In practice, we symmetrize the numerical results so that $w_- = w_+ = 0.5$, $\mu_- = -\mu_+$, and

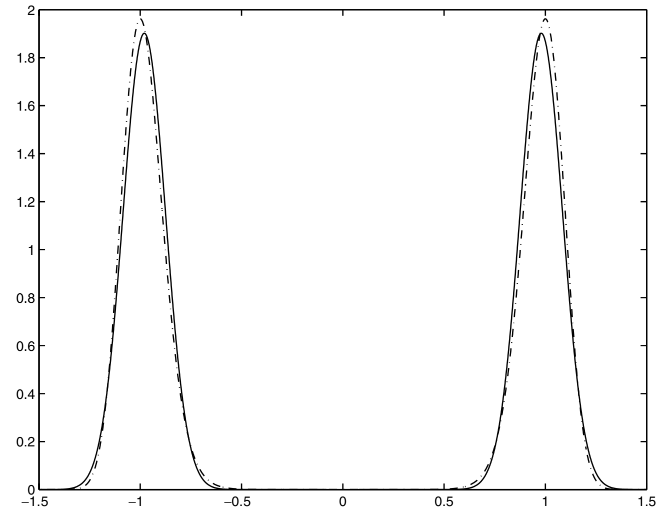

Fig. 1. Exact steady state density for DW model, $\kappa = 0.4$ (*dashed line*) and mixture model with $M = 2$ (*solid line*).

$C_- = C_+$. Our mixture model is then

$$Q_M(x) = w_- N(x; \mu_-, C_-) + w_+ N(x; \mu_+, C_+). \tag{34}$$

By construction, (34) has the same mean and variance of x as does the exact invariant measure. For the noise strength $\kappa = 0.4$ the densities of the invariant measure and the mixture model with $\mu_+ = 0.98, C_+ = 0.011$ are plotted in Fig. 1. Clearly, the mixture model (34) is here a quite good approximation.

4.1.1. Experiment A

Our first estimation experiment is for model (32) with this value, $\kappa = 0.4$. The "true" sample path was chosen to start in the positive well at $x = +1$. A realization starting in one well remains there an amount of time on average $\overline{\tau}$ which can be estimated from a weak-noise asymptotics, the Kramer formula:

$$\overline{\tau} \sim \frac{2\pi}{\sqrt{U''(1)|U''(0)|}} \exp\left(\frac{2\Delta U}{\kappa^2}\right), \quad \Delta U = U(0) - U(1), \tag{35}$$

valid as $\kappa \to 0$.[30,54] For our choice of parameters in this experiment, $\overline{\tau} \approx 3 \times 10^5$. On the other hand, when transitions occur, they require only about 5–6 time units to complete. Hence, the dynamics of this system consists of long periods of random diffusion about the bottom of a "well" interspersed with relatively rapid transitions, occupying a fraction of only about 10^{-4} of the total time. In our study we follow just the first such transition for a time-interval of 20 units around the point where the solution $x(t)$ passes through the unstable equilibrium at $x = 0$. On that interval we take seven measurements of the state $x(t)$ separated in time by $\Delta T = 2$ and contaminated with normal random errors of variance $R = 0.04$.

In this concrete setting, let us remind the reader of the specific steps that are taken in our new entropy-based filters.

We first consider MEF. At each of the seven measurement times t, we must choose a maximum-entropy distribution (12) to match the current particle ensemble $\{x_n(t^-) : n = 1, \ldots, N\}$ in the moments $\eta_{t^-} = \langle x(t^-)\rangle$, $H_{t^-} = \langle x^2(t^-)\rangle$ of the measured variable $h_t(x) = x$. The matching is accomplished by carrying out the minimization in (14). The "free-energy" function $F_t(\lambda, \Lambda)$ that appears there is now a function of just two real variables and this function and its derivatives are calculated from the formulas (70), (72)–(74) in Appendix B. Since the domain of the convex function F_t has a non-empty complement, we use a conjugate gradient scheme with a feasible Armijo line-search for the minimization in (14). This yields the parameters λ_{t^-}, Λ_{t^-} that are then updated to λ_{t^+}, Λ_{t^+} by Bayes' rule as in (17). The new maximum-entropy distribution with the updated parameters must lastly be resampled to yield the post-measurement ensemble $\{x_n(t^+) : n = 1, \ldots, N\}$. This is done by using the mixture representation (15) for $m = \pm 1$, with weights, means, and covariances given by (60), (64), (71), respectively, from Appendix B. These quantities are now all trivial to compute, since vectors are 1-dimensional and matrices 1×1. As discussed in Section 3.1, we can finally obtain the updated ensemble by choosing, for each $n = 1, \ldots, N$, one of the components $m = \pm 1$ of the mixture with probability $w_m(\lambda_{t^+}, \Lambda_{t^+})$—call it m_n—and then setting

$$x_n(t^+) = \mu_{m_n}(\lambda_{t^+}, \Lambda_{t^+}) + \sqrt{C_{m_n}(\Lambda_{t^+})}\xi_n, \qquad (36)$$

where ξ_n are i.i.d. $N(0, 1)$ random variables for $n = 1, \ldots, N$. Equation (36) is the analogue of (19) for our problem. This new set of samples is then evolved forward with the Eq. (32) to the next measurement time and the process repeated.

The procedure for MFF is similar and even somewhat simpler. A maximum-entropy distribution of the form (21) is chosen to match the pre-measurement ensemble $\{x_n(t^-) : n = 1, \ldots, N\}$ in just the first moment $\eta_{t^-} = \langle x(t^-)\rangle$. The matching is accomplished by carrying out the minimization over the single variable λ in (22), where the "free-energy" function $F_t(\lambda)$ and its first derivative are calculated from the formulas (79), (80) in Appendix B. We again use a conjugate-gradient algorithm for the minimization, yielding the parameter λ_{t^-}. However, unlike MEF, the update step to calculate λ_{t^+} is now carried out by a second minimization, as in (29). For this purpose the second derivative of F_t is also needed, in addition to the function and its first derivative, and this is given by (81) in Appendix B. Resampling the updated distribution is very similar as in MEF but is even more elementary, since the weights, means, and covariances of the two Gaussian components are given by the simpler formulas (76), (77), (78) in Appendix B.

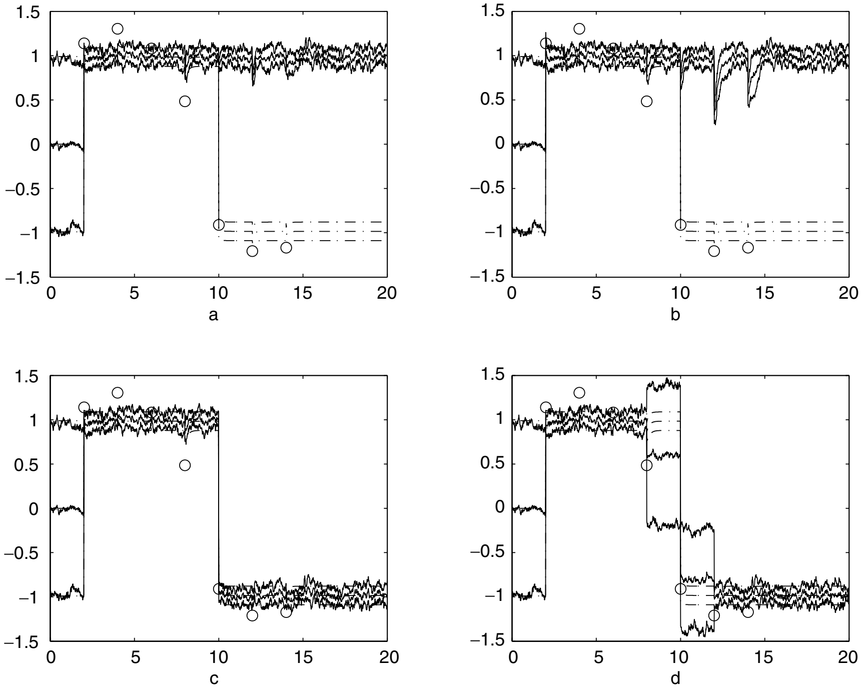

Fig. 2. Particle filter results for Experiment A with $N = 10$ samples: (a) WRF, (b) EnKF, (c) MEF, (d) MFF. The *circles* represent measurements taken from one sample path and the *solid lines* are the mean and mean ± standard deviations of the approximate filters. The *dot-dashed lines* are the mean and mean ± standard deviations from the Fokker-Planck solution of ref. 21.

In particular, the variances $C_m,\ m = \pm 1$ are not changed at all in the update. The formula (36) is used finally, just as in MEF, in order to generate the new ensemble $\{x_n(t^+) :\ n = 1, \ldots, N\}$ of analysis samples.

In Fig. 2 we show the results of applying the four particle methods, WRF, EnKF, MEF and MFF, to the DW model with such a set of measurements, using $N = 10$ particles. All the methods are initialized by sampling from the exact invariant measure using a Metropolis-Hastings algorithm. Therefore, all the methods show the same behavior, nearly zero mean and standard deviation close to one, before the first measurement. Up to the time of the transition at about $t = 10$, they continue to be very similar, except MFF, which shows a much larger variance than the others at times $t = 8$–12. After the transition, the methods all differ considerably. WRF and EnKF completely miss the transition and show almost no evidence of its existence. MEF and MFF capture the transition and estimate well its time and duration. MEF, in particular, is quite close to the optimal filter result, which is included in Fig. 2 for comparison. Interestingly, WRF performs the poorest of all the particle methods, despite its being the only one of the four which is convergent to the optimal result in the limit as $N \to \infty$. This exemplifies a general difficulty with WRF when the number of samples is small and a state ($x = -1$) is very improbable before a measurement, but very probable afterward. After several measurements at times t =2–8 indicating the state is near $x = +1$, all 10 particles are in that well. When the measurement comes at $t = 10$ indicating a transition, there is no particle in the well at $x = -1$ to carry the weight. EnKF fails for a related reason, because it models the system statistics by a Gaussian density with mean $\approx +1$ and small standard deviation ≈ 0.1 before the measurement. The Kalman gain is essentially a ratio between this standard deviation and the standard deviation of the measurement error, here 0.2. Therefore, there is insufficient gain at the observed transition to switch any of the 10 particles in EnKF to the other well at $x = -1$. MFF tracks the transition well, but is a little "premature" in suggesting a transition at time $t = 8$–10. On the other hand, the mean stays positive there and, despite the downward shift, the large standard deviation is consistent with the trajectory remaining near $x = +1$. At time $t = 10$, the mean becomes close to -1, faithfully reflecting the transition. As discussed in more detail in ref. 21, MFF in general follows observations too closely when the data lie near to the background average (here $x = 0$) and it tends to overpredict variances during transitions. MEF performs so well in this experiment that it would be hard to improve upon it.

In Fig. 3 we show the results of the four methods for the same estimation experiment using $N = 10^2$ particles. Except for reduced fluctuations in the ensemble-averages, there is little difference from the results for $N = 10$ in Fig. 2. This is to be expected. The calculation from formula (35) shows that only about 1 sample in 10^4 selected from the steady-state ensemble will be in transition between the two wells at any single time. Thus, $N = 100$ particles is not enough for WRF— after repeated measurements of the state in one well—to have any ensemble member making the transition to the other well. The situation with EnKF is even worse, because increasing N has almost no effect on the Kalman gain. The only useful consequence of increasing N is to generate a few samples sufficiently near to transition that the small gain from the measurements can shift them to the other well. Finally, we see that the results of MEF and MFF for $N = 10^2$ are also little changed from those for $N = 10$. Like EnKF, these methods perform very similarly for all values of N. However, unlike EnKF, MEF and MFF both succeed well in tracking the transitions.

In Fig. 4 we show the results of the four methods for the same estimation experiment using $N = 10^4$ particles. It now becomes clear that WRF is a convergent scheme, as it begins to approach closely the optimal filter results. With $N = 10^4$, there are enough particles either remaining in the well at $x = -1$ or switching back to that well in order for WRF to catch the transition at time $t = 10$. EnKF now indicates that there is a transition, but it lags the actual one by four time units. Because EnKF is "overconfident" that the state is near $x = +1$, two additional

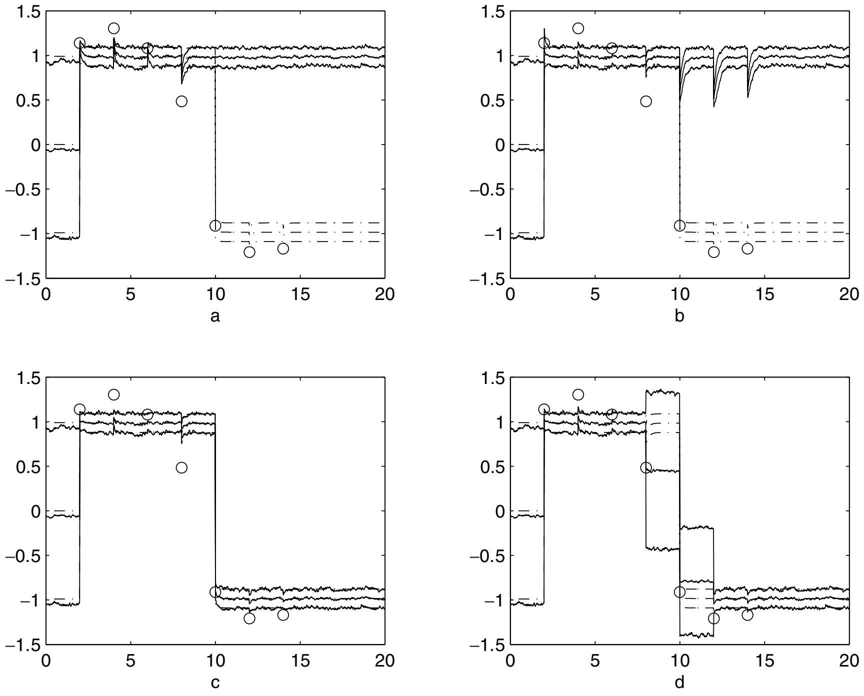

Fig. 3. Particle filter results for Experiment A with $N = 10^2$ samples: (a) WRF, (b) EnKF, (c) MEF, (d) MFF. Symbols as in Fig. 2.

measurements indicating that the state is in the other well are necessary to nudge some particles to make the transition. These results do not change much when N is further increased and seem to represent the limit for EnKF as $N \to \infty$. The results of both MEF and MFF for $N = 10^4$ are again little changed from those for $N = 10$ or 10^2, except that fluctuations are even smaller and the curves even smoother. It is one of the virtues of these methods that they very rapidly achieve their asymptotic $N \to \infty$ limit, already for relatively small N.

The conclusions that we have reached by examination of the plots can be confirmed quantitatively by considering the relative mean error, defined, for any quantity $\xi(t)$ over the time interval $0 < t < T$, as

$$\int_0^T dt\ |\xi_{ap}(t) - \xi_{ex}(t)| \Big/ \int_0^T dt\ |\xi_{ex}(t)|$$

where ξ_{ex} is exact and ξ_{ap} is approximate. We give these values in Table I for the mean of the state variable $x(t)$ over the interval $0 < t < 20$: For $N = 10$ and $N = 10^2$, WRF and EnKF results are poor, MFF reasonable, and MEF very good.

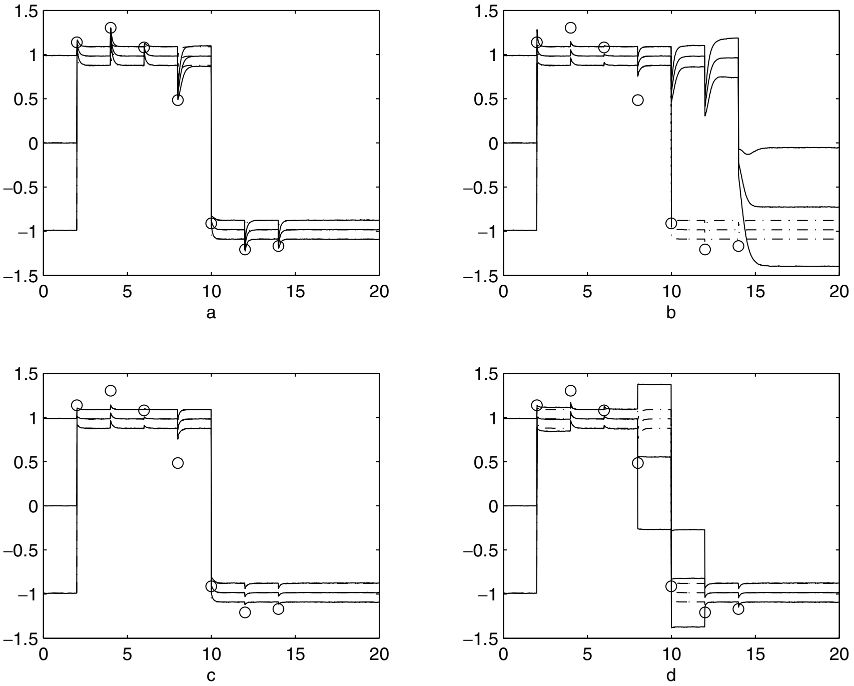

Fig. 4. Particle filter results for Experiment A with $N = 10^4$ samples: (a) WRF, (b) EnKF, (c) MEF, (d) MFF. Symbols as in Fig. 2.

For $N = 10^4$, MEF and MFF results are very similar to those for $N = 10^2$, while the results for the convergent scheme WRF are much improved (but not quite as good as those for MEF).

