A Maximum Entropy Method for Particle Filtering

Gregory L. Eyink* and Sangil Kim†

Abstract

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.

**corresponding author:* Department of Applied Mathematics and Statistics, The Johns Hopkins University, Baltimore, MD 21218 U.S.A. Email: eyink@ams.jhu.edu, Tel:410-516-7201, Fax: 410-516-7459

†Department of Mathematics, University of Arizona, Tucson, AZ 85721 U.S.A.

running head: *Maximum-Entropy 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 $P(\mathbf{x}_0, ..., \mathbf{x}_t | \mathbf{y}_1, ..., \mathbf{y}_t)$, given the observations up to that time. Such posterior probabilities can be obtained 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 $P(\mathbf{x}_t | \mathbf{y}_1, ..., \mathbf{y}_t)$. 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 Ulam and Von Neumann (37), but computing resources available at the time did not lend to their widespread use. However, following the seminal paper of Gordon, Salmon and Smith (15), particle filtering methods have attracted strong general interest. Several recent books (19; 6) and review articles (4) 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 (25; 28) or (4) 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 only as many as $N = 100$ 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 (14). 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 use some prior knowledge about the statistics of the system in order to choose most judiciously the members of the small ensemble.

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(8; 9). 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. Therefore, there is motivation to generalize this approach in order 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 (24). 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.

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. 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 (3; 20). 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 (18).

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 prior 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 (19; 6)). 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 can generally be assumed that $q < p$. The signal process is a Markov process with initial distribution $\mathcal{P}_0(d\mathbf{x}_0)$ and transition probability $\mathcal{P}_{t+1|t}(d\mathbf{x}_{t+1}|\mathbf{x}_t)$. We shall usually assume that these probability measures have densities $P_0(\mathbf{x}_0)$ and $P_{t+1|t}(\mathbf{x}_{t+1}|\mathbf{x}_t)$ 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

$$\mathbf{x}_{t+1} = \mathbf{f}_t(\mathbf{x}_t, \boldsymbol{\eta}_t), \quad t = 0, 1, ..., T \tag{1}$$

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

$$P_{t+1|t}(\mathbf{x}_{t+1}|\mathbf{x}_t) = \int \Pi_t(d\boldsymbol{\eta}_t)\delta^p(\mathbf{x}_{t+1} - \mathbf{f}_t(\mathbf{x}_t, \boldsymbol{\eta}_t)). \tag{2}$$

A special case of (1) of great practical importance is that when the equation is deterministic, i.e. the distribution $\Pi_t(d\boldsymbol{\eta}_t)$ is a delta-measure and $\boldsymbol{\eta}_t$ 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 $\mathcal{G}_t(\mathbf{y}_t \in A|\mathbf{x}_t) = \int_A d^q\mathbf{y}_t \; G_t(\mathbf{y}_t|\mathbf{x}_t)$. A simple example is provided by the following measurement model

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

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 $R_t(\boldsymbol{\epsilon}_t)$. In that case,

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

This framework includes the case that measurements are made only at a subset of times $\mathcal{T}_M = \{t_m, \; m = 1, ..., M\}$ (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 $P(\mathbf{x}_t|\mathbf{y}_1, ..., \mathbf{y}_t)$. These may be obtained by a standard recursive application of Bayes' Theorem. To state the algorithm, we introduce the following notation for the filter densities before and after measurements:

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

We also use the convention that $P(\mathbf{x}, 0^+) = P_0(\mathbf{x})$. 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 $t = 1, ..., T$.

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* $G_{1:T}(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T)$ 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 $\mathcal{N}_t = G_t(\mathbf{y}_t|\mathbf{y}_1, ..., \mathbf{y}_{t-1})$ which, taken together, yield

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

This is called the *innovation form* of the likelihood function (19).

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) 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 Priors

The first step in our construction of an appropriate parametric model is to develop a representation for the prior distribution of the stochastic process in the absence of any measurements. We shall denote by $Q(\mathbf{x}, t)$ this *a priori* distribution. Its importance is due to the fact that, at long times between measurements and for sufficiently mixing Markov processes, the posterior distribution $P(\mathbf{x}, t)$ is expected to relax back to this prior. That is, the memory of information gained from observations is expected to fade between measurements. In general, the prior 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.

Within our general scheme, various approaches may be followed for modeling the prior distribution $Q(\mathbf{x}, t)$. We shall consider here only one possibility, the use of *mixture models*. In this approach, the model of the prior distribution is

taken to be a convex combination of a finite number of normal distributions. In other words, the model density is of the form

$$Q_M(\mathbf{x},t) = \sum_{m=1}^{M} w_m(t) N(\mathbf{x}; \boldsymbol{\mu}_m(t), \mathbf{C}_m(t)) \tag{8}$$

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 (24) for a comprehensive, current introduction to the literature. Methodologies for consistent estimation of a mixing distribution are discussed e.g. in (17) and its references. We obtain the weights and components in the mixtures for our numerical examples, discussed later, by conditional averaging of a large ensemble of realizations of the signal process solving (1), for carefully chosen conditions that characterize the components. Thus, the components of the mixture model represent different "regimes" of the system. A mixture model density can be constructed to converge to the density of the true prior distribution by increasing the mixture complexity M. Methods to estimate mixture complexity are discussed in more detail in (24; 29; 30).

There are a number of practical advantages of mixture models for our purposes. One of these is that they are relatively easy to sample, by simply selecting among the M components with probabilities $w_m,\ m = 1, ..., M$ and then sampling 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 may also be used for constructing realizations of random vectors $\mathbf{x}$ chosen from the distribution $N(\boldsymbol{\mu}_m, \mathbf{C}_m)$. For example, one may take

$$\mathbf{x} = \boldsymbol{\mu}_m + \mathbf{S}_m \cdot \boldsymbol{\xi} \tag{9}$$

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 $\mathbf{D}_m = \mathrm{diag}(\gamma_m^1, ..., \gamma_m^p)$ is the diagonal matrix of eigenvalues of $\mathbf{C}_m$ and $\mathbf{O}_m = [\hat{\mathbf{e}}_m^1, ..., \hat{\mathbf{e}}_m^p]$ is the orthogonal matrix whose columns are the orthonormal set of eigenvectors. In that case,

$$\mathbf{x} = \boldsymbol{\mu}_m + \sum_{a=1}^{p} \xi_a \sqrt{\gamma_m^a}\ \hat{\mathbf{e}}_m^a, \tag{10}$$

where $\xi_a,\ a = 1, ..., p$ 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 (10) is the

Karhunen-Loève (K-L) representation of Gaussian random vector $\mathbf{x}$; see (22). In addition to the ease in sampling, mixture models possess other advantages, which appear in the next subsection.