The likelihoods $G_{1:t}(\mathbf{y}_1, \ldots, \mathbf{y}_t)$ of the first t measured values are additional quantities that are approximated by the various filtering schemes, whose accuracy we would like to compare. Filtering methods supply the likelihoods in the

Table I. Relative errors in Experiment A

N	WRF	EnKF	MEF	MFF
(a) Filter Mean				
10^2	1.12297975	1.10542974	0.01507894	0.09369440
10^4	0.01067231	0.52647795	0.00187048	0.06703097
(b) Filter Standard Deviation				
10^2	0.03442152	0.03567663	0.03216601	0.57671912

innovation form $G_{1:t} = \prod_{s=1}^{t} \mathcal{N}_s$. Note that in WRF the innovation $\mathcal{N}_t$ is obtained from the normalization in (44). In EnKF the innovation may be consistently taken to be a Gaussian

$$\mathcal{N}_t = \frac{\exp\left[-\frac{1}{2}\left(\mathbf{y}_t - \boldsymbol{\mu}_{t-}^{Y}\right)^{\top}\left(\mathbf{C}_{t-}^{Y}\right)^{-1}\left(\mathbf{y}_t - \boldsymbol{\mu}_{t-}^{Y}\right)\right]}{\sqrt{(2\pi)^q \mathrm{Det}\left(\mathbf{C}_{t-}^{Y}\right)}}, \tag{37}$$

with $\boldsymbol{\mu}_{t-}^{Y} = \boldsymbol{\mu}_{t-}^{H}, \ \mathbf{C}_{t-}^{Y} = \mathbf{C}_{t-}^{H} + \mathbf{R}_t$. This is the standard result in the Kalman formalism for linear problems with normal statistics.[24] We have already discussed how to obtain the likelihoods—or, rather, their logarithms—in the entropy-based methods, MEF and MFF. In Figs. 5 and 6 we plot the log-likelihoods $L_{1:t} = \ln G_{1:t}$ of the four methods plotted against t, with a linear interpolation between measurement times. For comparison, we show the results obtained when the innovations $\mathcal{N}_t$ are calculated using the Fokker-Planck solution of ref. 21. Since the outcomes for $N = 10$ and $N = 10^2$ are very similar, we only show the latter. The results are quite consistent with those we saw in Figs. 2–4 for the means and standard deviations. We see in Fig. 5 that MEF is already very accurate for $N = 10^2$ and MFF reasonably good, but WRF and EnKF are much worse. Both MEF and MFF show a slight drop around $t \sim 10$, associated to the transition between wells. Because they miss the transition, WRF and EnKF show a continual, sharp decrease, indicating that —from the point of view of these approximations—the measurements in the other "wrong" well are very unlikely. For $N = 10^4$, the plots in Fig. 6 show that WRF and MEF now both give very good results, MFF still reasonably good and EnKF very poor. The results for MEF and MFF with $N = 10^4$ are both very close to those with $N = 10^2$. The underestimation of $L_{1:t}$ by MFF is consistent with its overestimation of the variance $\sigma^2(t)$, since the increased spread of the probability density implies lower values of the density and thus decreased likelihoods. We should caution that absolute values $L_{1:t}$ of the log-likelihood are of less interest than differences $\Delta L_{1:t}$ for the purpose of parameter estimation by a maximum likelihood criterion.

Finally, we shall plot in Figs. 7 and 8 the relative entropy $H(t)$ as a function of time for both the MEF and MFF methods, again with $N = 10^2$ and 10^4. For comparison, we show the exact relative entropy calculated by a discretization of the integral (10) using the Fokker-Planck solution from the scheme of ref. 21. Furthermore, we also calculate a relative entropy from EnKF using the formula for a pair of normal densities $P = N(\boldsymbol{\mu}, \mathbf{C})$, $Q = N(\boldsymbol{\mu}_*, \mathbf{C}_*)$ that

$$H(P|Q) = \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_*)^{\top}\mathbf{C}_*^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_*) + \frac{1}{2}\mathrm{Tr}[\mathbf{C}\mathbf{C}_*^{-1} - \mathbf{I}] - \frac{1}{2}\ln\left(\frac{\mathrm{Det}\mathbf{C}}{\mathrm{Det}\mathbf{C}_*}\right).$$

We take $\boldsymbol{\mu}$, $\mathbf{C}$ to be the mean and covariance from EnKF and $\boldsymbol{\mu}_*$, $\mathbf{C}_*$ to be the mean and covariance for the invariant measure, calculated from long-time averages.

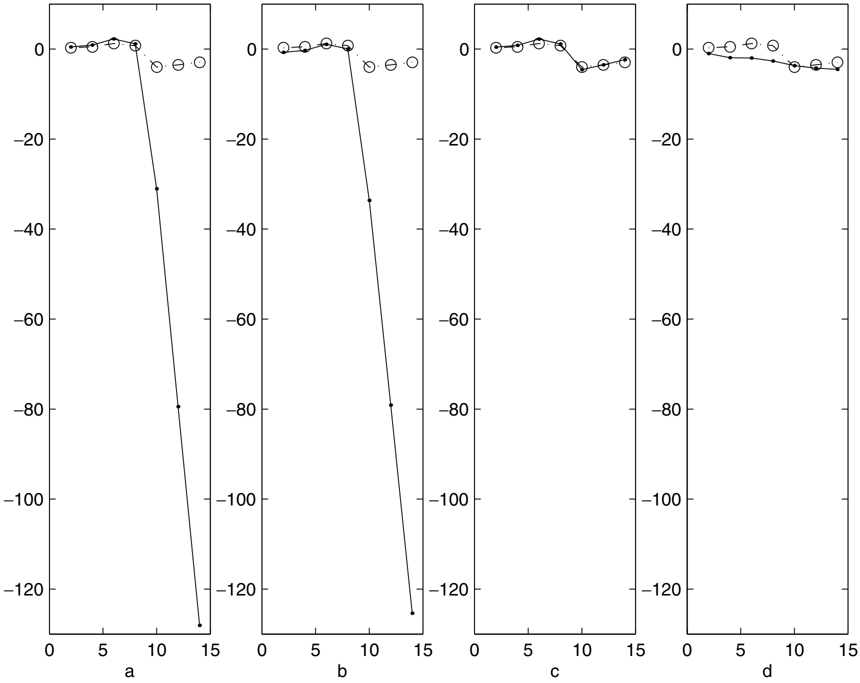

Fig. 5. Log-likelihoods $L_{1:t}$ versus time t for Experiment A with $N = 10^2$ samples: (a) WRF, (b) EnKF, (c) MEF, (d) MFF. The *circles* (joined by *dotted lines*) are the exact values from the Fokker–Planck solution of ref. 21, and the *black dots* (joined by *solid lines*) are the approximate values from the particle filters.

This formula is consistent with the basic assumption of the EnKF method that statistics of the system are Gaussian. Note, however, that it is not practical to use this formula for EnKF when the dimension of the state-space p is large, since the calculation of the determinant Det$\mathbf{C}$ at each time-step would cost $O(p^3)$ multiplications. In Fig. 7 we plot the entropies for $N = 10^2$ and in Fig. 8 for $N = 10^4$. Consistent with the results for the means and variances, we see that for MEF and MFF there is little difference in these plots at different N, except that the results for the smaller N are more random and rougher. All of the methods agree in assigning a high information content to the final measurements, slow to decay to zero, although EnKF poorly predicts the level. In general, EnKF consistently underpredicts the relative entropy and furthermore its approximation to the entropy often increases between measurements, violating the H-theorem.[9] During transitions the MFF method also underpredicts the information content of measurements because it (falsely) interprets them as a return to the steady-state statistics described by the invariant measure rather than the passage of the system

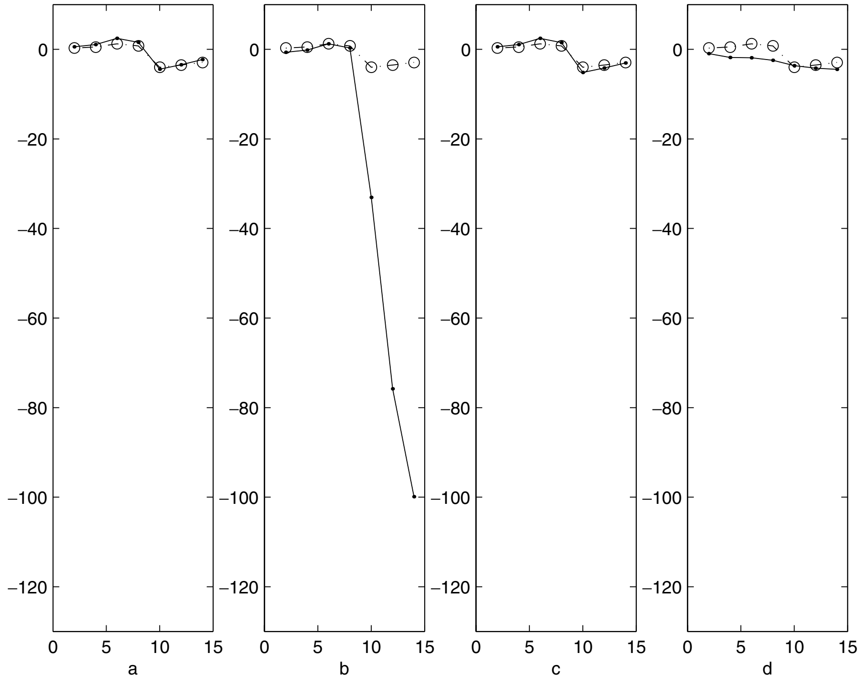

Fig. 6. Log-likelihoods $L_{1:t}$ versus time t for Experiment A with $N = 10^4$ samples: (a) WRF, (b) EnKF, (c) MEF, (d) MFF. Symbols as in Fig. 5.

through the rare saddle-point state at $x = 0$. Away from transitions, the results for MFF are similar to those for MEF. The entropy from MEF is very close to the exact entropy.

4.1.2. Experiment B

Our second estimation experiment is for the same stochastic model (32) but now with noise strength $\kappa = 0.7$. Because of the greater value of the diffusion, transitions from one well to another are much more frequent and the mean residence time in a well, as calculated from Eq. (35), is now $\overline{\tau} \approx 65.8$. The time required to make a transition is also somewhat shorter, taking about 1–2 time units, but the fraction of time spent in transitions is greatly increased, to about 0.01–0.03. Thus, out of 100 randomly selected particles, a small handful may be expected to be in the process of switching to the other well. Based upon our considerations in the preceding subsection, we can expect that each of the particle filtering methods will work well in this situation, using as few as just 100 samples. We carry out Experiment B in order to verify this expectation.

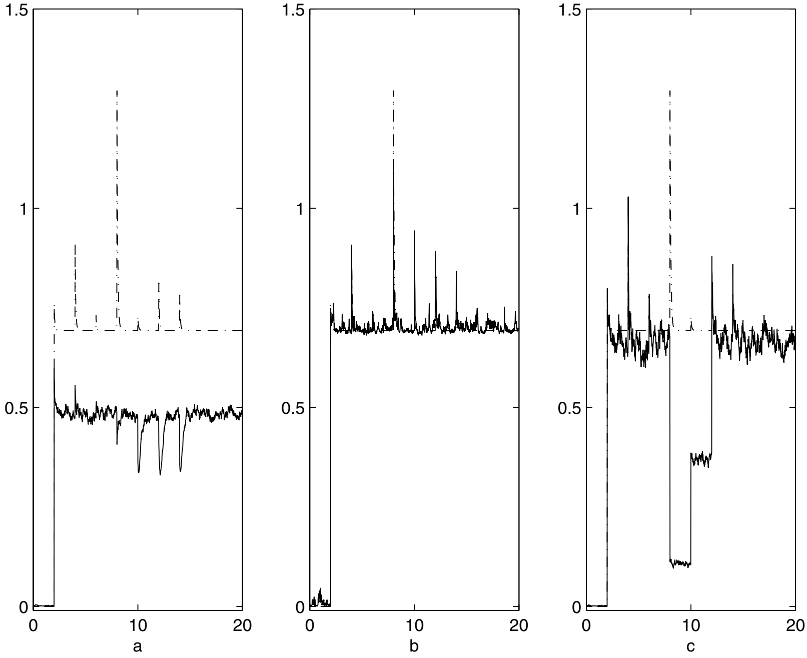

Fig. 7. Relative entropy for Experiment A with $N = 10^2$ samples: (a) EnKF, (b) MEF, (c) MFF. The *solid line* is the approximation and the *dashed line* is the exact result from the Fokker–Planck solution.

We consider again a 20 unit time-segment of a single realization, in this case containing a transition of the sample out of the well at $x = -1$ and then a second transition back into it. As before, seven measurements are taken separated by $\Delta T = 2$ time units and contaminated with normal random errors of mean zero and variance $R = 0.04$. In MEF and MFF we use $\mu_+ = 0.9322$ and $C_+ = 0.0477$ in the mixture model, calculated as discussed previously. We present results of all the particle filters only for $N = 10^2$, since the curves simply become smoother for increasing N and are otherwise unchanged. The means and standard deviations are plotted in Fig. 9. Of the four methods, they may be rated in order as MEF, WRF, EnKF, and MFF, from best to worst. However, all of the methods are relatively successful here and give quite similar approximations to the filter mean. The worst failing of the MFF method is that it, as usual, tends to overestimate the variance. These conclusions from inspection of the graphs are made quantitative by calculation of the relative mean errors, presented in Table II (for both $N = 10^2$ and $N = 10^4$).

We next consider the log-likelihoods $L_{1:t}$ of the four particle filtering methods, presented versus time t as before. The results are plotted in Fig. 10, again

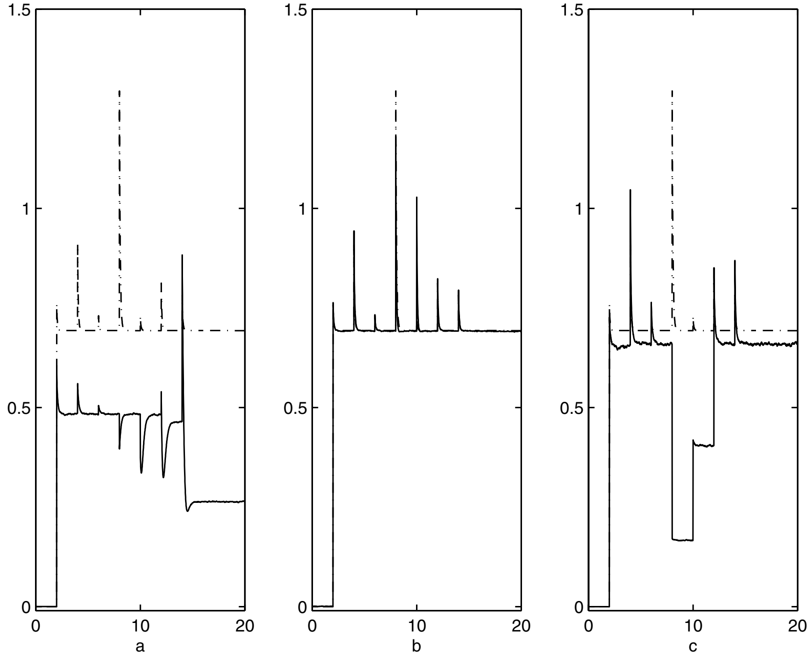

Fig. 8. Relative entropy for Experiment A with $N = 10^4$ samples: (a) EnKF, (b) MEF, (c) MFF. *Symbols* as in Fig. 7.

just for $N = 10^2$. We now see that all of the methods work reasonably well, but that WRF and MEF are both especially accurate. The remaining discrepancies between the results of these two methods and those of the Fokker-Planck solution for the log-likelihoods are apparently due only to statistical errors in the former and discretization errors in the latter. The MFF results again slightly underestimate the true log-likelihoods, consistent with the overestimate of the variances seen in Fig. 9. However, the MFF approximation is reasonably good here. Of all the

Table II. Relative Errors in Experiment B

N	WRF	EnKF	MEF	MFF
(a) Filter Mean				
10^2	0.04648154	0.07271474	0.04568027	0.13560871
10^4	0.00739388	0.05204111	0.00687383	0.09932946
(b) Filter Standard Deviation				
10^2	0.08999746	0.18219459	0.08845087	0.404094241
10^4	0.01668710	0.15788042	0.01384682	0.379310678

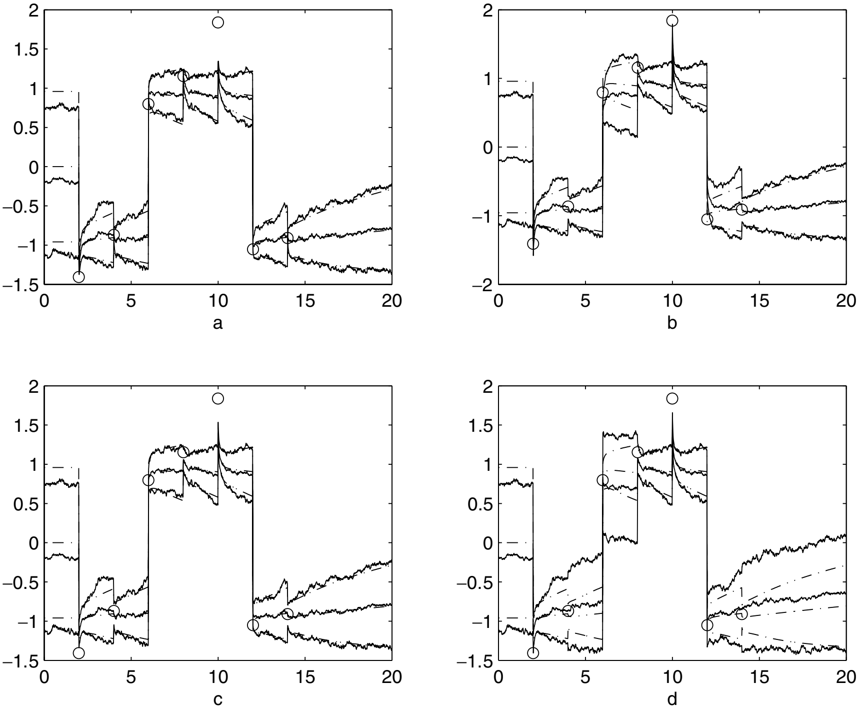

Fig. 9. Particle filter results for Experiment B with $N = 10^2$ samples: (a) WRF, (b) EnKF, (c) MEF, (d) MFF. *Symbols* as in Fig. 2.

particle filtering methods, EnKF gives the worst approximation to the log-likelihoods. This may be somewhat surprising, in view of the fact that its approximations to the means and variances in Fig. 9 are relatively accurate (better than those of MFF, for example). This poor performance should be viewed as a failure of the Gaussian assumption for the statistics, embodied in the standard Kalman formula (37) that we adopted for EnKF. The true likelihoods are not normal distributions, as assumed in (37). We should caution again that a better test of the methods from the point of view of maximum-likelihood estimation would be to compare their maximizers over a set of parameters, for a given set of observations $\mathbf{y}_1, \ldots, \mathbf{y}_T$. For this purpose, only increments or differences of the log-likelihoods matter, not the absolute values.