3.2 Maximum-Entropy Distributions

We now consider the problem of modeling the filter or posterior density $P(\mathbf{x}, t)$, given a model of the prior 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 its prior $Q(\mathbf{x}, t)$. A measure of this is the *relative entropy* or *Kullback-Leibler distance*, defined as

$$H(P(t)|Q(t)) = \int d\mathbf{x}\ P(\mathbf{x}, t) \ln\left(\frac{P(\mathbf{x}, t)}{Q(\mathbf{x}, t)}\right). \tag{11}$$

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 (33; 32) and (3), 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 posterior 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 denote as $P(\mathbf{x}, t^-)$ the filter distribution just before those measurements. The moments in that distribution of the measured variable,

$$\boldsymbol{\eta}_{t^-} = \langle \mathbf{h}_t \rangle_{t^-},\ \mathbf{H}_{t^-} = \langle \mathbf{h}_t \mathbf{h}_t^\top \rangle_{t^-}, \tag{12}$$

represent the *measurement forecast* at the time t, both the mean $\boldsymbol{\eta}_{t^-}$ and the covariance matrix $\mathbf{C}^H_{t^-} = \mathbf{H}_{t^-} - \boldsymbol{\eta}_{t^-}\boldsymbol{\eta}_{t^-}^\top$. 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. However, it would entail prohibitive computational costs to enforce those when the dimension p of the state space is very large.

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 (11) with the moments (12) 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:

$$P(\mathbf{x},t;\boldsymbol{\lambda},\boldsymbol{\Lambda}) = \frac{\exp[\boldsymbol{\lambda}\cdot\mathbf{h}_t(\mathbf{x}) + \frac{1}{2}\boldsymbol{\Lambda}:\mathbf{h}_t(\mathbf{x})\mathbf{h}_t^\top(\mathbf{x})]}{Z_t(\boldsymbol{\lambda},\boldsymbol{\Lambda})} Q(\mathbf{x},t). \tag{13}$$

Note that $Z_t(\boldsymbol{\lambda},\boldsymbol{\Lambda})$ is a normalization factor to ensure that (13) 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 (12) are obtained by taking derivatives, as follows:

$$\eta_i = \frac{\partial F_t}{\partial \lambda_i}(\boldsymbol{\lambda},\boldsymbol{\Lambda}),\;\; H_{ij} = \frac{\partial F_t}{\partial \Lambda_{ij}}(\boldsymbol{\lambda},\boldsymbol{\Lambda}),\;\; i \neq j,\;\; \frac{1}{2}H_{ii} = \frac{\partial F_t}{\partial \Lambda_{ii}}(\boldsymbol{\lambda},\boldsymbol{\Lambda}), i = j \tag{14}$$

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

$$H_t(\boldsymbol{\eta},\mathbf{H}) = \sup_{\boldsymbol{\lambda},\boldsymbol{\Lambda}} \{\boldsymbol{\eta}\cdot\boldsymbol{\lambda} + \frac{1}{2}\mathbf{H}:\boldsymbol{\Lambda} - F_t(\boldsymbol{\lambda},\boldsymbol{\Lambda})\} \tag{15}$$

which gives the relative entropy for the model density (13). Numerically, this optimization may be carried out by efficient algorithms due to the convexity of the function F_t. In our work below, we shall employ the 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 $q \ll p$. The maximum-entropy approximation could in principle be systematized by considering sequences of moment-constraints involving polynomials $x_{i_i}, x_{i_1}x_{i_2}, ..., x_{i_1}\cdots x_{i_n}$ 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. (12)). 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 $p \gg 1$.

Particular simplifications in the maximum-entropy formalism occur when the model prior density $Q_M(\mathbf{x},t)$ is a mixture of Gaussians, as in (8), 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 $t = 1, ..., T$, as in EnKF. In that case, as shown in Appendix B, , the cumulant-generating function $F_t(\boldsymbol{\lambda},\boldsymbol{\Lambda}) = \log Z_t(\boldsymbol{\lambda},\boldsymbol{\Lambda})$ can be calculated explicitly. The domain of this convex function, $\mathrm{dom}(F_t)$, 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 optimization in (15) must ensure that iterates stay within the feasible region $\mathrm{dom}(F_t)$. As shown in Appendix B, inside the domain it is possible to calculate exactly the gradients of F_t in (14), which can be

used in minimization by descent algorithms. In our experiments below, we shall use the CG algorithm with a feasible Armijo line-search, so that the iterates never go outside of $\mathrm{dom}(F_t)$ (27). 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.

Another advantage of the mixture model in (8) when the measurement function is affine is that the maximum-entropy densities (13) are also mixture models:

$$P_M(\mathbf{x}, t; \boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \sum_{m=1}^{M} w_m(t; \boldsymbol{\lambda}, \boldsymbol{\Lambda}) N(\mathbf{x}; \boldsymbol{\mu}_m(t; \boldsymbol{\lambda}, \boldsymbol{\Lambda}), \mathbf{C}_m(t; \boldsymbol{\Lambda})), \tag{16}$$

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.

3.3 The Maximum Entropy Filter

We can now outline the basic steps in the Maximum Entropy Filter (MEF) method. Between measurements, the particles $\mathbf{x}^{(n)}(t),\ n = 1, ..., N$ 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 practical implementation of Bayes theorem in the MEF method:

(i) Matching: The moments $\boldsymbol{\eta}_{t^-}, \mathbf{H}_{t^-}$ in (12) are determined by averaging over the N-particle ensemble $\mathbf{x}_{t^-}^{(n)},\ n = 1, ..., N$. A maximum-entropy density (13) is matched to these forecast statistics, with fitting parameters $(\boldsymbol{\lambda}_{t^-}, \boldsymbol{\Lambda}_{t^-})$ determined from the optimization in (15).

(ii) Updating: Bayes theorem is now applied, which, for normal error statistics, yields another maximum-entropy distribution (13) with parameters $(\boldsymbol{\lambda}_{t^+}, \boldsymbol{\Lambda}_{t^+})$ given by

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

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

(iii) Resampling: A new N-sample ensemble $\mathbf{x}_{t^+}^{(n)},\ n = 1, ..., N$ is created, by sampling from the model posterior $P(\mathbf{x}, t^+)$, the maximum-entropy distribution (13) with updated parameters $(\boldsymbol{\lambda}_{t^+}, \boldsymbol{\Lambda}_{t^+})$.

When the prior distribution is represented by a mixture model $Q_M(\mathbf{x}, t)$ and the measurement function is affine, then the matching step *(i)* can be carried out by a feasible CG method, as discussed earlier. The resampling step *(iii)* can also be carried out efficiently using the mixture representation (16), at least 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 prior distribution $Q_M(\mathbf{x}, t)$ consists of a single Gaussian component and if the model posterior distribution $P_M(\mathbf{x}, t)$ is a maximum-entropy distribution constrained by the full state statistics, $\boldsymbol{\mu}_{t^-} = \langle \mathbf{x} \rangle_{t^-}$, $\mathbf{M}_{t^-} = \langle \mathbf{x}\mathbf{x}^\top \rangle_{t^-}$, of second order, then the MEF method is equivalent to the Ensemble Kalman Filter (EnKF) (9; 38; 2) (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 (13) to calculate the normalization $\mathcal{N}_t$ as

$$\log \mathcal{N}_t = \Delta F_t - \frac{1}{2}\mathbf{y}_t^\top \mathbf{R}_t^{-1}\mathbf{y}_t - \frac{1}{2}\log[(2\pi)^q \mathrm{Det}\mathbf{R}_t], \tag{18}$$

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. meteorological weather forecasting—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 computational cost $O(Mq^3)$, growing rapidly with q (see Appendix D). Even EnKF requires $O(p^2 q)$ multiplications in order to compute the Kalman gain matrix, and this will also not be practical when $p \gg q \gg 1$. To deal with such cases, 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). Both the matching step *(i)* and the update step *(ii)* are now 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 equation (12), i.e. $\boldsymbol{\eta}_{t^-} = \langle \mathbf{h}_t \rangle_{t^-}$, as constraints. This density is a member of the exponential family

$$P(\mathbf{x}, t; \boldsymbol{\lambda}) = \frac{1}{Z_t(\boldsymbol{\lambda})} \exp[\boldsymbol{\lambda} \cdot \mathbf{h}_t(\mathbf{x})] \cdot Q(\mathbf{x}, t) \tag{19}$$

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

$$H_t(\boldsymbol{\eta}) = \sup_{\boldsymbol{\lambda}} \{\boldsymbol{\eta} \cdot \boldsymbol{\lambda} - F_t(\boldsymbol{\lambda})\} \tag{20}$$

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

$$\boldsymbol{\eta}_{t^+} = \operatorname*{arginf}_{\boldsymbol{\eta}} \left\{ H_t(\boldsymbol{\eta}|\boldsymbol{\eta}_{t^-}) + \frac{1}{2}[\boldsymbol{\eta} - \mathbf{y}_t]^\top \mathbf{R}_t^{-1} [\boldsymbol{\eta} - \mathbf{y}_t] \right\} \tag{21}$$

where

$$H_t(\boldsymbol{\eta}|\boldsymbol{\eta}_{t^-}) = H_t(\boldsymbol{\eta}) - H_t(\boldsymbol{\eta}_{t^-}) - (\boldsymbol{\eta} - \boldsymbol{\eta}_{t^-}) \cdot \boldsymbol{\lambda}_{t^-}. \tag{22}$$

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 (21).

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 (19) for $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{t^+}$ can be sampled using its representation by a mixture model [equation (16) with $\boldsymbol{\Lambda}$ set to zero].

The meaning of the new update procedure is best seen from the significance of the entropy function (22) in large deviations theory (7). If N independent samples $\mathbf{x}_{t-}^{(n)},\ n = 1, ..., N$ are drawn from the model distribution $P(\mathbf{x}, t; \boldsymbol{\lambda}_{t^-})$, the large-deviations result is, roughly speaking, that

$$\text{Prob}\left\{ \frac{1}{N} \sum_{n=1}^{N} \mathbf{h}_t(\mathbf{x}_{t^-}^{(n)}) \approx \boldsymbol{\eta} \right\} \sim \exp[-N \cdot H_t(\boldsymbol{\eta}|\boldsymbol{\eta}_{t^-})] \tag{23}$$

as $N \to \infty$, with $H_t(\boldsymbol{\eta}|\boldsymbol{\eta}_{t^-})$ as in (22). We can also take an i.i.d. set $\{\boldsymbol{\epsilon}_t^{(n)},\ n = 1, ..., N\}$ of $N(\mathbf{0}, \mathbf{R}_t)$ random variables, representing observation errors, and define the ensemble of measured values $\mathbf{y}_t^{(n)} = \mathbf{h}_t(\mathbf{x}_{t^-}^{(n)}) + \boldsymbol{\epsilon}_t^{(n)},\ n = 1, ..., N$. Then a large deviations result holds also for the joint probability as $N \to \infty$

$$\text{Prob}\left\{ \frac{1}{N} \sum_{n=1}^{N} \mathbf{h}_t(\mathbf{x}_{t^-}^{(n)}) \approx \boldsymbol{\eta}, \frac{1}{N} \sum_{n=1}^{N} \mathbf{y}_t^{(n)} \approx \mathbf{y} \right\} \sim \exp[-N \cdot H_t(\boldsymbol{\eta}, \mathbf{y}|\boldsymbol{\eta}_{t^-})] \tag{24}$$

where the joint-entropy $H_t(\boldsymbol{\eta}, \mathbf{y}|\boldsymbol{\eta}_{t^-})$ is the function in curly brackets in (21). It follows that the value $\boldsymbol{\eta}_{t^+}$ defined in (21) 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 $n = 1, ..., N$. 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 more discussion of this mean-field approximation, see (10).

There is also a natural mean-field analogue of the log-likelihood. It follows directly from (24) by a steepest descent result (contraction principle) that

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

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 (21). From (25) it is reasonable to define

$$\ln \mathcal{N}_t = -H_t^Y(\mathbf{y}_t|\boldsymbol{\eta}_{t^-}) \tag{26}$$

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 (26) 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. equation (35) below). Because the large-deviations result (25) has only logarithmic accuracy, one should expect to miss such constant terms. This does not detract necessarily from the utility of (26) to make maximum-likelihood estimates of parameters for distinct sequences $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T$ of observations.