Finally, we consider the relative entropy as calculated approximately by EnKF, MEF, and MFF with $N = 10^2$. The results are plotted in Fig. 11. EnKF and MFF perform very similarly, with both somewhat underpredicting the entropy. EnKF also violates the H-theorem by yielding occasionally an increasing relative

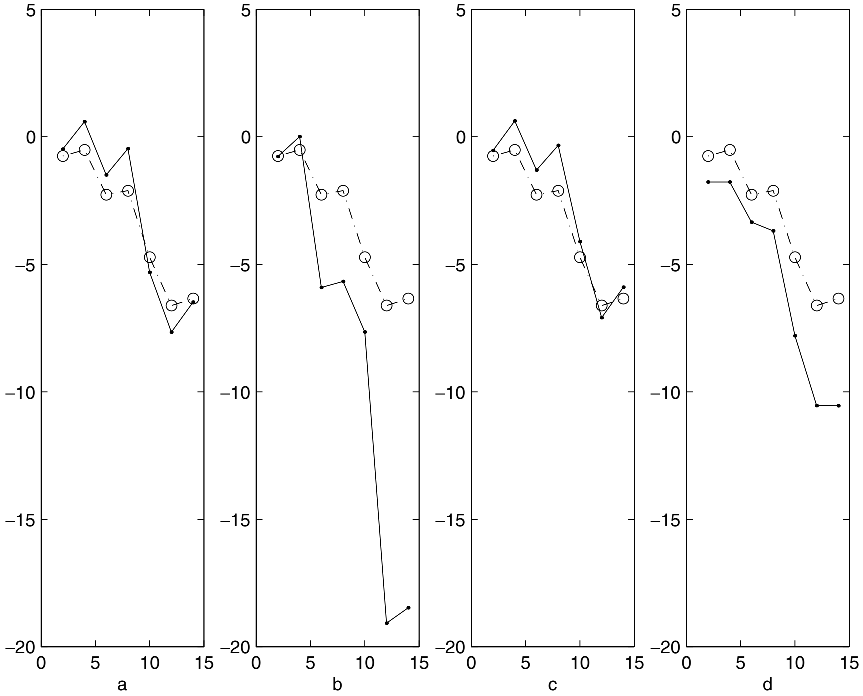

Fig. 10. Log-likelihoods $L_{1:t}$ versus time t for Experiment B with $N = 10^2$ samples: (a) WRF, (b) EnKF, (c) MEF, (d) MFF. *Symbols* as in Fig. 5.

entropy. MEF gives a quite good approximation to the exact entropy calculated from the Fokker-Planck solution. Increasing N further makes all the curves in Fig. 11 smoother but does not otherwise change the results for any of the methods. Although WRF gives outcomes in Experiment B comparable to those of MEF in all other respects, MEF has the advantage that it provides also an accurate approximation of relative entropy.

4.2. Lorenz Model

Our last experiment will be for the chaotic 3-dimensional dynamical system of Lorenz,[(42)] given by the differential equations:

$$\frac{dx}{dt} = \sigma(y - x), \qquad \frac{dy}{dt} = (r - z)x - y, \qquad \frac{dz}{dt} = xy - bz, \tag{38}$$

with coefficients classically chosen as $\sigma = 10, r = 28, b = 8/3$. We include this example to illustrate a set of issues in the application of the particle filtering methods to deterministic dynamical systems. *A priori* this will be a stringent test

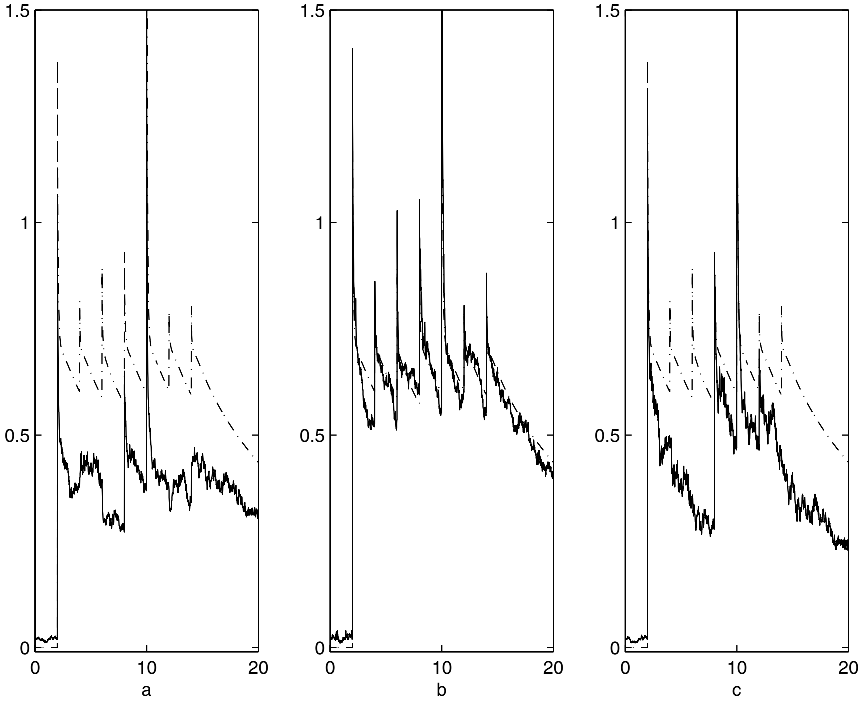

Fig. 11. Relative entropy for Experiment B with $N = 10^2$ samples: (a) EnKF, (b) MEF, (c) MFF. *Symbols* as in Fig. 7.

of the entropy-based methods, for a couple of reasons. First, the probability densities in phase space for a deterministic dynamics evolve in time under the Liouville equation[14] and the relative entropy of two solutions of this hyperbolic PDE is conserved in time.[9] In this sense, the H-theorem that we have exploited for stochastic systems holds for a deterministic dynamics in only a trivial sense. Furthermore, the invariant measure of a dissipative dynamical system lives generally on a strange attractor with zero Lebesgue measure. For example, the Lorenz dynamics at long times lives on the famous butterfly attractor with fractal dimension about 2.06. As a consequence, any measure absolutely continuous with respect to Lebesgue has infinite entropy relative to the invariant measure. Thus, we cannot even define (a finite) relative entropy with respect to the long-term "climate"' for most reasonable initial distributions.

The answer to both of these problems, we shall see, is obtained by considering "reduced" (or "marginal" or "coarse-grained") densities of a subset of variables. It is generally expected that "sufficiently chaotic" deterministic dynamics will have natural invariant measures on their attractors—called SRB measures—which are smooth along unstable manifolds.(68) (Note that this has recently been rigorously proved for the Lorenz model in ref. 65). This SRB measure will have reduced densities absolutely continuous with respect to Lebesgue for subsets of variables whose number is less than the fractal dimension of the attractor. Thus, it will be possible to define relative entropies with respect to "climate" for such reduced sets of variables. Furthermore, it is also expected that "coarse-graining" of high-dimensional chaotic dynamical systems will produce effective dynamics of such subsets of variables which is essentially stochastic.(19,40) Insofar as the Markov property is valid, an H-theorem will hold in the reduced phase-space of this subset of variables. In that case, the relative entropy of marginal densities will decrease between measurements and our maximum-entropy/minimum information ansatz for filter densities is still well-motivated. Note that, even if the H-theorem is not valid, the reduced densities for smooth initial probability distributions will converge in general to the reduced densities of the invariant measure at long times. Thus, the relative entropy of marginals with respect to "climate" will generally converge to zero, even if not monotonically.

For these reasons, we can expect that our entropy-based methods will work well also for "sufficiently chaotic" deterministic dynamics and we shall illustrate this point with the Lorenz 1963 model (38). Note, however, this system is not a particularly good showcase for our methods. The Lorenz model, with the standard choice of parameters, should be generally quite similar in its behavior to the stochastic Double Well model in the Experiment B considered in the last section. The phase point of the system switches chaotically from wing to wing of the attractor, with a residence time on each wing of similar order as the time to make the transition. Thus, we expect that all of the parametric resampling methods (including EnKF) will be able to track the transitions with a relatively small number of samples, $N = 10^2$, say. Our entropy methods apply to the Lorenz model but are unlikely to perform substantially better here than EnKF.

We shall compare all of the particle methods with the results of a convergent scheme, WRF, for a large number of samples. Since the Lorenz dynamics is deterministic, we use a density kernel estimator to improve the representation of the filter density, as discussed in Appendix A.1. To determine an optimal value of the kernel width δ_N for sample size N, we employ a *double density method*.(11) In this approach, the kernel width is chosen to minimize the difference between the density kernel estimates for two different choices of the kernel function K. For our application, we take

$$\delta_N = \underset{\delta}{\operatorname{argmin}} \left\{ \int_0^T dt \|\boldsymbol{\mu}_N(t; G, \delta) - \boldsymbol{\mu}_N(t; U, \delta)\| \right\} \tag{39}$$

where $\boldsymbol{\mu}_N(t; K, \delta)$ is the N-sample empirical mean of the state vector $\mathbf{x}(t)$ for the density kernel K with width δ, and G and U are Gaussian and uniform densities, respectively, with mean 0 and standard deviation 1. We shall verify numerically that the WRF results with δ_N chosen by (39) converge as $N \to \infty$. These results will then be taken as the exact conditional statistics for comparison with the parametric particle filtering methods.

To apply the entropy filtering schemes, MEF and MFF, to the Lorenz Eq. (38) we must build a mixture model $Q_M(\mathbf{x}, t)$ for the reference distribution. In our experiment below, we shall sample the initial conditions from the invariant measure on the strange attractor, which is thus the time-independent background. Because we are interested mainly in the switching transitions from one wing to another of the attractor, we shall employ a Gaussian mixture of complexity $M = 2$. We construct the component weights, means, and covariances by using the function $I(x, y, z) = \text{sign}\,(x + y)$ as an "index" to characterize the regimes of the model. Thus, we consider the complementary sets $\{\text{sign}\,(x + y) = \pm 1\}$, which each contain one wing of the attractor. We then compute a single long time-trajectory of the Lorenz system (38) for an initial condition on the attractor and extract from it the probabilities $w_\pm$ of these two events, and the conditional means $\boldsymbol{\mu}_\pm$ and covariance matrices $\mathbf{C}_\pm$:

$$\boldsymbol{\mu}_\pm = (\bar{x}_\pm, \bar{y}_\pm, \bar{z}_\pm)^\top = (\pm 6.36389,\ \pm 6.69471602,\ 23.5506805)^\top \quad (40)$$

$$\mathbf{C}_\pm = \begin{pmatrix} 22.3056857 & 20.2011608 & \pm 24.9259341 \\ 20.2011608 & 36.3717702 & \pm 1.57754284 \\ \pm 24.9259341 & \pm 1.57754284 & 74.3283071 \end{pmatrix} \quad (41)$$

The numerical results have been symmetrized under the reflection $(x, y, z) \to (-x, -y, z)$ that maps one wing to the other. We then construct the mixture model with $w_\pm = 0.5$ and with $\boldsymbol{\mu}_\pm$, $\mathbf{C}_\pm$ from (40), (41). This construction guarantees that the model has the same second-order statistics (mean and covariance) as the exact invariant measure. In Fig. 12 we compare the mixture model with two Gaussian components and the Lorenz butterfly attractor. Although it is relatively crude, the mixture model captures the dominant bimodality of the Lorenz model statistics.

4.2.1. Experiment C

In our numerical experiment we integrate the system of Eq. (38) by the 4th-order Runge-Kutta method with an integration step of $\Delta t = 1/60$. We take as "reality" a particle started at $\mathbf{x} = (1.508870, -1.531271, 25.46091)^\top$. Measurements on the first two components (x, y) are collected at sampling intervals of $\Delta T = \frac{1}{6}, \frac{1}{3}, \frac{2}{3}$, and $\frac{4}{3}$ time units over the interval $0 < t < 16$ and then contaminated with Gaussian errors of mean zero $\mathbf{0}$ and covariance $\mathbf{R} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 4 \end{smallmatrix}\right)$ for all measurement times. We seek the conditional statistics given this "observational" data.

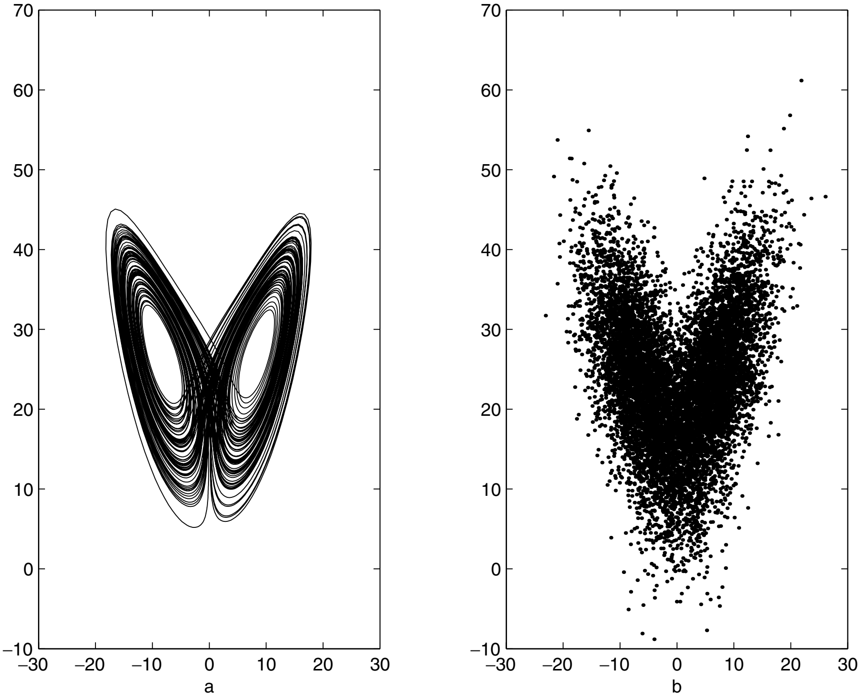

Fig. 12. (a) The Lorenz attractor and (b) scatterplot of samples from the mixture model ($M = 2$), both projected to the xy-plane.

We should say a few words about the implementation of MEF and MFF in this context, since this example is a little less trivial and thus more instructive than the double-well system considered in the previous experiments. The measured variable $\mathbf{h}_t(\mathbf{x}) = (x, y)^\top$ is now a 2-vector, so that the dual variable $\boldsymbol{\lambda}_t$ is also a 2-vector and $\boldsymbol{\Lambda}_t$ is a symmetric 2×2 matrix. Thus, the function $F_t(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ that appears in the minimization (14) in the matching step of MEF depends upon 5 variables, while the function $F_t(\boldsymbol{\lambda})$ used in (22) for MFF depends upon 2 variables. The evaluation of the functions and their derivatives using the formulas in Appendix B thus involved 2×2 matrix operations for the former (e.g. Cholesky decomposition and matrix inversion) and operations on 2-vectors for the latter (e.g. multiplication by a known 2×2 matrix). The minimizations in each case were carried out with a conjugate gradient scheme, using a feasible Armijo line-search in (14) for MEF, since the domain of the convex function $F_t(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ has a non-empty complement.