In a practical implementation of the MFF method, one can avoid the calculation of $\boldsymbol{\eta}_{t^+}$ in (21). Instead, one can calculate $\boldsymbol{\lambda}_{t^+}, H_t^Y(\mathbf{y}_t|\boldsymbol{\eta}_{t^-})$ directly by combining (20) and (21) into a single optimization:

$$\begin{aligned}\boldsymbol{\lambda}_{t^+} = \ \underset{\boldsymbol{\lambda}}{\text{arginf}}\Bigg\{ & \boldsymbol{\eta}_t(\boldsymbol{\lambda}) \cdot (\boldsymbol{\lambda} - \boldsymbol{\lambda}_{t^-}) - F_t(\boldsymbol{\lambda}) + F_t(\boldsymbol{\lambda}_{t-}) \\ & + \frac{[\boldsymbol{\eta}_t(\boldsymbol{\lambda}) - \mathbf{y}_t]^\top \mathbf{R}_t^{-1}[\boldsymbol{\eta}_t(\boldsymbol{\lambda}) - \mathbf{y}_t]}{2}\Bigg\}\end{aligned} \tag{27}$$

with also

$$\begin{aligned}H_t^Y(\mathbf{y}_t|\boldsymbol{\eta}_{t^-}) = \ \inf_{\boldsymbol{\lambda}}\Bigg\{ & \boldsymbol{\eta}_t(\boldsymbol{\lambda}) \cdot (\boldsymbol{\lambda} - \boldsymbol{\lambda}_{t^-}) - F_t(\boldsymbol{\lambda}) \\ & + F_t(\boldsymbol{\lambda}_{t-}) + \frac{[\boldsymbol{\eta}_t(\boldsymbol{\lambda}) - \mathbf{y}_t]^\top \mathbf{R}_t^{-1}[\boldsymbol{\eta}_t(\boldsymbol{\lambda}) - \mathbf{y}_t]}{2}\Bigg\}\end{aligned} \tag{28}$$

where $\boldsymbol{\eta}_t(\boldsymbol{\lambda}) = \frac{\partial F_t}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda})$. To carry out the optimization in (27) 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 equation (27):

$$\frac{\partial^2 F_t}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}) \left\{ \boldsymbol{\lambda} - \boldsymbol{\lambda}_{t^-} + \mathbf{R}_t^{-1} [\boldsymbol{\eta}_t(\boldsymbol{\lambda}) - \mathbf{y}_t] \right\}. \tag{29}$$

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(23). 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 $\kappa > 0$

$$dx = f(x)dt + \kappa dW(t), \tag{30}$$

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. 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 $x(t)$ itself 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 (21)

$$\begin{aligned} x(t_{k+1}) &= x(t_k) + f(x(t_k))\Delta t + \kappa N_k \sqrt{\Delta t}, \\ t_{k+1} &= t_k + \Delta t \end{aligned} \tag{31}$$

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 (10).

In order to apply the MEF method, we must construct a model for the prior $Q(x,t)$. Here we shall assume that the initial condition x_0 is drawn from the invariant measure $P_*(x)$, so that the prior 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 $M = 2$. To construct the weights, means, and variances of the components, we compute a single realization for a long time and gather probabilities for the complementary events $\{\text{sign}(x) = \pm 1\}$, and means and variances conditioned on these two events. Then we take $w_\pm$ to be the probabilities, and $\mu_\pm, C_\pm$ to be the conditional means and variances. In practice, we symmetrize the numerical results so that $w_- = w_+ = 0.5, \mu_- = -\mu_+$, and $C_- = C_+$. Our mixture model is then

$$Q_M(x) = w_- N(x; \mu_-, C_-) + w_+ N(x; \mu_+, C_+). \tag{32}$$

By construction, (32) 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 (32) is a quite good approximation in this example.

4.1.1 Experiment A:

Our first estimation experiment is for model (30) 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:

$$\overline{\tau} \sim \frac{2\pi}{\sqrt{U''(1)|U''(0)|}} \exp\left(\frac{2\Delta U}{\kappa^2}\right), \qquad \Delta U = U(0) - U(1), \tag{33}$$

valid as $\kappa \to 0$ (31; 16). 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 $5 \sim 6$ 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 (13) to match the current particle ensemble $\{x_n(t^-) : \ n = 1, ..., N\}$ 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 (15). 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 (68),(70),(71), (72) 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 (15). This yields the parameters $\lambda_{t^-}, \Lambda_{t^-}$ that are then updated to $\lambda_{t^+}, \Lambda_{t^+}$ 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 $\{x_n(t^+) : \ n = 1, ..., N\}$. This is done by using the mixture representation (16) for $m = \pm 1$, with weights, means, and covariances given by (69),(62),(58), 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 $n = 1, ..., N$, 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

$$x_n(t^+) = \mu_{m_n}(\lambda_{t^+}, \Lambda_{t^+}) + \sqrt{C_{m_n}(\Lambda_{t^+})}\xi_n, \tag{34}$$

where ξ_n are i.i.d. $N(0, 1)$ random variables for $n = 1, ..., N$. (34) is the analogue

of (10) for our problem. This new set of samples is then evolved forward with the equation (30) 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 (19) is chosen to match the pre-measurement ensemble $\{x_n(t^-) : \ n = 1, ..., N\}$ 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 (20), where the "free-energy" function $F_t(\lambda)$ and its first derivative are calculated from the formulas (77),(78) 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 (27). 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 (79) 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 (74), (75),(76) in Appendix B. In particular, the variances C_m, $m = \pm 1$ are not changed at all in the update. The formula (34) is used finally, just as in MEF, in order to generate the new ensemble $\{x_n(t^+) : \ n = 1, ..., N\}$ of post-measurement 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 $N = 10^2$ 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 $t = 6 \sim 12$. 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 $t = 2 \sim 8$ indicating the state is near $x = +1$, all 100 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 100 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 $t = 8 \sim 10$. 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 (10), MFF in general follows observations too closely when the data lie near to the prior 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 $N = 10^4$ particles. It becomes clear now 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 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 very little changed from those for $N = 10^2$, except that fluctuations are smaller and the curves are 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 1 for the mean of the state variable $x(t)$ over the interval $0 < t < 20$:

For $N = 10^2$, WRF and EnKF results are poor, MFF reasonable, and MEF very good. 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 $G_{1:t}(\mathbf{y}_1, ..., \mathbf{y}_t)$ 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 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 (42). In EnKF the innovation may be

consistently taken to be a Gaussian

$$\mathcal{N}_t = \frac{\exp\left[-\frac{1}{2}(\mathbf{y}_t - \boldsymbol{\mu}_{t-}^Y)^\top (\mathbf{C}_{t-}^Y)^{-1}(\mathbf{y}_t - \boldsymbol{\mu}_{t-}^Y)\right]}{\sqrt{(2\pi)^q \mathrm{Det}(\mathbf{C}_{t-}^Y)}}, \tag{35}$$

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 (13). We have already discussed how to obtain the likelihoods—or, rather, their logarithms—in the entropy-based methods, MEF and MFF. In Figs. 4-5 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 (10). The results are quite consistent with those we saw in Figs. 2-3 for the means and standard deviations. For $N = 10^2$, the plots in Fig. 4 show that MEF is already very accurate 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. 5 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. 6-7 the relative entropy $H(t)$ as a function of time for both the MEF and MFF methods. For comparison, we show the exact relative entropy calculated by a discretization of the integral (11) using the Fokker-Planck solution from the scheme of (10). 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. 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. 6 we plot the entropies for $N = 10^2$ and in Fig. 7 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 (3). 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 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 (30) 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 equation (33), is now $\overline{\tau} \approx 65.8$. The time required to make a transition is also somewhat shorter, taking about $1 \sim 2$ time units, but the fraction of time spent in transitions is greatly increased, to about $0.01 \sim 0.03$. 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.

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. The results of the experiment for means and standard deviations are shown in Fig. 8 with $N = 10^2$ and in Fig. 9 with $N = 10^4$. There is little difference between the two sets of figures, except that the first is rougher and more random. Of the four methods, they may be rated in order as WRF, MEF, 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:

We next consider the log-likelihoods $L_{1:t}$ of the four particle filtering methods, presented versus time t as before. We consider in this Experiment B the results of the methods only with $N = 10^2$, since those with $N = 10^4$ do not differ substantially. 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 Figs. 8-9. However, the MFF approximation is reasonably good here. Of all the 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 Figs. 8-9 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 (35) that we adopted for EnKF. The true likelihoods are not normal distributions, as assumed in (35). 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 $\mathbf{y}_1, ..., \mathbf{y}_T$. 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. The results for $N = 10^2$ are given in Fig. 11 and for $N = 10^4$ in Fig. 12. The two sets of figures again are quite close, with just an increase of smoothness for larger N. EnKF and MFF are similar in their results, with both somewhat underpredicting the entropy. EnKF also violates the H-theorem by yielding occasionally an increasing relative entropy. MEF gives a quite good approximation to the exact entropy calculated from the Fokker-Planck solution.

4.2 Lorenz Model

Our last experiment will be for the chaotic 3-dimensional dynamical system of Lorenz(23), given by the differential equations:

$$\frac{dx}{dt} = \sigma(y - x), \ \frac{dy}{dt} = (r - z)x - y, \ \frac{dz}{dt} = xy - bz, \tag{36}$$

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 of the entropy-based methods, since the relative entropy of two solutions of the Liouville equation —for the probability densities in phase space— is conserved in time. In this sense, the H-theorem holds here in a trivial sense only. Furthermore, the invariant measure of the system lives on a strange attractor with fractal dimension about 2.06, the famous Lorenz butterfly attractor. Thus, the relative entropy of any measure absolutely continuous with respect to Lebesgue and the invariant measure is infinite. On the other hand, this system is generally quite similar in its behavior to the stochastic Double Well model considered in the last section, especially in Experiment B. 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 will be able to track the transitions with a relatively small number of samples, $N = 10^2$, say.

We shall compare the parametric 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* (5). 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

$$\delta_N = \underset{\delta}{\operatorname{argmin}} \left\{ \int_0^T dt \|\boldsymbol{\mu}_N(t; G, \delta) - \boldsymbol{\mu}_N(t; U, \delta)\| \right\} \tag{37}$$

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 (37) 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 equations (36) we must build a mixture model $Q_M(\mathbf{x}, t)$ for the prior 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 prior. 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 considering the complementary sets $\{\text{sign}\,(x + y) = \pm 1\}$, which each contain one wing of the attractor. We then consider a single long time-trajectory of the Lorenz system (36) for an initial condition on the attractor and extract from it the probabilities $w_\pm$ of the two events, and the conditional means $\boldsymbol{\mu}_\pm$ and

covariance matrices $\mathbf{C}_\pm$:

$$\boldsymbol{\mu}_\pm = (\bar{x}_\pm, \bar{y}_\pm, \bar{z}_\pm)^\top = (\pm 6.36389,\ \pm 6.69471602,\ 23.5506805)^\top \tag{38}$$

$$\mathbf{C}_\pm = \begin{pmatrix} 22.3056857 & 20.2011608 & \pm 24.9259341 \\ 20.2011608 & 36.3717702 & \pm 1.57754284 \\ \pm 24.9259341 & \pm 1.57754284 & 74.3283071 \end{pmatrix} \tag{39}$$

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 (38),(39). This construction guarantees that the model has the same second-order statistics (mean and covariance) as the exact invariant measure. In Fig. 13 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 equations (36) 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 every $\Delta T = \frac{2}{3}$ time units over the interval $0 < t < 16$ and then contaminated with Gaussian errors of mean zero $\mathbf{0}$ and covariance $\mathbf{R} = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$ for all measurement times. We seek the conditional statistics given this "observational" data.

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 (15) in the matching step of MEF depends upon 5 variables, while the function $F_t(\boldsymbol{\lambda})$ used in (20) 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 (15) 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 (16). 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 $\mathbf{C}_m(\Lambda_{t^+})$, $m = 1, ..., M$ at every measurement time t. Thus, we have employed instead the 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" (81), 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:

$$\mathbf{x}' = \boldsymbol{\mu}_m(\boldsymbol{\lambda}_{t^+}, \boldsymbol{\Lambda}_{t^+}) + \sum_{a=1}^{3} \xi_a \sqrt{\gamma_m^{(a)}} \hat{\mathbf{e}}_m^{(a)}, \;\; m = \pm. \tag{40}$$

Here $\boldsymbol{\mu}_m(\boldsymbol{\lambda}_{t^+}, \boldsymbol{\Lambda}_{t^+})$ is the vector given by (62) in Appendix B, ξ_a, $a = 1, 2, 3$ are i.i.d. N(0,1) random variables,

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

are the eigenvalues of the fixed matrices $\mathbf{C}_m$, $m = \pm$ in (39), and

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

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 (40) 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 (39). 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. 14-15 we illustrate the convergence of the WRF method. The optimization in (37) gives $\delta_N = 0.6$ for $N = 10^2$, and $\delta_N = 0.1$ for $N = 10^4$. The plots in Fig. 14(a)-(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. Fig. 15(a)-(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 $N = 10^4$

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. 14(c) and 15(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. 14(a) and 15(a) with $N = 10^2$, we had to scan over about 100 values of δ to find the approximate minimum in (37). This is essentially 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.

In Figs. 16 and 17 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 $\sigma_x(t)$. We will not show the results for $N = 10^4$, because they are almost identical to those for $N = 10^2$. We can thus assume that the results in Figs. 16-17 well represent the converged approximations of the methods as $N \to \infty$. This is supported by the data in Table 3, which gives the relative mean errors for both sample-sizes $N = 10^2$ and $N = 10^4$. 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 $10 \sim 20$ 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 $2 \sim 3$ 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 (36). We find that for the means MEF is somewhat better than EnKF, which is itself slightly better than MFF, but all three perform reasonably well. All three methods give standard deviations that are too large, but MFF much larger than the other two.