Once the matching and updating steps were carried out by means of these mimimizations, the new maximum-entropy distributions with updated parameters were resampled. We might have used the same procedure for the Lorenz model as we did earlier for the double-well model, based upon Karhunen–Loève expansions for the Gaussian components of the mixture model (15). However, in realistic applications of MEF this will not be practical, since it would require computing the eigenvalues and eigenvectors of each of the symmetric matrices $\mathbf{C}_m(\Lambda_{t^+})$, $m = 1, \ldots, M$ at every measurement time t. Thus, we have employed instead a more economical sampling scheme discussed in Appendix C, using a Metropolis-Hastings algorithm to sample the Gaussian component $N(\boldsymbol{\mu}_m(\boldsymbol{\lambda}_{t^+}, \boldsymbol{\Lambda}_{t^+}), \mathbf{C}_m(\Lambda_{t^+}))$ for $m = \pm$. That is, based upon the "Hamiltonian" (83), we accepted or rejected proposals sampled from $N(\boldsymbol{\mu}_m(\boldsymbol{\lambda}_{t^+}, \boldsymbol{\Lambda}_{t^+}), \mathbf{C}_m)$, with *fixed* covariance $\mathbf{C}_m$, via its Karhunen-Loève expansion:

$$\mathbf{x}' = \boldsymbol{\mu}_m(\boldsymbol{\lambda}_{t^+}, \boldsymbol{\Lambda}_{t^+}) + \sum_{a=1}^{3} \xi_a \sqrt{\gamma_m^{(a)}}\hat{\mathbf{e}}_m^{(a)}, \qquad m = \pm. \tag{42}$$

Here $\boldsymbol{\mu}_m(\boldsymbol{\lambda}_{t^+}, \boldsymbol{\Lambda}_{t^+})$ is the vector given by (64) in Appendix B, ξ_a, $a = 1, 2, 3$ are i.i.d. $N(0, 1)$ random variables,

$$\gamma_\pm^{(1)} = 86.1296844, \quad \gamma_\pm^{(2)} = 44.5555211, \quad \gamma_\pm^{(3)} = 2.3205575$$

are the eigenvalues of the fixed matrices $\mathbf{C}_m$, $m = \pm$ in (41), and

$$\hat{\mathbf{e}}_\pm^{(1)} = \begin{bmatrix} 0.4096545 \\ 0.1945717 \\ \pm 0.8912491 \end{bmatrix}, \quad \hat{\mathbf{e}}_\pm^{(2)} = \begin{bmatrix} 0.3744077 \\ 0.8550490 \\ \mp 0.3587618 \end{bmatrix}, \quad \hat{\mathbf{e}}_\pm^{(3)} = \begin{bmatrix} 0.8318666 \\ -0.4806589 \\ \mp 0.2774256 \end{bmatrix}$$

are the corresponding eigenvectors, or conditional EOF's for the Lorenz model. For full details of this sampling algorithm, see Appendix C. In the case of MFF we could resample using (42) directly, without an accept/reject criterion, because in MFF the covariances of the Gaussian mixture components for the updated distribution are just the constant matrices $\mathbf{C}_\pm$ in (41). This is true in general for MFF and is another simplifying feature of that method.

Now let us consider results for Experiment C obtained by the different particle filtering schemes.

In Figs. 13 and 14 we illustrate the convergence of the WRF method. The optimization in (39) gives $\delta_N = 0.6$ for $N = 10^2$, and $\delta_N = 0.1$ for $N = 10^4$. The plots in Fig. 13(a) and (b) show the results for $\bar{x}(t)$, the conditional mean of the first coordinate as a function of time, with both numbers of samples. Clearly, there is little difference between the WRF results with $N = 10^2$ and $N = 10^4$. For comparison, we have plotted the original solution trajectory from which measurements were extracted. As can be seen, the "real" solution is here nearly recoverable from the measurements. Figure 14(a) and (b) shows $\sigma_x(t)$, the conditional standard deviation of the first coordinate, for both values of N, and these also differ very little. Similar results have also been found for statistical moments of the other variables y, z of the system. Thus, the WRF results for

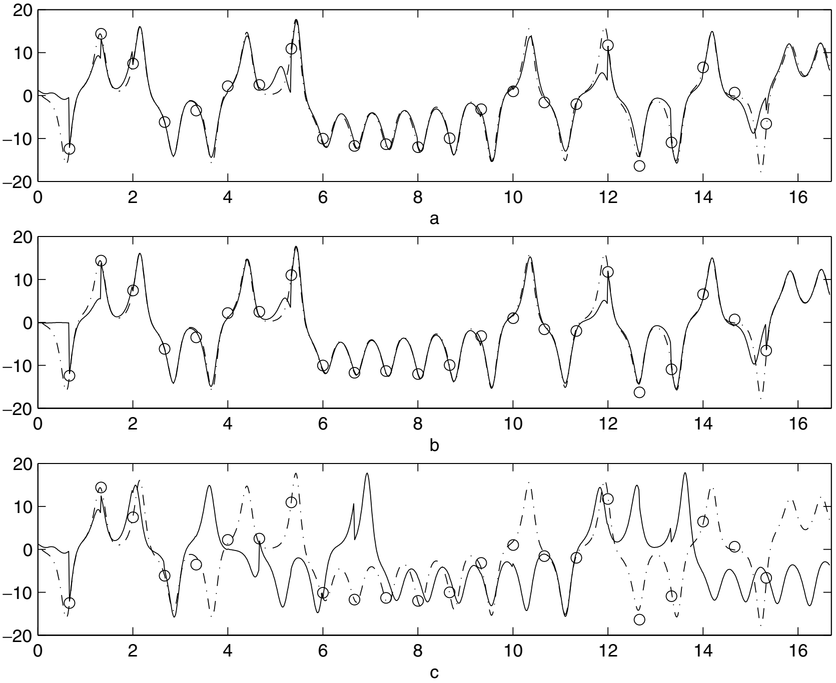

Fig. 13. WRF results for $\bar{x}(t)$ in Experiment C with $\Delta T = 2/3$. (a)$N = 10^2$, $\delta_N = 0.6$, (b) $N = 10^4$, $\delta_N = 0.1$, and (c) $N = 10^2$, $\delta_N = 0.1$. Measurement data shown as *circles*, mean as *solid line*, original solution trajectory as *dot-dashed line*.

$N = 10^4$ appear to be converged, and we shall take them as the exact conditional statistics of the Lorenz model with these measurements and use them as a standard of comparison for the other particle filters. It should be noted that, although the WRF results with $N = 10^2$ are already quite accurate, this depends crucially upon the choice of kernel width as the optimal value δ_N. In Figs. 13(c) and 14(c) we show that the WRF results for $\bar{x}(t)$ and $\sigma_x(t)$ with $N = 10^2$ are quite different if we use instead $\delta = 0.1$, for example. In general, the results of WRF in this deterministic model depend quite sensitively on the choice of kernel width δ. To get the good results in Figs. 13(a) and 14(a) with $N = 10^2$, we had to scan over about 100 values of δ to find the approximate minimum in (39). This is the same amount of work as to carry out the calculation with $N = 10^4$ for just a single kernel width. Thus, for very high-dimensional deterministic dynamics WRF as employed here would not be a practical filtering method.

We shall present results for EnKF, MEF, and MFF with $N = 10^2$, all compared with results of the optimally-tuned WRF method for $N = 10^4$, taken as exact. We will not show results of EnKF, MEF and MFF for $N = 10^4$, because they

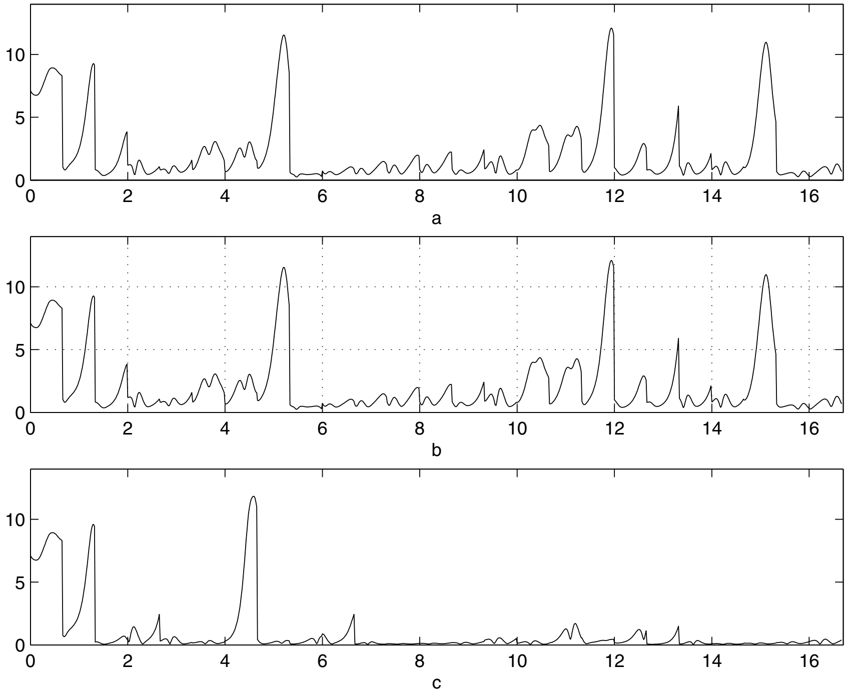

Fig. 14. WRF results for $\sigma_x(t)$ in Experiment C with $\Delta T = 2/3$. (a) $N = 10^2,\ \delta_N = 0.6$, (b) $N = 10^4,\ \delta_N = 0.1$, and (c) $N = 10^2,\ \delta_N = 0.1$. The conditional standard deviation is plotted versus time as a *solid line*.

are almost identical to those for $N = 10^2$. We can thus assume that the latter well represent the converged approximations of the parametric methods as $N \to \infty$.

We first give in Table III the relative errors in means and standard deviations of the x-coordinate for each of the three methods with $N = 10^2$, as functions of the sampling interval ΔT. We see that MFF gives the least good results for all ΔT considered. EnKF and MEF switch roles, with EnKF performing better for small $\Delta T \leq 1/3$ and MEF better for large $\Delta T \geq 2/3$. The slight superiority of EnKF over MEF for small ΔT must be due to the fact that, at measurement times, EnKF uses a gaussian model for the forecast density whose mean and covariance are those of the full three-dimensional state vector $\mathbf{x} = (x, y, z)^\top$. In contrast, MEF uses the mean and covariance only of the measured vector $\mathbf{h}_t(\mathbf{x}) = (x, y)^\top$ to construct its model prior. When measurements are very frequent, this additional information incorporated in EnKF proves advantageous. Of course, the MEF method could be modified to use as well the mean and covariance of the full state vector in order to construct its model prior, although this might prove expensive in practice. When the sampling interval ΔT is large, then the system has more

Table III. Relative Errors in Experiment C

ΔT	EnKF	MEF	MFF
(a) Filter Mean			
1/6	0.0950	0.2458	0.4229
1/3	0.2211	0.3329	0.6081
2/3	0.6341	0.5558	0.7489
4/3	0.8164	0.7498	0.8315
(b) Filter Standard Deviation			
1/6	1.0457	1.7846	7.6775
1/3	1.5034	1.6041	4.4428
2/3	1.1548	1.0200	1.8964
4/3	0.7215	0.6529	1.0528

time between observations to relax to its invariant distribution on the two-winged butterfly attractor and, in that case, MEF profits from its two-component mixture model of the forecast density. In the rest of this section we shall consider only the case $\Delta T = 2/3$ where MEF is marginally superior to EnKF.

In Figs. 15 and 16 we show the results of the EnKF, MEF, and MFF methods with $N = 10^2$ for the mean $\bar{x}(t)$ and the standard deviation $\sigma_x(t)$. As expected, we see that all of the methods do a reasonable job of approximating the filter mean, with the MEF errors smaller by about 10–20 percentage points than those of the other two, and with the EnKF errors just slightly smaller than those for MFF. All of the methods underestimate—or even miss—a few transitions that occur in the exact filter mean but follow its general trends. For the filter standard deviation, MEF and EnKF perform quite similarly, but MFF is considerably worse. While all three methods tend to overestimate the standard deviations, those for MFF are 2–3 times too large. The results that we see here are generally consistent with those obtained in other estimation experiments we have performed on the Lorenz model (38). We find that for the means MEF is somewhat better at larger sampling intervals than EnKF, which is itself slightly better than MFF, but all three perform rather well. All three methods give standard deviations too large, but MFF much larger than the other two.

Although only modest gains have been achieved for the filter mean and variance by the use of our entropy methods, it may be that their performance is substantially better for the log-likelihood and relative entropy, as we found in the previous double-well diffusion model. In Fig. 17 we plot the results for the log-likelihoods of the three methods EnKF, MEF, and MFF with $N = 10^2$ for Experiment C with $\Delta T = 2/3$. The results of WRF with $N = 10^4$ are again taken as exact. By comparison, MEF performs best, MFF second best, and EnKF least well. It is interesting to note that MFF overestimates the log-likelihoods, although it also overestimates the variances in Fig. 17. This seems to be due to MFF's

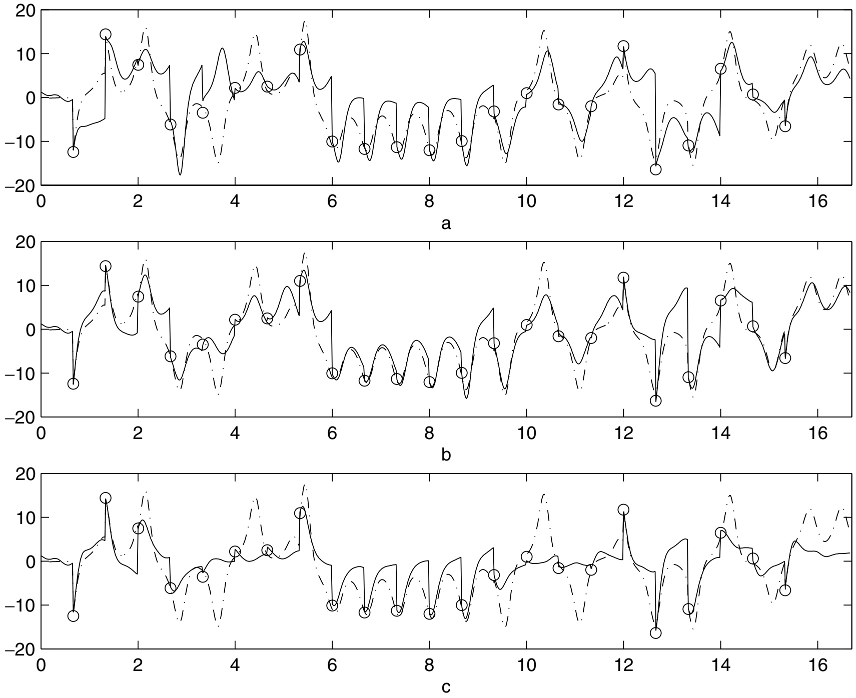

Fig. 15. Approximate filter means $\bar{x}(t)$ for Experiment C with $\Delta T = 2/3$ and $N = 10^2$. (a) EnKF, (b) MEF, and (c) MFF. Measurement data shown as *circles*, approximations as *solid lines*, exact filter result (WRF) as *dot-dashed line*.

missing prefactors in its estimate of $\ln \mathcal{N}_t$. We can make a crude estimate of the correction as $-\frac{1}{2}\ln[(2\pi)^q \mathrm{Det}\mathbf{C}^Y_{t-}]$, which is exact for linear dynamics. If we add this correction to the MFF result for the log-likelihood (not shown), then it also becomes an underestimate and lies between the results of MEF and EnKF.

Lastly, we consider the relative entropy approximated using the three methods, EnKF, MEF and MFF. These are plotted in Fig. 18(a)–(c) for each method with $N = 10^2$. As for the means and variances, the results on entropy with $N = 10^2$ were so similar to those for $N = 10^4$ that the latter need not be considered here. It can be seen that all of the approximate entropies behave qualitatively similarly, rising discontinuously at measurements and then decaying between measurements, but non-monotonically. This behavior may seem paradoxical, in view of the fact that the exact relative entropy $H(P(t)|Q)$—where $P(t)$ is the filter measure on the Lorenz attractor and $Q = P_*$ is the invariant measure—does not change in time between measurements. However, the approximate entropies that have been constructed all have the property that they must converge to zero at long times between

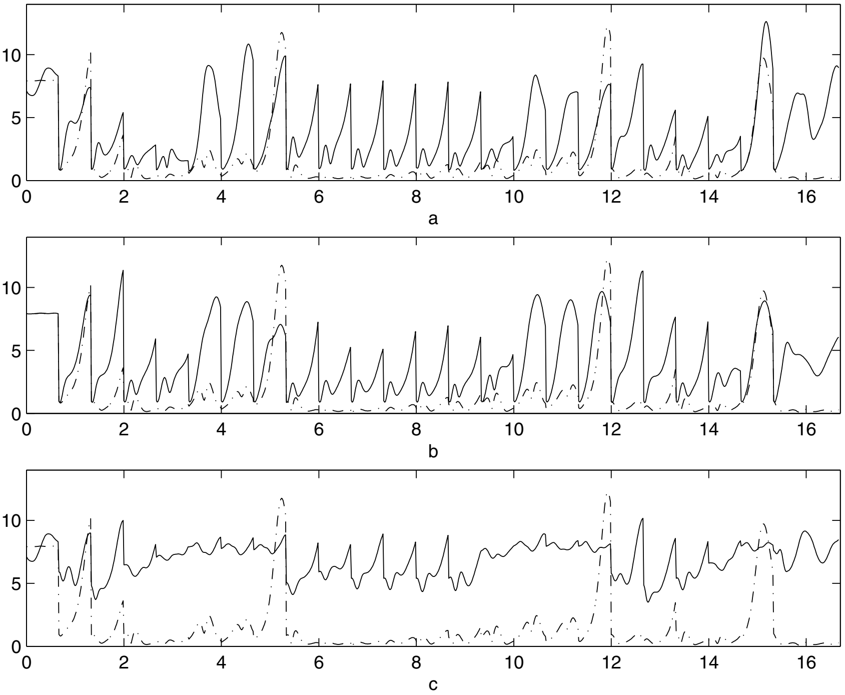

Fig. 16. Approximate filter standard deviations $\sigma_x(t)$ for Experiment C with $\Delta T = 2/3$ and $N = 10^2$. (a) EnKF, (b) MEF, and (c) MFF. Approximations shown as *solid lines* and exact filter result (WRF) as *dot-dashed line*.

measurements, because the moments employed, such as $\overline{x}(t), \overline{y}(t), \overline{x^2}(t), \overline{y^2}(t)$, etc. all converge as $t \to \infty$ to the corresponding averages in the invariant measure $\overline{x}_*, \overline{y}_*$, etc. As a consequence, $\lim_{t\to\infty} P_M(t) = Q_M$, so also $H(P_M(t)|Q_M) \to 0$. Since the approximate entropies are based upon only a few statistical moments of the Lorenz system, this amounts to an implicit "coarse-graining" of the entropy.