In Fig. 18 we plot the results for the log-likelihoods of the three methods EnKF, MEF, and MFF with $N = 10^2$, where the results of WRF with $N = 10^4$ are 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 missing prefactors in its estimate of $\ln \mathcal{N}_t$. We can make a crude estimate of the correction as $-\frac{1}{2} \ln[(2\pi)^q \mathrm{Det} \mathbf{C}_{t-}^Y]$, 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 approximations previously proposed for the relative entropy using the three methods, EnKF, MEF and MFF. These are plotted in Fig. 19(a)-(c) for each method with $N = 10^2$. As above, the results with $N = 10^4$ were so similar that they 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 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 (13) 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 (36; 40), 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 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 (11). The results are shown in Fig. 19(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 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), 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 but is also more expensive. If $q \gg 1$, then both MEF and EnKF may be too costly to apply and 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 $N = 10^2$ 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 prior 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 stochastic dynamical systems.

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.

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 $\mathbf{x}_t^{(n)},\ n = 1, ..., N$ of solutions of (1) with independent realizations of the noise in order to approximate the filter densities by empirical measures

$$P^{(N)}(\mathbf{x}, t) = \sum_{n=1}^{N} w_t^{(n)} \delta^p(\mathbf{x} - \mathbf{x}_t^{(n)}). \tag{41}$$

The non-negative real numbers $w_t^{(n)},\ n = 1, ..., N$ 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

$$w_{t^+}^{(n)} = \frac{G_t(\mathbf{y}_t|\mathbf{x}_t^{(n)})}{\mathcal{N}_t} w_{t^-}^{(n)},\quad n = 1, ..., N \tag{42}$$

which is determined so that the N-sample approximations (41) 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 $\mathbf{x}_0^{(n)},\ n = 1, ..., N$ 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 $n = 1, ..., N$. Initialized in this manner, the algorithm outlined above provides a systematic approach to approximating the filter distributions, via (41). For convenient reference, we shall call this simple standard particle method the Weighted Ensemble Filter (WEF). For more details, see (19; 6). It has been proved by Moral(26) 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 (42), 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 (42) in the above method may be augmented with a *resampling* step, as was originally suggested by Ulam and von Neumann (37). That is, a new ensemble

$\mathbf{x}_{t^+}^{(n)}$, $n = 1, ..., N$ 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 $n' = 1, ..., N$. 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 (41) may be improved with *kernel smoothing:*

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

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 (35; 34; 39). Resampling from a distribution like (43) may be accomplished in two steps: first, select an index $n' = 1, ..., N$ 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 $n = 1, ..., N$. 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 (43) will also converge to the true density $P(\mathbf{x}, t)$ as $N \to \infty$. We refer to standard texts (35; 34; 39) for more details.

The algorithm with resampling as described above is one of the most widely used particle filtering methods (e.g. see (6)), 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 $N \to \infty$ ((25; 28) or (4) 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 (11) 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 (8; 9) (with an important correction in Burgers et al.(2)). 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,

$$\mathbf{h}_t(\mathbf{x}) = \boldsymbol{\mathcal{H}}_t\mathbf{x} + \mathbf{d}_t, \tag{44}$$

where $\boldsymbol{\mathcal{H}}_t$ is a $q \times p$ matrix and $\mathbf{d}_t$ is a q-vector for each time t. Otherwise, Taylor expansions of $\mathbf{h}_t(\mathbf{x})$ and $\log G_t(\mathbf{y}|\mathbf{x})$ (to 1st and 2nd order, resp.) must be employed. As in all Kalman filtering schemes, the statistical basis of this method 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:

$$\boldsymbol{\mu}_{t^-} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_{t^-}^{(n)}, \qquad \mathbf{M}_{t^-} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_{t^-}^{(n)}[\mathbf{x}_{t^-}^{(n)}]^\top \tag{45}$$

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

$$\mathbf{y}_t^{(n)} = \mathbf{y}_t + \boldsymbol{\epsilon}_t^{(n)},\ n = 1, ..., N \tag{46}$$

where $\boldsymbol{\epsilon}_t^{(n)},\ n = 1, ..., N$ 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

$$\mathbf{x}_{t^+}^{(n)} = \mathbf{x}_{t^-}^{(n)} + \boldsymbol{\mathcal{K}}_t[\mathbf{y}_t^{(n)} - \mathbf{h}_t(\mathbf{x}_{t^-}^{(n)})], \tag{47}$$

with

$$\boldsymbol{\mathcal{K}}_t = \mathbf{C}_{t^-}\boldsymbol{\mathcal{H}}_t^\top[\boldsymbol{\mathcal{H}}_t\mathbf{C}_{t^-}\boldsymbol{\mathcal{H}}_t^\top + \mathbf{R}_t]^{-1}, \tag{48}$$

the *Kalman gain matrix*

An easy computation (2) shows that the mean and covariance of the updated ensemble are given in the limit $N \to \infty$ by

$$\boldsymbol{\mu}_{t^+} = \boldsymbol{\mu}_{t^-} + \boldsymbol{\mathcal{K}}_t[\mathbf{y}_t - \boldsymbol{\mu}_{t^-}^H] \tag{49}$$

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

where $\boldsymbol{\mu}_{t^-}^H = \boldsymbol{\mathcal{H}}_t\boldsymbol{\mu}_{t^-} + \mathbf{d}_t$ and $\mathbf{C}_{t^-}^H = \boldsymbol{\mathcal{H}}_t\mathbf{C}_{t^-}\boldsymbol{\mathcal{H}}_t^\top$ 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 (13). Notice, however, that the posterior density $P(\mathbf{x}, t^+)$ represented by the samples $\mathbf{x}_{t^+}^{(n)},\ n = 1, ..., N$ is non-Gaussian, because the original samples $\mathbf{x}_{t^-}^{(n)},\ n = 1, ..., N$ 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.

B Thermodynamics of Maximum-Entropy Models

In this appendix we derive the equations of a "thermodynamic formalism" for maximum-entropy mixture models. We assume that the measurement function is affine, $\mathbf{h}(\mathbf{x}) = \boldsymbol{\mathcal{H}}\mathbf{x} + \mathbf{d}$, and consider the mixture model (8), with

$$N(\mathbf{x}; \boldsymbol{\mu}_m, \mathbf{C}_m) = \frac{\exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_m)^\top \mathbf{C}_m^{-1}(\mathbf{x} - \boldsymbol{\mu}_m)\right]}{\sqrt{(2\pi)^p \mathrm{Det}\mathbf{C}_m}} \tag{51}$$

We first prove the following simple but useful lemma:

$$\begin{aligned} &\exp[\boldsymbol{\lambda}\cdot\mathbf{h}(\mathbf{x}) + \frac{1}{2}\boldsymbol{\Lambda}\mathbf{:}\mathbf{h}(\mathbf{x})\mathbf{h}^\top(\mathbf{x})]N(\mathbf{x}; \boldsymbol{\mu}_m, \mathbf{C}_m) \\ &\quad = Z_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})N(\mathbf{x}; \boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}), \mathbf{C}_m(\boldsymbol{\Lambda})) \end{aligned} \tag{52}$$

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 (52) is a product of Gaussians, the equality is proved easily by completing the square, with the results:

$$\mathbf{C}_m(\boldsymbol{\Lambda}) = (\mathbf{C}_m^{-1} - \boldsymbol{\mathcal{H}}^\top\boldsymbol{\Lambda}\boldsymbol{\mathcal{H}})^{-1} \tag{53}$$

$$\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \boldsymbol{\mu}_m + \mathbf{C}_m(\boldsymbol{\Lambda})\boldsymbol{\mathcal{H}}^\top(\boldsymbol{\lambda} + \boldsymbol{\Lambda}\boldsymbol{\mu}_m^H) \tag{54}$$

$$Z_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \sqrt{\frac{\mathrm{Det}\mathbf{C}_m(\boldsymbol{\Lambda})}{\mathrm{Det}\mathbf{C}_m}} \exp\left[\boldsymbol{\lambda}^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \boldsymbol{\Lambda}\mathbf{d} \right. \\ \left. -\frac{1}{2}\boldsymbol{\mu}_m^\top \mathbf{C}_m^{-1}\boldsymbol{\mu}_m + \frac{1}{2}\boldsymbol{\mu}_m^\top(\boldsymbol{\lambda}, \boldsymbol{\Lambda})\mathbf{C}_m^{-1}(\boldsymbol{\Lambda})\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})\right]. \tag{55}$$

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:

$$(\mathbf{A}^{-1} + \mathbf{B}^\top\mathbf{C}^{-1}\mathbf{B})^{-1} = \mathbf{A} - \mathbf{A}\mathbf{B}^\top(\mathbf{B}\mathbf{A}\mathbf{B}^\top + \mathbf{C})^{-1}\mathbf{B}\mathbf{A}, \tag{56}$$

$$(\mathbf{A}^{-1} + \mathbf{B}^\top\mathbf{C}^{-1}\mathbf{B})^{-1}\mathbf{B}^\top = \mathbf{A}\mathbf{B}^\top(\mathbf{B}\mathbf{A}\mathbf{B}^\top + \mathbf{C})^{-1}\mathbf{C}. \tag{57}$$

These identities are standard in the Kalman filtering literature (13). From (56) it follows immediately that

$$\mathbf{C}_m(\boldsymbol{\Lambda}) = \mathbf{C}_m + \mathbf{C}_m\boldsymbol{\mathcal{H}}^\top\boldsymbol{\Gamma}_m^H[\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}]^{-1}\boldsymbol{\Lambda}\boldsymbol{\mathcal{H}}\mathbf{C}_m, \tag{58}$$

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 (57) gives likewise

$$\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \boldsymbol{\mu}_m + \mathbf{C}_m\boldsymbol{\mathcal{H}}^\top\boldsymbol{\Gamma}_m^H[\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}]^{-1}(\boldsymbol{\lambda} + \boldsymbol{\Lambda}\boldsymbol{\mu}_m^H) \tag{59}$$

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 (1; 38). [In fact, our calculations here are a natural generalization of the representer solution for Gaussian mixture models; e.g. see (62) 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 (54) for $\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$. Cancelling many terms, one finds finally that

$$Z_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \sqrt{\frac{\mathrm{Det}\boldsymbol{\Gamma}_m^H}{\mathrm{Det}(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda})}} \exp\left[-\frac{1}{2}(\boldsymbol{\mu}_m^H)^\top\boldsymbol{\Gamma}_m^H\boldsymbol{\mu}_m^H \right. \\ \left. +\frac{1}{2}(\boldsymbol{\Gamma}_m^H\boldsymbol{\mu}_m^H + \boldsymbol{\lambda})^\top(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda})^{-1}(\boldsymbol{\Gamma}_m^H\boldsymbol{\mu}_m^H + \boldsymbol{\lambda})\right]. \tag{60}$$

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

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

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

$$(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}) \cdot \boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \boldsymbol{\Gamma}_m^H \boldsymbol{\mu}_m^H + \boldsymbol{\lambda}. \tag{61}$$

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 (60) must be a convex function of $\boldsymbol{\lambda}$ for realizability to hold; moreover, (53),(57) 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 (59) gives

$$\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \boldsymbol{\mu}_m + \mathbf{C}_m \boldsymbol{\mathcal{H}}^\top \boldsymbol{\Gamma}_m^H \cdot [\boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) - \boldsymbol{\mu}_m^H] \tag{62}$$

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 (58) and (60) for $\mathbf{C}_m(\boldsymbol{\Lambda})$ and $Z_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$, respectively. In fact, we may rewrite (60) somewhat to eliminate the inverse matrix:

$$\begin{aligned}
Z_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = &\sqrt{\frac{\mathrm{Det}\boldsymbol{\Gamma}_m^H}{\mathrm{Det}(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda})}} \times \\
&\exp\left[-\frac{1}{2}(\boldsymbol{\mu}_m^H)^\top \boldsymbol{\Gamma}_m^H \boldsymbol{\mu}_m^H + \frac{1}{2}(\boldsymbol{\Gamma}_m^H \boldsymbol{\mu}_m^H + \boldsymbol{\lambda})^\top \boldsymbol{\eta}_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda})\right].
\end{aligned} \tag{63}$$

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

$$\begin{aligned}
\ln Z_m(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) &= \frac{1}{2}(\boldsymbol{\Gamma}_m^H \boldsymbol{\mu}_m^H + \boldsymbol{\lambda})^\top (\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda})^{-1} (\boldsymbol{\Gamma}_m^H \boldsymbol{\mu}_m^H + \boldsymbol{\lambda}) \\
&- \frac{1}{2}(\boldsymbol{\mu}_m^H)^\top \boldsymbol{\Gamma}_m^H \boldsymbol{\mu}_m^H - \frac{1}{2}\mathrm{Tr}[\ln(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda}) - \ln \boldsymbol{\Gamma}_m^H]
\end{aligned} \tag{64}$$