In fact, it follows from the exponential formula (12) for the maximum-entropy distribution that, in the MEF method, $H(P_M(t)|Q_M) = H(\widetilde{P}_M(t)|\widetilde{Q}_M)$, where $\widetilde{P}_M, \widetilde{Q}_M$ are the marginal distributions of P_M, Q_M on measured variables (here, x and y). Because the invariant measure $Q = P_*$ of the Lorenz model is smooth on unstable manifolds,[65,68] the marginals $\widetilde{P}(t), \widetilde{Q}$ both have densities with respect Lebesgue measure on the 2-dimensional space of x, y coordinates and $H(\widetilde{P}(t)|\widetilde{Q})$ is finite. However, unlike $H(P(t)|Q)$, which is time-independent, the relative entropy of the marginal measures $\widetilde{P}(t), \widetilde{Q}$ is expected between measurements to converge toward zero (but not necessarily monotonically). Thus, it is more proper to compare the MEF entropy with $H(\widetilde{P}(t)|\widetilde{Q})$. We have approximated the

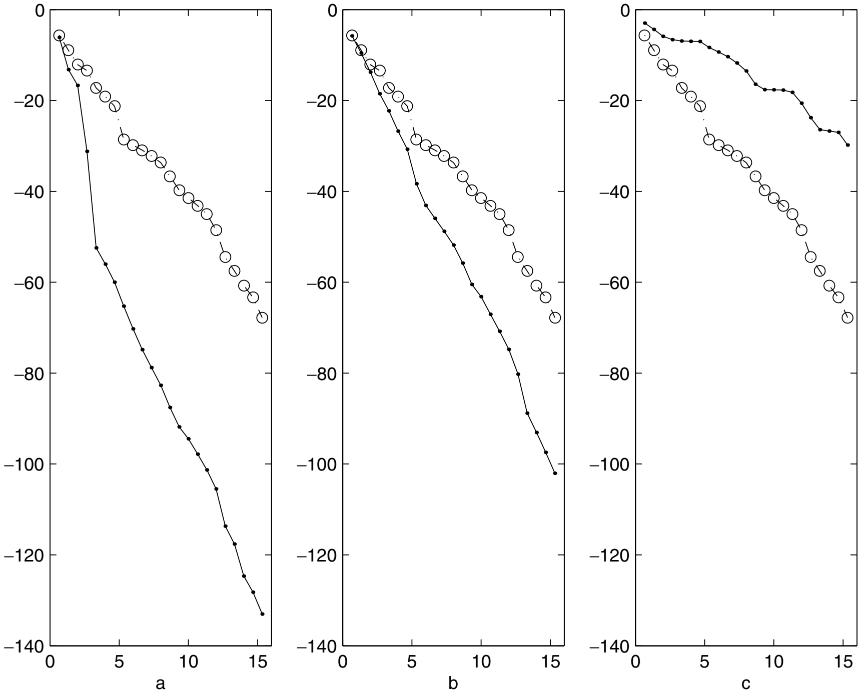

Fig. 17. Log-likelihoods $L_{1:t}$ versus time t for Experiment C with $\Delta T = 2/3$ and $N = 10^2$ samples: (a) EnKF, (b) MEF, (c) MFF. The *circles* (joined by *dotted lines*) are the exact values from the WRF method with $N = 10^4$, and the *black dots* (joined by *solid lines*) are approximations from the other particle filters.

latter using the WRF solution to construct histograms that represent the densities $\widetilde{P}(x, y; t)$, $\widetilde{Q}(x, y)$ in the x, y-plane and then used a discrete quadrature formula for the integral (10). The results are shown in Fig. 18(d) for histograms on a 40×40 grid in the rectangle $-20 < x < 20$, $-28 < y < 28$ with $N = 10^4$ samples. The resolution is low and the statistical fluctuations are still quite large, but these results will suffice for a rough comparison. While the MEF entropy does not agree perfectly with this relative entropy of the marginals, it does show qualitatively similar behavior and it is more accurate quantitatively than either EnKF (which is too large) and MFF (which is too small).

5. SUMMARY AND CONCLUSIONS

In this paper we have introduced two new entropy-based particle filtering schemes. The first method uses maximum-entropy parametric models to

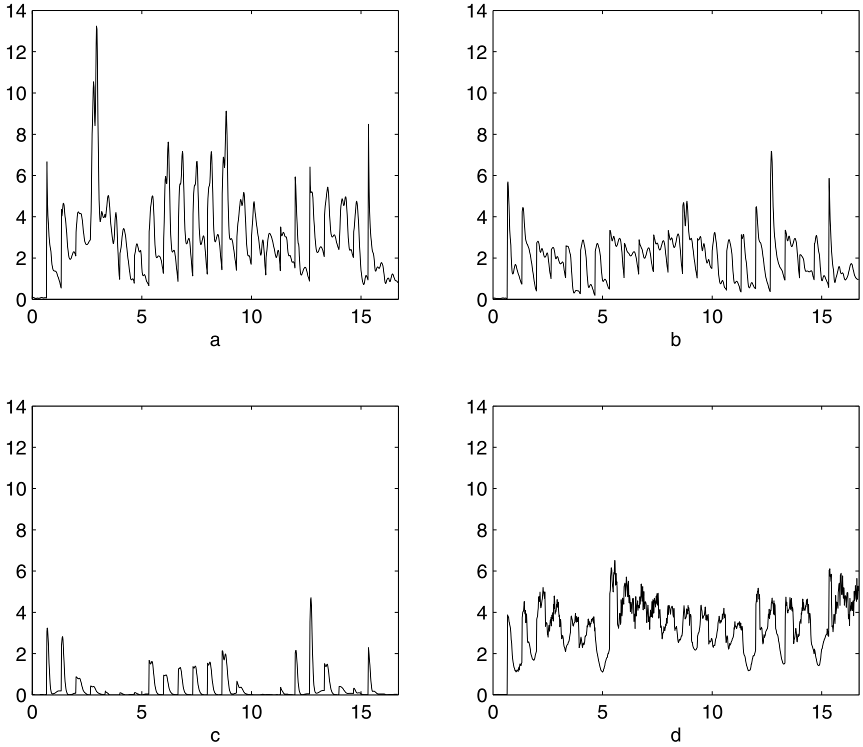

Fig. 18. Entropy for Experiment C with $\Delta T = 2/3$ and $N = 10^2$. (a) EnKF; (b) MEF; and (c) MFF; (d) Relative entropy of xy-marginals using WRF with $N = 10^4$.

implement the update by Bayes' Theorem at measurement times, and has been called by us the Maximum Entropy Filter (MEF). The second method updates the filter densities by a maximum-entropy criterion that implements Bayes' Theorem only in a mean-field sense, and was called by us the Mean-Field Filter (MFF). We have compared these new methods with two standard ensemble/particle filters, the Weight Resampling Filter (WRF) and the Ensemble Kalman Filter (EnKF), which are reviewed in Appendix A. In terms of the computational cost to implement them for a fixed number of samples, the methods can be ranked in order, from cheapest to most expensive, as WRF, MFF, EnKF, and MEF, when $p \gg q \gg 1$. Here the integer p is the dimension of the state space and q is the dimension of the measured random variable. See Appendix D.

With small samples sizes N, the standard methods perform very poorly when there are subsets of the state space that have low priori but high posteriori probabilities, as in our Experiment A. Although WRF gives optimal results in the limit $N \to \infty$, events of low priori probability are insufficiently represented when the number of samples is too small. The method may thus converge only with quite large N. In EnKF, the probability density in state space before the measurement is modelled by a Gaussian density centered in the region of high prior probability. Unless subsequent measurements are very accurate (lower variance than the Gaussian model) or very frequent, the gain from the measurements is insufficient to shift the state to the regions of high posterior probability. The parametric resampling methods that have been introduced in this paper were designed to work better precisely in the circumstance where there is a large disparity between priori and posteriori probabilities. This superior performance has been confirmed in our experiment A with $N = 10^2$, where MEF performed the best and MFF second best of all the four methods. In circumstances such as these, MEF should be preferred, but MFF is an acceptable, cheaper substitute if the former is unaffordable.

Even when the ratios between prior and posterior probabilities are not large, the optimally convergent WRF scheme encounters another difficulty with deterministic dynamics, as in our Experiment C, because the resampling step is ineffectual in such cases. We have found that a modification using a density kernel method to represent the filter density in state space may give good results with small N. However, the accuracy of this representation can depend sensitively on the kernel width, and it will not be practical to search for the optimal width if only a small number of samples is available. Here, EnKF can work well with moderate N, and, based on the Experiment C that we performed, it can be recommended, at least when the number of measurements q is not too large. MEF gives somewhat better results at large sampling intervals but is also more expensive. If $q \gg 1$, then both MEF and EnKF may be too costly to apply, unless further improvements of the algorithms are made.(3,18,31,34) In that case, MFF is a practical substitute.

Where these difficulties do not occur, as in our Experiment B, WRF can give the optimal results economically with small N. In such cases, or where large numbers of samples are readily available, it is the preferred method.

The parametric resampling methods introduced in this work are very robust. While WRF and EnKF may or may not perform well, depending on the circumstances, MEF gave results of quality from excellent to good in all of the experiments we performed. The results of MFF were less accurate, but generally acceptable and less costly. Both MEF and MFF converge rapidly as the number of samples N is increased, and, except for larger fluctuations, gave nearly the same results for $N = 10 \sim 10^2$ as for $N = 10^4$. The price that must be paid for these advantages is that the parametric methods cannot be carried out "blindfolded" but require some prior knowledge of the system. In the method as presented here, we constructed the parametric densities by minimizing the information relative to a carefully chosen model of the background distribution (with no measurements whatsoever). This is a well-motivated choice for Markov stochastic processes, because of an "H-theorem" which requires that the relative entropy decreases monotonically in time. Even for deterministic dynamics, the relative entropies of marginal distributions—which are the only statistics practically accessible for large-scale systems—converge to zero. Our maximum-entropy parametric models have other important practical advantages for the filtering problem: matching parameters to ensemble statistics can be carried out by minimization of a convex function; the Bayes update is implemented by a trivial change of parameters; and, efficient methods exist for sampling from the maximum-entropy distributions. The algorithms yield as side-products the log-likelihood and relative entropy, which are of independent interest. We believe these features should make the maximum-entropy filtering methods very useful in a variety of applications to high-dimensional nonlinear dynamical systems.

The initial tests that we have made here are only for quite low-dimensional dynamics, although the formalism has been developed with high (or even infinite-dimensional) dynamical systems in mind. In a paper in preparation, we demonstrate the good performance of our entropy-based filtering schemes for a stochastic PDE model of oceanic thermohaline circulation.[8,20] This model has two stable steady states of circulation and a consequent bimodality of its statistics that is readily described by a two-component mixture. We anticipate that our methods will work well, in general, for large-scale systems whose statistics are well-described by a Gaussian mixture model of relatively low complexity.

ACKNOWLEDGMENTS

We wish to thank F. J. Alexander, T. W. N. Haine, G. Johnson, D. J. Lea, J. M. Restrepo, and C. Wunsch for helpful discussions. This work was supported by NSF grant # DMS-0113649.

APPENDIX A: STANDARD ENSEMBLE FILTERS

In this appendix, we briefly review some of the standard particle/ensemble methods that have been proposed to solve the optimal filtering problem.

A.1. Convergent Particle Schemes for Optimal Filtering

The basic idea of all ensemble/particle methods is to employ an ensemble $\mathbf{x}_t^{(n)}$, $n = 1, \ldots, N$ of solutions of (1) with independent realizations of the noise in order to approximate the filter densities by empirical measures

$$P^{(N)}(\mathbf{x}, t) = \sum_{n=1}^{N} w_t^{(n)} \delta^p\left(\mathbf{x} - \mathbf{x}_t^{(n)}\right). \tag{43}$$

The non-negative real numbers $w_t^{(n)}$, $n = 1, \ldots, N$ are called *importance weights* and must satisfy $\sum_{n=1}^N \ w_t^{(n)} = 1$. In common to all these methods is the very desirable property that they implement the prediction step (5) in the Bayes recursion exactly, at least in the limit $N \to \infty$. Various methods differ in how they approximate the update step (6).

In the simplest approach, the sample weights are updated by the formula

$$w_{t^+}^{(n)} = \frac{G_t\left(\mathbf{y}_t|\mathbf{x}_t^{(n)}\right)}{\mathcal{N}_t} w_{t^-}^{(n)}, \qquad n = 1, \ldots, N \tag{44}$$

which is determined so that the N-sample approximations (43) satisfy (6) exactly. As in (6), $\mathcal{N}_t$ is a normalization factor to ensure that $\sum_{n=1}^N \ w_{t^+}^{(n)} = 1$ for all times t. If the initial samples are chosen so that $\mathbf{x}_0^{(n)}$, $n = 1, \ldots, N$ are i.i.d. distributed according to $\mathcal{P}_0$, then the initial weights may be taken to be $w_0^{(n)} = 1/N$ for all $n = 1, \ldots, N$. Initialized in this manner, the algorithm outlined above provides a systematic approach to approximating the filter distributions, via (43). For convenient reference, we shall call this simple standard particle method the Weighted Ensemble Filter (WEF). For more details, see refs. 12, 36. It has been proved by Moral[(46)] that the approximate filter densities produced by this method converge (weakly) to the optimal filter density as $N \to \infty$. Unfortunately, despite being a convergent method, WEF often performs poorly in practice. As can be seen from (44), if measurements are very accurate or if the outcomes of measurement are very far from the predictions of the samples, then updated weights may be very small. In that case, the effective size of the ensemble can be much less than N, since samples with small importance weights do not contribute significantly to any averages. Therefore, the convergence of the WEF algorithm is often quite slow.

To overcome the difficulty with non-uniform weights, the update (44) in the above method may be augmented with a *resampling* step, as was originally suggested by Metropolis and Ulam.[(44)] That is, a new ensemble $\mathbf{x}_{t^+}^{(n)}$, $n = 1, \ldots, N$ with uniform weights $1/N$ may be selected independently from the set of pre-measurement samples $\mathbf{x}_{t^-}^{(n')}$ with probabilities $w_{t^+}^{(n')}$, for $n' = 1, \ldots, N$. In the process, realizations with high probability are multiply resampled and "cloned," while states with low probability are not sampled at all and become "extinct". In the case of genuinely stochastic dynamics, the resampling procedure described above may already suffice. However, for deterministic dynamics, "cloned" individuals have identically the same behavior in the future and act collectively as a single sample with high weight. In that case, the results with resampling are equivalent to those for WEF. To deal with this situation, the representation of the filter density by means of the empirical measure (43) may be improved with *kernel smoothing:*

$$P^{(N,\delta)}(\mathbf{x},t) = \sum_{n=1}^{N} w_t^{(n)} K_\delta^p\left(\mathbf{x} - \mathbf{x}_t^{(n)}\right), \tag{45}$$

where the "density kernel" $K_\delta^p(\mathbf{x} - \mathbf{x}')$ is an approximate delta function in $\mathbb{R}^p$ with width proportional to δ.[57,58,67] Resampling from a distribution like (45) may be accomplished in two steps: first, select an index $n' = 1, \ldots, N$ in the sum with probability $w_t^{(n')}$ and, second, select a random sample $\mathbf{x}_t^{(n)} = \mathbf{x}_t^{(n')} + \boldsymbol{\rho}^{(n)}$ where $\boldsymbol{\rho}^{(n)}$ are i.i.d. samples drawn from $K_\delta^p(\boldsymbol{\rho})$, successively for $n = 1, \ldots, N$. If the density kernel is of a simple standard type, such as a multivariate Gaussian or a uniform distribution on a hypercube, then there are efficient algorithms for drawing the random samples from $K_\delta^p(\boldsymbol{\rho})$. In this way, the problem of "cloned" samples is eliminated by the random perturbations or "mutations" of each sample. If the kernel width is chosen δ_N as a function of N so that $\delta_N \to 0$ suitably as $N \to \infty$, then the kernel density estimator (45) will also converge to the true density $P(\mathbf{x}, t)$ as $N \to \infty$. We refer to standard texts[57,58,67] for more details.

The algorithm with resampling as described above is one of the most widely used particle filtering methods (e.g. see ref. 12), which we shall refer to, for convenience, as the Weighted Resampling Filter (WRF). As with the simpler WEF method, the approximate density from WRF has been proved to converge weakly to the optimal filter density as $N \to \infty$ (refs. 45, 47 or 10 for a recent review.) The rate of convergence of the approximation error to zero is the same as for standard Monte Carlo, i.e. $O(N^{-1/2})$. In order to choose a large enough value of N, one may simply monitor the convergence of statistics of interest for increasing number of samples. Alternatively, a recent paper[22] has proposed a method whereby the number of samples required for convergence may be determined automatically. The WRF method has become popular because it is simple to use, handles with ease nonlinearity of the dynamics and non-Gaussianity of statistics, and gives optimal results under conditions that are frequently achievable in practice. It is possible to construct examples that "break" this method, even with N quite large, but WRF is probably the method of choice in cases where a large number of samples are readily available.

A.2. Ensemble Kalman Filter

The Ensemble Kalman Filter (EnKF) was proposed by Evensen[16,17] (with an important correction in Burgers et al.[7]). This is also a sequential particle filtering method, like those discussed in the previous subsection. However, the update by Bayes' theorem is only implemented approximately. It applies in the most straightforward form only when observation errors are normal random vectors with mean $\mathbf{0}$ and covariance $\mathbf{R}_t$ and when measurement functions are affine,

$$\mathbf{h}_t(\mathbf{x}) = \mathcal{H}_t \mathbf{x} + \mathbf{d}_t, \tag{46}$$

where $\mathcal{H}_t$ is a $q \times p$ matrix and $\mathbf{d}_t$ is a q-vector for each time t. If the measured variables are not given by an affine state function, then an extension of the method is required. One approach is to linearize the measurement function by a Taylor expansion around the mean, with $\mathbf{h}_t(\mathbf{x}) = \mathbf{h}_t(\overline{\mathbf{x}}) + \mathbf{J}_t \cdot (\mathbf{x} - \overline{\mathbf{x}}) + O(|\mathbf{x} - \overline{\mathbf{x}}|^2)$ truncated to linear terms.(33) However, this approach requires a calculation of the Jacobian matrix $\mathbf{J}_t$ and, furthermore, appeals to a linear approximation which may be inaccurate. A second and more theoretically satisfactory approach is to extend the state-space by working with the joint state-observation vector $\mathbf{z} = (\mathbf{x}, \mathbf{h}_t(\mathbf{x}))$.(3,63) In this extended state-space the measurement function is always given by a linear projection, i.e. $\mathbf{h}_t(\mathbf{x}, \mathbf{y}) = \mathbf{y}$. The major virtue of this method is that it treats the nonlinearity in the measurements exactly.

As in all Kalman filtering schemes, the statistical basis of EnKF is the use of a Gaussian model for the prior $P(\mathbf{x}, t^-)$. The mean and covariance of this model are obtained by empirical averages over the N-sample ensemble. A new ensemble is then generated by performing, for each sample state, a linear interpolation between the original state and a sample measurement, weighted by the so-called "Kalman gain matrix". The EnKF update algorithm may be divided into three steps, as follows:

(i) *Matching:* The mean $\boldsymbol{\mu}_{t^-}$ and covariance $\mathbf{C}_{t^-}$ before the measurement are obtained from particle averages:

$$\boldsymbol{\mu}_{t^-} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_{t^-}^{(n)}, \quad \mathbf{M}_{t^-} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_{t^-}^{(n)}\left[\mathbf{x}_{t^-}^{(n)}\right]^\top \tag{47}$$

with $\mathbf{C}_{t^-} = \mathbf{M}_{t^-} - \boldsymbol{\mu}_{t^-}\boldsymbol{\mu}_{t^-}^\top$.

(ii) *Resampling:* An N-sample ensemble of measurement outcomes is generated from

$$\mathbf{y}_t^{(n)} = \mathbf{y}_t + \boldsymbol{\epsilon}_t^{(n)}, \quad n = 1, \dots, N \tag{48}$$

where $\boldsymbol{\epsilon}_t^{(n)}$, $n = 1, \dots, N$ are i.i.d. $N(\mathbf{0}, \mathbf{R}_t)$ random vectors.

(iii) *Updating:* A new N-sample ensemble of state vectors is obtained from

$$\mathbf{x}_{t^+}^{(n)} = \mathbf{x}_{t^-}^{(n)} + \mathcal{K}_t\left[\mathbf{y}_t^{(n)} - \mathbf{h}_t\left(\mathbf{x}_{t^-}^{(n)}\right)\right], \tag{49}$$

with

$$\mathcal{K}_t = \mathbf{C}_{t^-}\mathcal{H}_t^\top[\mathcal{H}_t\mathbf{C}_{t^-}\mathcal{H}_t^\top + \mathbf{R}_t]^{-1}, \tag{50}$$

the *Kalman gain matrix*

An easy computation(7) shows that the mean and covariance of the updated ensemble are given in the limit $N \to \infty$ by

$$\boldsymbol{\mu}_{t^+} = \boldsymbol{\mu}_{t^-} + \mathcal{K}_t\big[\mathbf{y}_t - \boldsymbol{\mu}^H_{t^-}\big] \tag{51}$$

$$\mathbf{C}_{t^+} = \mathbf{C}_{t^-} - \mathcal{K}_t\big[\mathbf{C}^H_{t^-} + \mathbf{R}_t\big]\mathcal{K}_t^\top, \tag{52}$$

where $\boldsymbol{\mu}^H_{t^-} = \mathcal{H}_t\boldsymbol{\mu}_{t^-} + \mathbf{d}_t$ and $\mathbf{C}^H_{t^-} = \mathcal{H}_t\mathbf{C}_{t^-}\mathcal{H}_t^\top$ are the mean and covariance, respectively, of the measured variable $\mathbf{h}_t(\mathbf{x})$ in the N-sample ensemble before the measurement. These formulas are the well-known results of the Kalman filtering procedure, which are derived by applying Bayes' theorem to a Gaussian prior.(24) Notice, however, that the posterior density $P(\mathbf{x}, t^+)$ represented by the samples $\mathbf{x}^{(n)}_{t^+}$, $n = 1, \ldots, N$ is non-Gaussian, because the original samples $\mathbf{x}^{(n)}_{t^-}$, $n = 1, \ldots, N$ are drawn from a non-Gaussian density $P(\mathbf{x}, t^-)$. The Gaussian model $N(\mathbf{x}; \boldsymbol{\mu}_{t^-}, \mathbf{C}_{t^-})$ for the prior distribution has the same mean and covariance as the N-sample ensemble before the measurement, but other moments of the two distributions will be generally unequal. Thus, the Ensemble Kalman Filter is only guaranteed to give the correct conditional statistics, in the limit $N \to \infty$, when the system statistics are indeed Gaussian. Otherwise, its estimates of the conditional mean and covariance converge to suboptimal values.

APPENDIX B: MAXIMUM-ENTROPY THERMODYNAMICS

In this appendix we derive the equations of a "thermodynamic formalism" for maximum-entropy mixture models. We assume that the measurement function is affine, $\mathbf{h}(\mathbf{x}) = \mathcal{H}\mathbf{x} + \mathbf{d}$, and consider the mixture model (9), with

$$N(\mathbf{x}; \boldsymbol{\mu}_m, \mathbf{C}_m) = \frac{\exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_m)^\top \mathbf{C}_m^{-1}(\mathbf{x} - \boldsymbol{\mu}_m)\right]}{\sqrt{(2\pi)^p \mathrm{Det}\mathbf{C}_m}}. \tag{53}$$

If the measurement function is not affine, then the procedures discussed below can still be applied, by employing similar devices as those discussed in Appendix A for Kalman filtering, including linearization(33) and extended state-space formulations.(3,63)

We first prove the following simple but useful lemma:

$$\exp\left[\boldsymbol{\lambda}\boldsymbol{\cdot}\mathbf{h}(\mathbf{x}) + \frac{1}{2}\boldsymbol{\Lambda}\boldsymbol{:}\mathbf{h}(\mathbf{x})\mathbf{h}^\top(\mathbf{x})\right] N(\mathbf{x}; \boldsymbol{\mu}_m, \mathbf{C}_m) = Z_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) N(\mathbf{x}; \boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}), \mathbf{C}_m(\boldsymbol{\Lambda})) \tag{54}$$

with functions $Z_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$, $\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$, $\mathbf{C}_m(\boldsymbol{\Lambda})$ described in detail below. Since the lefthand side in (54) is a product of Gaussians, the equality is proved easily by completing the square, with the results:

$$\mathbf{C}_m(\boldsymbol{\Lambda}) = \left(\mathbf{C}_m^{-1} - \boldsymbol{\mathcal{H}}^\top \boldsymbol{\Lambda} \boldsymbol{\mathcal{H}}\right)^{-1} \tag{55}$$

$$\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \boldsymbol{\mu}_m + \mathbf{C}_m(\boldsymbol{\Lambda})\boldsymbol{\mathcal{H}}^\top\left(\boldsymbol{\lambda} + \boldsymbol{\Lambda}\boldsymbol{\mu}_m^H\right) \tag{56}$$

$$Z_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \sqrt{\frac{\mathrm{Det}\mathbf{C}_m(\boldsymbol{\Lambda})}{\mathrm{Det}\mathbf{C}_m}} \exp\left[\boldsymbol{\lambda}^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \boldsymbol{\Lambda}\mathbf{d}\right.$$

$$\left. -\frac{1}{2}\boldsymbol{\mu}_m^\top \mathbf{C}_m^{-1}\boldsymbol{\mu}_m + \frac{1}{2}\boldsymbol{\mu}_m^\top(\boldsymbol{\lambda}, \boldsymbol{\Lambda})\mathbf{C}_m^{-1}(\boldsymbol{\Lambda})\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})\right]. \tag{57}$$

Here we introduce $\boldsymbol{\mu}_m^H = \boldsymbol{\mathcal{H}}\boldsymbol{\mu}_m + \mathbf{d}$ and $\mathbf{C}_m^H = \boldsymbol{\mathcal{H}}\mathbf{C}_m\boldsymbol{\mathcal{H}}^\top$, the mean and covariance of $\mathbf{h}(\mathbf{x})$ for $\mathbf{x}$ an $N(\boldsymbol{\mu}_m, \mathbf{C}_m)$ random variable.

These formulas can be simplified by using the following matrix identities, valid for $\mathbf{A}$ and $\mathbf{C}$ any $p \times p$ and $q \times q$ non-singular matrices, respectively, and $\mathbf{B}$ an arbitrary $q \times p$ matrix:

$$(\mathbf{A}^{-1} + \mathbf{B}^\top\mathbf{C}^{-1}\mathbf{B})^{-1} = \mathbf{A} - \mathbf{A}\mathbf{B}^\top(\mathbf{B}\mathbf{A}\mathbf{B}^\top + \mathbf{C})^{-1}\mathbf{B}\mathbf{A}, \tag{58}$$

$$(\mathbf{A}^{-1} + \mathbf{B}^\top\mathbf{C}^{-1}\mathbf{B})^{-1}\mathbf{B}^\top = \mathbf{A}\mathbf{B}^\top(\mathbf{B}\mathbf{A}\mathbf{B}^\top + \mathbf{C})^{-1}\mathbf{C}. \tag{59}$$

These identities are standard in the Kalman filtering literature.(24) From (58) it follows immediately that

$$\mathbf{C}_m(\boldsymbol{\Lambda}) = \mathbf{C}_m + \mathbf{C}_m\boldsymbol{\mathcal{H}}^\top\boldsymbol{\Gamma}_m^H\left[\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}\right]^{-1}\boldsymbol{\Lambda}\boldsymbol{\mathcal{H}}\mathbf{C}_m, \tag{60}$$

where we have defined $\boldsymbol{\Gamma}_m^H = [\mathbf{C}_m^H]^{-1}$. Note that we have written this formula so that it is valid even if $\boldsymbol{\Lambda}$ is singular. Applying (59) gives likewise

$$\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \boldsymbol{\mu}_m + \mathbf{C}_m\boldsymbol{\mathcal{H}}^\top\boldsymbol{\Gamma}_m^H\left[\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}\right]^{-1}\left(\boldsymbol{\lambda} + \boldsymbol{\Lambda}\boldsymbol{\mu}_m^H\right) \tag{61}$$

In these formulas, the combination $\boldsymbol{\mathcal{K}}_m = \mathbf{C}_m\boldsymbol{\mathcal{H}}^\top\boldsymbol{\Gamma}_m^H(\boldsymbol{\Lambda} - \boldsymbol{\Gamma}_m^H)^{-1}\boldsymbol{\Lambda}$ is the analogue of the *Kalman gain matrix* and $\mathbf{r}_m = \boldsymbol{\mathcal{H}}\mathbf{C}_m$ the *representer* of the mth mixture component.(6,66) [In fact, our calculations here are a natural generalization of the representer solution for Gaussian mixture models; e.g. see (64) below.] Finally, simplifications can be made in the exponent of the normalization factor $Z_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ by using the formula

$$\mathbf{C}_m^{-1}(\boldsymbol{\Lambda})\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \mathbf{C}_m^{-1}\boldsymbol{\mu}_m + \boldsymbol{\mathcal{H}}^\top(\boldsymbol{\lambda} + \boldsymbol{\Lambda}\mathbf{d})$$

and dotting with formula (56) for $\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$. Cancelling many terms, one finds finally that

$$Z_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \sqrt{\frac{\mathrm{Det}\boldsymbol{\Gamma}_m^H}{\mathrm{Det}(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda})}} \exp\Big[-\frac{1}{2}\big(\boldsymbol{\mu}_m^H\big)^\top \boldsymbol{\Gamma}_m^H \boldsymbol{\mu}_m^H + \frac{1}{2}\big(\boldsymbol{\Gamma}_m^H \boldsymbol{\mu}_m^H + \boldsymbol{\lambda}\big)^\top \big(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}\big)^{-1} \big(\boldsymbol{\Gamma}_m^H \boldsymbol{\mu}_m^H + \boldsymbol{\lambda}\big) \Big]. \tag{62}$$

We have also used the identity, for $p \times q$ matrix $\mathbf{A}$ and $q \times p$ matrix $\mathbf{B}$,

$$\mathrm{Det}(\mathbf{I} - \mathbf{AB}) = \exp\left[-\sum_{k=1}^{\infty} \frac{1}{k} \mathrm{Tr}((\mathbf{AB})^k) \right] = \mathrm{Det}(\mathbf{I} - \mathbf{BA}),$$

by cyclicity of the trace, in order to write

$$\begin{aligned} \mathrm{Det}(\mathbf{C}_m)/\mathrm{Det}(\mathbf{C}_m(\boldsymbol{\Lambda})) &= \mathrm{Det}(\mathbf{C}_m) \cdot \mathrm{Det}\big(\mathbf{C}_m^{-1} - \boldsymbol{\mathcal{H}}^\top \boldsymbol{\Lambda} \boldsymbol{\mathcal{H}}\big) \\ &= \mathrm{Det}(\mathbf{I} - \mathbf{C}_m \boldsymbol{\mathcal{H}}^\top \boldsymbol{\Lambda} \boldsymbol{\mathcal{H}}) = \mathrm{Det}(\mathbf{I} - \boldsymbol{\Lambda} \boldsymbol{\mathcal{H}} \mathbf{C}_m \boldsymbol{\mathcal{H}}^\top) \\ &= \mathrm{Det}\big[\mathbf{I} - \boldsymbol{\Lambda}\big(\boldsymbol{\Gamma}_m^H\big)^{-1}\big] = \mathrm{Det}\big(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}\big)/\mathrm{Det}\big(\boldsymbol{\Gamma}_m^H\big). \end{aligned}$$

For purposes of numerical evaluations on the computer, it is convenient to introduce $\boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ as the solution of the linear equation

$$\big(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}\big) \cdot \boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \boldsymbol{\Gamma}_m^H \boldsymbol{\mu}_m^H + \boldsymbol{\lambda}. \tag{63}$$

This equation can be solved using a Cholesky factorization of $\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}$, since this matrix must be positive-definite in order for the model density to be statistically realizable with the given matrix $\boldsymbol{\Lambda}$. [Notice that $\ln Z_m$ in (62) must be a convex function of $\boldsymbol{\lambda}$ for realizability to hold; moreover, (55), (59) imply that $(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda})^{-1} = \boldsymbol{\mathcal{H}} \mathbf{C}_m(\boldsymbol{\Lambda}) \boldsymbol{\mathcal{H}}^\top = \mathbf{C}_m^H(\boldsymbol{\Lambda})$, which must be positive-definite.] Introducing the solution $\boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ into (61) gives

$$\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \boldsymbol{\mu}_m + \mathbf{C}_m \boldsymbol{\mathcal{H}}^\top \boldsymbol{\Gamma}_m^H \cdot \big[\boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) - \boldsymbol{\mu}_m^H\big] \tag{64}$$

The Cholesky factor can also be used to calculate the inverse $[\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}]^{-1}$ and the determinant $\mathrm{Det}(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda})$ that appear in the formulae (60) and (62) for $\mathbf{C}_m(\boldsymbol{\Lambda})$ and $Z_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$, respectively. In fact, we may rewrite (62) somewhat to eliminate the inverse matrix:

$$Z_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \sqrt{\frac{\mathrm{Det}\boldsymbol{\Gamma}_m^H}{\mathrm{Det}\big(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}\big)}} \times \exp\left[-\frac{1}{2}\big(\boldsymbol{\mu}_m^H\big)^\top \boldsymbol{\Gamma}_m^H \boldsymbol{\mu}_m^H + \frac{1}{2}\big(\boldsymbol{\Gamma}_m^H \boldsymbol{\mu}_m^H + \boldsymbol{\lambda}\big)^\top \boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) \right]. \tag{65}$$

It is important for the numerical feasibility of these calculations that $\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}$ is a $q \times q$ matrix, where we assume that $q \ll p$.

The derivatives of Z_m are also straightforward to evaluate. We use

$$\begin{aligned}\ln Z_m(\boldsymbol{\lambda},\boldsymbol{\Lambda}) &= \frac{1}{2}\big(\boldsymbol{\Gamma}_m^H\boldsymbol{\mu}_m^H+\boldsymbol{\lambda}\big)^{\top}\big(\boldsymbol{\Gamma}_m^H-\boldsymbol{\Lambda}\big)^{-1}\big(\boldsymbol{\Gamma}_m^H\boldsymbol{\mu}_m^H+\boldsymbol{\lambda}\big)\\ &\quad -\frac{1}{2}\big(\boldsymbol{\mu}_m^H\big)^{\top}\boldsymbol{\Gamma}_m^H\boldsymbol{\mu}_m^H-\frac{1}{2}\mathrm{Tr}\big[\ln\big(\boldsymbol{\Gamma}_m^H-\boldsymbol{\Lambda}\big)-\ln\boldsymbol{\Gamma}_m^H\big] \qquad (66)\end{aligned}$$

and two standard identities for differentiation of a matrix with respect to a parameter: $\frac{\partial}{\partial\lambda}\mathbf{A}^{-1}=-\mathbf{A}^{-1}\frac{\partial\mathbf{A}}{\partial\lambda}\mathbf{A}^{-1}$ and $\frac{\partial}{\partial\lambda}\mathrm{Tr}\ln\mathbf{A}=\mathrm{Tr}(\mathbf{A}^{-1}\frac{\partial\mathbf{A}}{\partial\lambda})$. Then

$$\frac{\partial Z_m}{\partial\boldsymbol{\lambda}}(\boldsymbol{\lambda},\boldsymbol{\Lambda})=Z_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})\boldsymbol{\eta}_m(\boldsymbol{\lambda},\boldsymbol{\Lambda}), \qquad (67)$$

for $i\neq j$,

$$\frac{\partial Z_m}{\partial\Lambda_{ij}}=Z_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})\big\{[\boldsymbol{\eta}_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})]_i[\boldsymbol{\eta}_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})]_j+\big[\big(\boldsymbol{\Gamma}_m^H-\boldsymbol{\Lambda}\big)^{-1}\big]_{ij}\big\}, \qquad (68)$$

and for $i=j$,

$$\frac{\partial Z_m}{\partial\Lambda_{ii}}=\frac{1}{2}Z_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})\big\{[\boldsymbol{\eta}_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})]_i[\boldsymbol{\eta}_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})]_i+\big[\big(\boldsymbol{\Gamma}_m^H-\boldsymbol{\Lambda}\big)^{-1}\big]_{ii}\big\}, \qquad (69)$$

We can now easily deduce the results claimed in the text. First we derive the mixture representation (15) of the maximum-entropy densities. This follows directly from the main lemma (54) with

$$Z(\boldsymbol{\lambda},\boldsymbol{\Lambda})=\sum_{m=1}^{M}w_m Z_m(\boldsymbol{\lambda},\boldsymbol{\Lambda}) \qquad (70)$$

and

$$w_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})=w_m\frac{Z_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})}{Z(\boldsymbol{\lambda},\boldsymbol{\Lambda})}, \qquad (71)$$

where $Z_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})$, $\boldsymbol{\mu}_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})$, $\mathbf{C}_m(\boldsymbol{\Lambda})$, $m=1,\ldots,M$ are given by (65), (64), (60). Second we derive the thermodynamic functions for the mixture model, starting with $F(\boldsymbol{\lambda},\boldsymbol{\Lambda})=\ln Z(\boldsymbol{\lambda},\boldsymbol{\Lambda})$ and $Z(\boldsymbol{\lambda},\boldsymbol{\Lambda})$ given in (70). The derivatives are then obtained from

$$\frac{\partial F}{\partial\boldsymbol{\lambda}}(\boldsymbol{\lambda},\boldsymbol{\Lambda})=\frac{1}{Z(\boldsymbol{\lambda},\boldsymbol{\Lambda})}\sum_{m=1}^{M}w_m\frac{\partial Z_m}{\partial\boldsymbol{\lambda}}(\boldsymbol{\lambda},\boldsymbol{\Lambda})$$

and the similar formula for the derivative with respect to $\boldsymbol{\Lambda}$. Using (67)–(69) one obtains:

$$\frac{\partial F}{\partial\boldsymbol{\lambda}}(\boldsymbol{\lambda},\boldsymbol{\Lambda})=\sum_{m=1}^{M}w_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})\boldsymbol{\eta}_m(\boldsymbol{\lambda},\boldsymbol{\Lambda}), \qquad (72)$$

for $i \neq j$,

$$\frac{\partial F}{\partial \Lambda_{ij}} = \sum_{m=1}^{M} w_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})\left\{\left[\left(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}\right)^{-1}\right]_{ij} + [\boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})]_i[\boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})]_j\right\} \tag{73}$$

and for $i = j$,

$$\frac{\partial F}{\partial \Lambda_{ii}} = \frac{1}{2}\sum_{m=1}^{M} w_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})\left\{\left[\left(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}\right)^{-1}\right]_{ii} + [\boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})]_i[\boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})]_i\right\}. \tag{74}$$

The computational cost to determine $F(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ and its derivatives is dominated by the Cholesky factorization, for which the operation count (number of multiplications) is $O(Mq^3)$. Calculation of the vectors $\boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ is $O(Mq^2)$ operations, while calculation of the inverses $(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda})^{-1}$ requires an additional $O(Mq^3)$ operations. This will be feasible as long as M and q are not too large. On the other hand, even if the $p \times q$ matrices $\mathbf{C}_m\boldsymbol{\mathcal{H}}^\top\boldsymbol{\Gamma}_m^H$ and $q \times p$ matrices $\boldsymbol{\mathcal{H}}\mathbf{C}_m$ are stored in advance, calculating $\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$, $\mathbf{C}_m(\boldsymbol{\Lambda})$ for $m = 1, \ldots, M$ from (64) and (60) requires $O(Mpq)$ and $O(Mp^2q)$ operations, respectively, in addition to the Cholesky factorization. These calculations are expensive if $p \gg q$.

All of the above formulae simplify considerably within the mean-field approximation, and, in fact, remain valid simply upon setting $\boldsymbol{\Lambda} = \mathbf{O}$. Thus, (63) becomes

$$\boldsymbol{\eta}_m(\boldsymbol{\lambda}) = \boldsymbol{\mu}_m^H + \mathbf{C}_m^H\boldsymbol{\lambda}, \tag{75}$$

(65) becomes

$$Z_m(\boldsymbol{\lambda}) = \exp\left[-\frac{1}{2}(\boldsymbol{\mu}_m^H)^\top\boldsymbol{\Gamma}_m^H\boldsymbol{\mu}_m^H + \frac{1}{2}\left(\boldsymbol{\Gamma}_m^H\boldsymbol{\mu}_m^H + \boldsymbol{\lambda}\right)^\top\boldsymbol{\eta}_m(\boldsymbol{\lambda})\right], \tag{76}$$

(64) becomes

$$\boldsymbol{\mu}_m(\boldsymbol{\lambda}) = \boldsymbol{\mu}_m + \mathbf{C}_m\boldsymbol{\mathcal{H}}^\top\boldsymbol{\lambda}, \tag{77}$$

and (60) becomes simply

$$\mathbf{C}_m(\boldsymbol{\lambda}) = \mathbf{C}_m. \tag{78}$$

The thermodynamics also simplifies, with

$$F(\boldsymbol{\lambda}) = \ln Z(\boldsymbol{\lambda}) = \ln\left(\sum_{m=1}^{M} w_m Z_m(\boldsymbol{\lambda})\right), \tag{79}$$

$$\frac{\partial F}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}) = \sum_{m=1}^{M} w_m(\boldsymbol{\lambda})\boldsymbol{\eta}_m(\boldsymbol{\lambda}) = \boldsymbol{\eta}(\boldsymbol{\lambda}), \tag{80}$$

$$\frac{\partial^2 F}{\partial \lambda_i \partial \lambda_j}(\boldsymbol{\lambda}) = \sum_{m=1}^{M} w_m(\boldsymbol{\lambda})\left\{\left[\mathbf{C}_m^H\right]_{ij} + [\boldsymbol{\eta}_m(\boldsymbol{\lambda}) - \boldsymbol{\eta}(\boldsymbol{\lambda})]_i[\boldsymbol{\eta}_m(\boldsymbol{\lambda}) - \boldsymbol{\eta}(\boldsymbol{\lambda})]_j\right\} \tag{81}$$

and $w_m(\boldsymbol{\lambda}) = w_m Z_m(\boldsymbol{\lambda})/Z(\boldsymbol{\lambda})$, as in (71). The cost to calculate $\boldsymbol{\eta}_m(\boldsymbol{\lambda})$, $Z_m(\boldsymbol{\lambda})$ for $m = 1, \ldots, M$ and $F(\boldsymbol{\lambda})$ and its first and second derivatives is $O(Mq^2)$, while the cost to calculate $\boldsymbol{\mu}_m(\boldsymbol{\lambda})$ is $O(Mpq)$. Thus, there are considerable savings with the mean-field approximation.

APPENDIX C: SAMPLING FROM MAXIMUM-ENTROPY MODELS

Direct sampling from the maximum-entropy distributions using the mixture representation (15) is prohibitively expensive when $p \gg 1$. For example, sampling the Gaussian components using their Karhunen-Loève expansions would require calculating the EOF's of the $p \times p$ covariance matrices $\mathbf{C}_m(\boldsymbol{\Lambda}_t^+)$, $m = 1, \ldots, M$ for every new value of $\boldsymbol{\Lambda}_t^+$. It might be possible to calculate the EOF's for the covariance matrices $\mathbf{C}_m(t)$, $m = 1, \ldots, M$ of the components of the mixture-model (9) for $Q_M(\mathbf{x}, t)$, especially if the latter is time-independent or changes sufficiently slowly in time that only a few representative values of t need be considered. In that case, a more efficient sampling strategy can be based upon the identity

$$\mathbf{C}_m^{-1}(\boldsymbol{\Lambda}) = \mathbf{C}_m^{-1} - \boldsymbol{\mathcal{H}}^\top \boldsymbol{\Lambda} \boldsymbol{\mathcal{H}}, \tag{82}$$

the inverse of (55) [where we drop from here on the explicit time label t]. This formula implies that the Gaussian component $N(\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}), \mathbf{C}_m(\boldsymbol{\Lambda}))$ can be sampled by the Metropolis-Hastings algorithm with $N(\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}), \mathbf{C}_m)$ as the proposal distribution and with

$$E(\mathbf{x}) = -\frac{1}{2}[\mathbf{h}(\mathbf{x}) - \boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})]^\top \boldsymbol{\Lambda} [\mathbf{h}(\mathbf{x}) - \boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})] \tag{83}$$

as the "energy function" to calculate acceptance probabilities. Using the Karhunen-Loève representation of $N(\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}), \mathbf{C}_m)$, proposed updates have the form

$$\mathbf{x}' = \boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) + \sum_{a=1}^{p} \xi_a \sqrt{\gamma_m^a}\, \hat{\mathbf{e}}_m^a, \tag{84}$$

where ξ_a, $a = 1, \ldots, p$ are i.i.d. normal random variables and γ_m^a, $\hat{\mathbf{e}}_m^a$ are the eigenvalues and eigenvectors of $\mathbf{C}_m$. Note that the eigensystems do not depend on $\boldsymbol{\Lambda}$ and that the vectors $\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ can be efficiently calculated from (64). The updates (84) are accepted with probability $\min\{1, e^{-\Delta E}\}$ to replace a current state vector $\mathbf{x}$, where $\Delta E = E(\mathbf{x}') - E(\mathbf{x})$.

An efficient algorithm to sample $P_M(\mathbf{x}; \boldsymbol{\lambda}, \boldsymbol{\Lambda})$ is then as follows: First, set $\mathbf{x}_m = \boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ as an initial guess of the state in the mth component for each $m = 1, \ldots, M$. Then, successively for $n = 1, \ldots, N$, choose $\mathbf{x}^{(n)}$ by first selecting a component index $m = 1, \ldots, M$ with probability $w_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$. For the selected m, generate a new trial state $\mathbf{x}_m'$ via (84) and then accept or reject it compared with the current state $\mathbf{x}_m$ by the Metropolis-Hastings algorithm. That is, replace $\mathbf{x}_m$ with $\mathbf{x}_m'$

if accepted and otherwise leave $\mathbf{x}_m$ intact. In either case, after completion of the trial, take $\mathbf{x}^{(n)} = \mathbf{x}_m$. In this way, the entire N-sample ensemble $\mathbf{x}^{(n)},\ n = 1, \ldots, N$ will be generated, distributed according to $P_M(\mathbf{x}; \boldsymbol{\lambda}, \boldsymbol{\Lambda})$. In practice, it is advisable to consider some number n_T of trial vectors $\mathbf{x}'_m$ for each selected component m and to successively accept or reject them, in order to generate each member of the N-sample ensemble. This will help to ensure better equilibration in the Metropolis-Hastings algorithm. Furthermore, it has the benefit for deterministic dynamics that it helps to guarantee that $\mathbf{x}^{(n)} = \mathbf{x}^{(n')}$ for $n \neq n'$, i.e. that members of the ensemble are not identical.

This Metropolis-Hastings scheme will work well if $\boldsymbol{\Lambda}$ is small, but rejection rates will be high if the values of the energy function E in (83) become large. This is precisely what occurs as a consequence of the Bayes' rule update (17), when measurements are very accurate. In fact, in the limit that $\|\mathbf{R}\|$ is small,

$$\boldsymbol{\lambda}^{+} = \boldsymbol{\lambda}^{-} + \mathbf{R}^{-1}\mathbf{y} \approx \mathbf{R}^{-1}\mathbf{y}, \quad \boldsymbol{\Lambda}^{+} = \boldsymbol{\Lambda}^{-} - \mathbf{R}^{-1} \approx -\mathbf{R}^{-1}, \tag{85}$$

and updated values of parameters, to leading order, are independent of their values $\boldsymbol{\lambda}^{-}$, $\boldsymbol{\Lambda}^{-}$ before the measurement. In that case, the mixture model $P_M(\mathbf{x}; \boldsymbol{\lambda}^{+}, \boldsymbol{\Lambda}^{+})$ simplifies considerably. It is easy to show using (60), (64), (65) that, as $\|\mathbf{R}\| \to 0$, the following asymptotic formulas hold for the component weights

$$w_m^{+} = \frac{w_m \exp\left\{-\frac{1}{2}\left[\mathbf{y} - \boldsymbol{\mu}_m^H\right]^{\top} \boldsymbol{\Gamma}_m^H \left[\mathbf{y} - \boldsymbol{\mu}_m^H\right]\right\}}{\mathcal{N}\sqrt{\mathrm{Det}\mathbf{C}_m^H}}[1 + O(\|\mathbf{R}\|)] \tag{86}$$

(where $\mathcal{N}$ is a normalization factor), for component means

$$\boldsymbol{\mu}_m^{+} = \boldsymbol{\mu}_m + \mathbf{C}_m \boldsymbol{\mathcal{H}}^{\top} \boldsymbol{\Gamma}_m^H \left(\mathbf{y} - \boldsymbol{\mu}_m^H\right) + O(\|\mathbf{R}\|) \tag{87}$$

and for component covariances

$$\mathbf{C}_m^{+} = \mathbf{C}_m - \mathbf{C}_m \boldsymbol{\mathcal{H}}^{\top} \left(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Gamma}_m^H \mathbf{R} \boldsymbol{\Gamma}_m^H\right) \boldsymbol{\mathcal{H}} \mathbf{C}_m + O(\|\mathbf{R}\|^2). \tag{88}$$

Notice that the exponential factor in (86) is the normal density $N(\mathbf{y}; \boldsymbol{\mu}_m^H, \mathbf{C}_m^H)$ of the measurement function $\mathbf{h}(\mathbf{x})$ in the mth component, evaluated at the measured value $\mathbf{y}$. Notice also that $\boldsymbol{\mathcal{H}}\boldsymbol{\mu}_m^{+} + \mathbf{d} = \mathbf{y} + O(\|\mathbf{R}\|)$ from (87) and that $\boldsymbol{\mathcal{H}}\mathbf{C}_m^{+}\boldsymbol{\mathcal{H}}^{\top} = \mathbf{R} + O(\|\mathbf{R}\|^2)$ from (88), as would be expected for the limit of very accurate measurements. A simple sampling scheme in this limit, therefore, is to choose components $m = 1, \ldots, M$ with the probabilities w_m^{+} in (86) and then to draw samples from the selected Gaussian component $N(\boldsymbol{\mu}_m^{+}, \mathbf{C}_m^{+})$ directly, e.g. using its Karhunen-Loève representation. For that purpose, the EOF's of the covariance matrices $\mathbf{C}_m^{+}$ may be calculated and stored in advance. It is crucial that $\mathbf{C}_m^{+}$ in (88) does not depend upon $\boldsymbol{\lambda}^{-}$, $\boldsymbol{\Lambda}^{-}$.

The two sampling schemes that have been discussed in this section of the Appendix are efficient and accurate for opposite limits of large $\|\mathbf{R}\|$ and small $\|\mathbf{R}\|$, respectively. Therefore, the best results should be obtained from a hybrid approach, that switches from the first method to the second as $\|\mathbf{R}\|$ decreases. As a practical criterion for switching, the rejection rate of the n_T proposals in the Metropolis-Hastings algorithm may be monitored and the second method employed when the rejection rate becomes too large in the first method.

APPENDIX D: COMPUTATIONAL COSTS OF THE METHODS

We here briefly compare the computational costs of the four main particle filtering methods considered: WRF, EnKF, MEF, and MFF.

WRF: The main cost lies in the computation of the probability density of measurement errors, $G_t(\mathbf{y}_t|\mathbf{x}^{(n)})$, $n = 1, \ldots, N$ for the update (44). When this density is Gaussian, $O(q^3)$ multiplications are required to calculate the determinant Det$\mathbf{R}_t$ and inverse $\mathbf{R}_t^{-1}$, then $O(Nq^2)$ multiplications to calculate the quadratic forms $[\mathbf{y}_t - \mathbf{h}_t(\mathbf{x}^{(n)})]^\top \mathbf{R}_t^{-1}[\mathbf{y}_t - \mathbf{h}_t(\mathbf{x}^{(n)})]$ and Npq multiplications to calculate the values $\mathbf{h}(\mathbf{x}^{(n)})$ of the linear measurement function (46), for $n = 1, \ldots, N$. Resampling requires just N independent random numbers.

EnKF: An efficient scheme to implement EnKF is the representer method.[6,66] In this approach, the $q \times p$ representer matrix $\mathbf{r} = \mathcal{H}\mathbf{C}_{t-}$ is evaluated as the N-sample covariance of the forecast measurement outcome and of the forecast state vector, requiring $O(Npq)$ multiplications. Likewise, $\mathbf{C}_{t-}^H = \mathcal{H}\mathbf{C}_{t-}\mathcal{H}^\top$ is calculated as the N-sample covariance of the forecast measurement outcome vector, requiring an additional $O(Nq^2)$ multiplications. In the limit $p \gg q$ which mostly concerns us, this cost is smaller than that to calculate the representer matrix. Calculation of the inverse and/or Cholesky factors of $\mathbf{C}_{t-}^H + \mathbf{R}_t$ is $O(q^3)$ multiplications and calculation of the N representer coefficient vectors $\mathbf{b}^{(n)} = [\mathbf{C}_{t-}^H + \mathbf{R}_t]^{-1}[\mathbf{y}^{(n)} - \mathbf{h}_t(\mathbf{x}^{(n)})]$ requires an additional $O(Nq^2)$ operations, either by matrix multiplication with the inverse or by backsubstitution using the Cholesky factors. Additionally, Nq random numbers must be generated for the measurement resampling in (48). Finally, update of $\mathbf{x}_{t-}^{(n)}$ to $\mathbf{x}_{t+}^{(n)}$ for $n = 1, \ldots, N$ is obtained via $\mathbf{x}_{t+}^{(n)} = \mathbf{x}_{t-}^{(n)} + \mathbf{r}^\top \mathbf{b}^{(n)}$, equivalent to (49). This is requires $O(Npq)$ operations, equal in cost to the calculation of the representer matrix.

Other implementations of EnKF[3,18,31,34] make different trade-offs and may have costs that scale differently in the parameters N, p, q. For example, ref. 18 has used a standard matrix identity to avoid inversion of the $q \times q$ matrix $\mathbf{C}_{t-}^H + \mathbf{R}_t$ at cost $O(q^3)$ and to replace it instead with inversion of an $N \times N$ matrix at cost $O(N^3)$. When $q \gg N$, the latter may be far smaller.

MEF: Calculating the measurement forecast moments in (11) requires $O(Nq^2)$ multiplications, which is smaller by a factor of q/p than the work required to obtain the representer matrix in EnKF. The matching and updating steps, (14) and (17), take place entirely in the space of $\frac{1}{2}q(q + 3)$ variables $(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$. Hence, these are relatively inexpensive when $q \ll p$. As discussed above, calculation of F_t and its gradients at one value of $(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ requires $O(Mq^3)$ multiplications. Hence, the total cost of the optimization in (14) by conjugate-gradient (CG) is $O(n_{CG}Mq^3)$, where n_{CG} is the number of iterations. For our convex cost function, CG is convergent globally but, generally, only linearly. Hence, it is better in practice to use a hybrid algorithm that switches to a superlinearly convergent, quasi-Newton method close to the solution, if storage of an approximate Hessian is affordable. The most expensive step of the algorithm, however, is the resampling. This can be accomplished, for example, using the standard sampling scheme (18) for the Gaussian components of the mixture model (15). As discussed above, calculating $\boldsymbol{\mu}_m(t; \boldsymbol{\lambda}, \boldsymbol{\Lambda})$ and $\mathbf{C}_m(t; \boldsymbol{\Lambda})$ for $m = 1, \ldots, M$ in (15) requires $O(Mpq)$ and $O(Mp^2q)$ operations, respectively, at each measurement time t. This is $(M/N)(p/q)$ times the cost to calculate the representer matrix in EnKF, with M/N small and p/q large. On the other hand, generating new samples by (18) requires Np random numbers and $O(Np^2)$ multiplications, always more expensive than the $O(Npq)$ multiplications for EnKF. Furthermore, a matrix square root (Cholesky factor or EOF's) of $\mathbf{C}_m(t; \boldsymbol{\Lambda})$ is required in (18), which costs $O(Mp^3)$ multiplications to calculate. This is very expensive for large p, too expensive in general to perform at each measurement time t.

If the model prior distribution $Q_M(\mathbf{x}, t)$ in (9) is time-independent (or varies sufficiently slowly in time), then there is the alternative sampling method discussed in Appendix C using a Metropolis-Hastings algorithm. In this scheme, proposals are generated from the Gaussian components in the mixture model for the prior (9). A number n_T of such trials are successively generated and accepted/rejected according to a Metropolis criterion, in order to produce each new ensemble member. An advantage of this approach is that one does not need to calculate the covariance matrices $\mathbf{C}_m(t; \boldsymbol{\Lambda})$ in (15) at all. One saves $O(Mp^2q)$ multiplications by avoiding the calculation of updated covariances. On the other hand, this alternative algorithm requires Npn_T random numbers and $O(Np^2n_T)$ multiplications to generate the new ensemble. The main savings lies in the fact that one needs only to calculate Cholesky factors or EOF's of the (time-independent) covariances $\mathbf{C}_m,\ m = 1, \ldots, M$ in (9) at the outset of the algorithm, a single-time cost of $O(Mp^3)$, rather than to calculate new matrix square roots at each measurement time. Since it will be true generally that $n_T \ll p$, this provides considerable economy when measurements are taken at many times.

Even with the most efficient implementations that we have been able to devise, this MEF algorithm is substantially more expensive than EnKF. The additional cost can only be justified by improved accuracy of the results. Substantial savings can be obtained by making some further approximations, for example, truncation of the K-L expansion (19) to a maximum number of EOF's $p_{\max} \ll p$. Finding just the $p_{\max}$ leading eigenvalues and eigenvectors of $\mathbf{C}_m$ for $m = 1, \ldots, M$ requires $O(Mp^2 p_{\max})$ operations, e.g. by iterative Arnoldi methods. This is smaller by a factor of $p_{\max}/p$ than the cost to determine all of the eigenvalues and eigenvectors.

Likewise, Metropolis sampling from the truncated K-L expansion uses $Np_{\max}n_T$ random numbers and $O(Npp_{\max}n_T)$ multiplications, smaller by the factor $p_{\max}/p$.

As with EnKF, alternative implementations of MEF can and should be explored, for which costs will scale differently than in the algorithms sketched above. We have only presented the most obvious numerical approaches and the costs that they entail.

MFF: The number of operations to calculate the function inside the brackets in (29) and its gradient in (31) is $O(Mq^2)$. Hence, the total cost of the matching step is $O(n_{CG}Mq^2)$ when using a conjugate-gradient algorithm. This is smaller by a factor of $1/q$ than the cost of the matching for full MEF and smaller by a factor $O(n_{CG}(M/N)(q/p))$ than the cost to calculate the representer matrix in EnKF. The resampling step in the mean-field MEF uses $O(Mpq)$ multiplications to calculate the means $\boldsymbol{\mu}_m(t;\boldsymbol{\lambda})$, $m = 1,\ldots,M$ in (15) [now depending only on $\boldsymbol{\lambda}$]. As in MEF with the Metropolis-Hastings sampling, there is a one-time expense of $O(Mp^3)$ to calculate square roots of the time-independent covariances $\mathbf{C}_m,\ m = 1,\ldots,M$. Also, Np random numbers and $O(Np^2)$ multiplications are needed to generate new samples by (18). Thus, resampling in MFF is cheaper than in full MEF by a factor of $1/n_T$ and more expensive than in EnKF by a factor of p/q. However, if a truncated K-L expansion is used with only $p_{\max}$ terms, as discussed above for MEF, then this latter factor is instead $p_{\max}/q$ and the cost will be similar as for EnKF if $p_{\max} \approx q$. In that case, MFF will be much cheaper overall than EnKF, the savings being that it avoids calculation of matrices such as the Kalman gain or the representer matrix.

REFERENCES

1. R. V. Abramov and A. J. Majda, Quantifying uncertainty for non-gaussian ensembles in complex systems. *Siam J. Set. Comp.* **26**:411–447 (2004).
2. D. L. Alspach and H. W. Sorenson, Nonlinear bayesian estimation using gaussian sum approximation. *IEEE Trans. Auto. Cont.* **17**:439–448 (1972).
3. J. L. Anderson, An ensemble adjustment filter for data assimilation. *Mon. Wea. Rev.* **129**:2884–2903 (2001).
4. J. L. Anderson and S. L. Anderson, A monte-carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.* **127**:2741–2758 (1999).
5. T. Bengtsson, C. Snyder, and D. Nychka, Toward a nonlinear ensemble filter for high-dimensional systems. *J. Geophys. Res.* **108**:8775 (2003).
6. A. F. Bennett, *Inverse Methods in Physical Oceanography* (Cambridge University Press, Cambridge, 1992).
7. G. Burgers, P. J. van Leeuwen, and G. Evensen, Analysis scheme in the ensemble kalman filter. *Mon. Wea. Rev.* **126**:1719–1724 (1998).
8. P. Cessi and W. R. Young, Multiple equilibria in two-dimensional thermo-haline circulation. *J. Fluid Mech.* **241**:291–309 (1992).
9. T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, 1991).
10. D. Crisan and A. Doucet, A survey of convergence results on particle filtering for practitioners. *IEEE Trans. Signal Process.* **50**:736–746 (2002).

11. L. Devroye, The double kernel method in density estimation. *Ann. Inst. H. Poincarè* **25**:533–580 (1989).
12. A. Doucet, N. de Freitas, and N. Gordon (eds.), *Sequential Monte Carlo Pract* (Springer-Verlag, 2001).
13. A. W. F. Edwards, *Likelihood* (The Johns Hopkins University Press, Baltimore, MD, 1992).
14. M. Ehrendorfer, The liouville equation and its potential usefulness for the prediction of forecast skill, part i: Theory. *Mon. Wea. Rev.* **122**:703–713 (1994).
15. R. Ellis, *Entropy, Large Deviations, and Statistical Mechanics* (Springer-Verlag, New York, 1985).
16. G. Evensen, Inverse methods and data assimilation in nonlinear ocean modles. *Physica D* **77**:108–129, (1994a).
17. G. Evensen, Sequential data assimilation with a nonlinear quasigeostrophic model using monte carlo methods to forecast error statistics. *J. Geophys. Res.* 99 (C5):10,143–10,162, (1994b).
18. G. Evensen, Sampling strategies and square-root analysis schemes for the enkf. *Ocean Model.* **54**:539–560 (2004).
19. G. Eyink, Turbulence noise. *J. Stat. Phys.* **83**:955–1019 (1996).
20. G. L. Eyink, Statistical hydrodynamics of the thermohaline circulation in a two-dimensional model. *Tellus A* **57**:100–115 (2005).
21. G. L. Eyink, J. M. Restrepo, and F. J. Alexander, A mean field approximation in data assimilation for nonlinear dynamics. *Physica D* **95**:347–368 (2004).
22. D. Fox, Adapting the sample size in particle filters through kid-sampling. *Int. J. Robot. Res.* **12**:985–1003 (2003).
23. M. Frontini and A. Tagliani, Maximum entropy in the finite stieltjes and hamburger moment problem. *J. Math. Phys.* **35**:6748–6756 (1994).
24. A. Gelb, *Applied Optimal Estimation* (The MIT Press, Cambridge, MA, 1974).
25. M. Ghil and K. Ide, Data assimilation in meteorology and oceanography: Theory and practice. *J. Meteor. Soc. Japan* **75**:111–496 (1997).
26. A. Golan, G. Judge, and D. Miller, *Maximum Entropy Econometrics*: *Robust Estimation with Limited Data* (John Wiley & Sons, New York, 1996).
27. N. J. Gordon, D. J. Salmond, and A. F. M. Smith, Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proc.-F* **140**:107–113 (1993).
28. A. Hannachi and A. O'Neill, Atmospheric multiple equilibria and non-gaussian behaviour in model simulations. *Quart. J. R. Meteor. Soc.* **127**:939–958 (2001).
29. K. Haven, A. J. Majda, and R. V. Abramov, Quantifying predictability through information theory: small sample estimation in a non-gaussian framework. *J. Comp. Phys.* **206**:334–362 (2005).
30. P. Hanggi, P. Talkner, and M. Borkovec, Reaction-rate theory: fifty years after kramers. *Rev. Mod. Phys.* **62**(2):251–341 (1990).
31. P. L. Houtekamer and H. L. Mitchell, A sequential ensemble kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.* **129**:123–137 (2001).
32. L. F. James, C. E. Priebe, and D. J. Marchette, Consistent estimation of mixture complexity. *Ann. Stat.* **29**(5):1281–1296 (2001).
33. A. H. Jazwinski, *Stochastic Processes and Filtering Theory* (Academic Press, NY, 1970).
34. C. L. Keppenne, Data assimilation into a primitive-equation model with a parallel ensemble kalman filter. *Mon. Wea. Rev.* **128**:1971–1981 (2000).
35. S. Kim, G. L. Eyink, J. M. Restrepo, F. J. Alexander, and G. Johnson, Ensemble filtering for nonlinear dynamics. *Mon. Wea. Rev.* **131**:2586–2594 (2003).
36. G. Kitagawa and W. Gersch, *Smoothness Priors Analysis of Time Series*, vol. 116 of *Lecture Notes in Statistics* (Springer-Verlag, 1996).
37. R. Kleeman, Measuring dynamical prediction utility using relative entropy. *J. Atmos. Sci.* **59**:2057–2072 (2002).

38. P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations* (Springer Verlag, 1997).
39. S. Kullback and R. A. Leibler, On information and sufficiency. *Ann. Math. Stat.* **22**:7986 (1951).
40. K. Lindenberg, B. J. West, and J. Kottalam, Fluctuations and dissipation in problems of geophysical fluid dynamics. In *Irreversible Phenomena and Dynamical Systems Analysis in Geo-sciences*, C. Nicolis and G. Nicolis (eds.) (NATO ASI, Series C. D. Reidel Publ. Co., 1987).
41. M. Loève, *Probability Theory*, 3rd ed. (Van Nostrand, New York, 1963).
42. E. N. Lorenz, Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**:130–141 (1963).
43. G. McLachlan and D. Peel, *Finite Mixture Models* (John Wiley & Sons, New York, 2000).
44. N. Metropolis and S. Ulam, The monte carlo method. *J. Amer. Stat. Assoc.* **44**:335–341 (1949).
45. P. Del Moral, Nonlinear filtering: Interacting particle solution. *Markov Proc. Rel. Fields.* **2**:555–579 (1996).
46. P. Del Moral, Nonlinear filtering using random particles. *Theor. Probab. Appl.* **40**:690–701 (1996).
47. P. Del Moral, J. Jacod, and Ph. Protter, The monte carlo method for filtering with discrete-time observations. *Probab. Theory. Rel.* **120**:346–368 (2001).
48. J. Namias, The index cycle and its role in the general circulation. *J. Meteorol.* **7**:130–139 (1950).
49. J. Nocedal and S. J. Wright, *Numerical Optimization* (Springer Series in Operations Research. Springer-Verlag, 1999).
50. Y. Pawitan, *In All Likelihood*: *Statistical Modelling and Inference Using Likelihood* (Oxford University Press, Oxford, UK, 2001).
51. C. E. Priebe, Adaptive mixtures. *J. Amer. Stat. Assoc.* **89**:796–806 (1994).
52. C. E. Priebe and D. J. Marchette, Alternating kernel and mixture density estimates. *Comput. Stat. Data. An.* **35**:43–65 (2000).
53. D. Rex, Blocking action in the middle troposphere and its effect upon regional climate. *Tellus* **2**:196–211 (1950).
54. H. Risken, *The Fokker-Planck Equation* (Springer-Verlag, New York, 1984).
55. F. Schoegl, Fluctuations in thermodynamic non-equilibrium states. *Z. Physik.* **244**:199–205 (1971).
56. F. Schoegl, On stability of steady states. *Z. Physik* **243**:303–310 (1971).
57. D. W. Scott, *Multivariate Density Estimation* (John Wiley, New York, 1992).
58. B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, New York, 1986).
59. P. Smyth, Model selection for probabilistic clustering using cross-validated likelihood. *Stat. Comput.* **10**:63–72 (2000).
60. P. Smyth, K. Ide, and M. Ghil, Multiple regimes in northern hemisphere height fields via mixture model clustering. *J. Atmos. Sci.* **56**:3704–3723 (1999).
61. D. B. Stephenson, A. Hannachi, and A. O Neill, On the existence of multiple climate regimes. *Q. J. R. Meteor. Soc.* **130**:583–605 (2004).
62. H. Stommel, Thermohaline convection with two stable regimes of flow. *Tellus* **13**:224–230 (1961).
63. A. Tarantola, *Inverse Problem Theory* (Elsevier Science, 1987).
64. H. Theil and D. G. Fiebig, *Exploiting Continuity*: *Maximum Entropy Estimation of Continuous Distributions* (Ballinger Publishing Co., Cambridge, MA, 1984).
65. W. Tucker, A rigorous ode solver and smale's 14th problem. *Found. Comput. Math.* **2**:53–117 (2002).
66. P. J. van Leeuwen and G. Evensen, Data assimilation and inverse methods in terms of a probabilistic formulation. *Mon. Wea. Rev.* **124**:2898–2913 (1996).
67. M. P. Wand and M. C. Jones, *Kernel Smoothing* (Chapman and Hall, London, 1995).
68. L.-S. Young, What are srb measures and which dynamical systems have them? *J. Stat. Phys.* **108**:733–754 (2002).