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 $\frac{\partial}{\partial\lambda}\mathrm{Tr}\ln\mathbf{A} = \mathrm{Tr}\left(\mathbf{A}^{-1}\frac{\partial\mathbf{A}}{\partial\lambda}\right)$. Then

$$\frac{\partial Z_m}{\partial\boldsymbol{\lambda}}(\boldsymbol{\lambda},\boldsymbol{\Lambda}) = Z_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})\boldsymbol{\eta}_m(\boldsymbol{\lambda},\boldsymbol{\Lambda}), \tag{65}$$

for $i \neq j$,

$$\frac{\partial Z_m}{\partial\Lambda_{ij}} = Z_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})\Big\{[\boldsymbol{\eta}_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})]_i[\boldsymbol{\eta}_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})]_j + [(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda})^{-1}]_{ij}\Big\}, \tag{66}$$

and for $i = j$,

$$\frac{\partial Z_m}{\partial\Lambda_{ii}} = \frac{1}{2}Z_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})\Big\{[\boldsymbol{\eta}_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})]_i[\boldsymbol{\eta}_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})]_i + [(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda})^{-1}]_{ii}\Big\}, \tag{67}$$

We can now easily deduce the results claimed in the text. First we derive the mixture representation (16) of the maximum-entropy densities. This follows directly from the main lemma (52) with

$$Z(\boldsymbol{\lambda},\boldsymbol{\Lambda}) = \sum_{m=1}^{M} w_m Z_m(\boldsymbol{\lambda},\boldsymbol{\Lambda}) \tag{68}$$

and

$$w_m(\boldsymbol{\lambda},\boldsymbol{\Lambda}) = w_m\frac{Z_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})}{Z(\boldsymbol{\lambda},\boldsymbol{\Lambda})}, \tag{69}$$

where $Z_m(\boldsymbol{\lambda},\boldsymbol{\Lambda}), \boldsymbol{\mu}_m(\boldsymbol{\lambda},\boldsymbol{\Lambda}), \mathbf{C}_m(\boldsymbol{\Lambda})$, $m = 1,...,M$ are given by (63),(62),(58). 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 (68). 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 (65)-(67) one obtains:

$$\frac{\partial F}{\partial\boldsymbol{\lambda}}(\boldsymbol{\lambda},\boldsymbol{\Lambda}) = \sum_{m=1}^{M} w_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})\boldsymbol{\eta}_m(\boldsymbol{\lambda},\boldsymbol{\Lambda}), \tag{70}$$

for $i \neq j$,

$$\frac{\partial F}{\partial\Lambda_{ij}} = \sum_{m=1}^{M} w_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})\Big\{[(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda})^{-1}]_{ij} + [\boldsymbol{\eta}_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})]_i[\boldsymbol{\eta}_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})]_j\Big\} \tag{71}$$

and for $i = j$,

$$\frac{\partial F}{\partial\Lambda_{ii}} = \frac{1}{2}\sum_{m=1}^{M} w_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})\Big\{[(\boldsymbol{\Gamma}_m^H - \boldsymbol{\Lambda})^{-1}]_{ii} + [\boldsymbol{\eta}_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})]_i[\boldsymbol{\eta}_m(\boldsymbol{\lambda},\boldsymbol{\Lambda})]_i\Big\}. \tag{72}$$

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 $m = 1, ..., M$ from (62) and (58) 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, (61) becomes

$$\boldsymbol{\eta}_m(\boldsymbol{\lambda}) = \boldsymbol{\mu}_m^H + \mathbf{C}_m^H \boldsymbol{\lambda}, \tag{73}$$

(63) becomes

$$Z_m(\boldsymbol{\lambda}) = \exp\left[-\frac{1}{2} (\boldsymbol{\mu}_m^H)^\top \boldsymbol{\Gamma}_m^H \boldsymbol{\mu}_m^H + \frac{1}{2} (\boldsymbol{\Gamma}_m^H \boldsymbol{\mu}_m^H + \boldsymbol{\lambda})^\top \boldsymbol{\eta}_m(\boldsymbol{\lambda}) \right], \tag{74}$$

(62) becomes

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

and (58) becomes simply

$$\mathbf{C}_m(\boldsymbol{\lambda}) = \mathbf{C}_m. \tag{76}$$

The thermodynamics also simplifies, with

$$F(\boldsymbol{\lambda}) = \ln Z(\boldsymbol{\lambda}) = \ln \left(\sum_{m=1}^M w_m Z_m(\boldsymbol{\lambda}) \right), \tag{77}$$

$$\frac{\partial F}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}) = \sum_{m=1}^M w_m(\boldsymbol{\lambda}) \boldsymbol{\eta}_m(\boldsymbol{\lambda}) = \boldsymbol{\eta}(\boldsymbol{\lambda}), \tag{78}$$

$$\frac{\partial^2 F}{\partial \lambda_i \partial \lambda_j}(\boldsymbol{\lambda}) = \sum_{m=1}^M w_m(\boldsymbol{\lambda}) \Big\{ [\mathbf{C}_m^H]_{ij} + [\boldsymbol{\eta}_m(\boldsymbol{\lambda}) - \boldsymbol{\eta}(\boldsymbol{\lambda})]_i [\boldsymbol{\eta}_m(\boldsymbol{\lambda}) - \boldsymbol{\eta}(\boldsymbol{\lambda})]_j \Big\} \tag{79}$$

and $w_m(\boldsymbol{\lambda}) = w_m Z_m(\boldsymbol{\lambda}) / Z(\boldsymbol{\lambda})$, as in (69). The cost to calculate $\boldsymbol{\eta}_m(\boldsymbol{\lambda}), Z_m(\boldsymbol{\lambda})$ for $m = 1, ..., M$ 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.

C Sampling from the Maximum-Entropy Models

Direct sampling from the maximum-entropy distributions using the mixture representation (16) 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 $\mathbf{C}_m(\mathbf{\Lambda}_t^+),\ m = 1, ..., M$ for every new value of $\mathbf{\Lambda}_t^+$. It might be possible to calculate the EOF's for the covariance matrices $\mathbf{C}_m(t),\ m = 1, ..., M$ of the components of the mixture-model (8) 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

$$\mathbf{C}_m^{-1}(\mathbf{\Lambda}) = \mathbf{C}_m^{-1} - \boldsymbol{\mathcal{H}}^\top \mathbf{\Lambda} \boldsymbol{\mathcal{H}}, \tag{80}$$

the inverse of (53) [where we drop from here on the explicit time label t]. This formula implies that the Gaussian component $N(\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \mathbf{\Lambda}), \mathbf{C}_m(\mathbf{\Lambda}))$ can be sampled by the Metropolis-Hastings algorithm with $N(\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \mathbf{\Lambda}), \mathbf{C}_m)$ as the proposal distribution and with

$$E(\mathbf{x}) = -\frac{1}{2}[\mathbf{h}(\mathbf{x}) - \boldsymbol{\eta}_m(\boldsymbol{\lambda}, \mathbf{\Lambda})]^\top \mathbf{\Lambda} [\mathbf{h}(\mathbf{x}) - \boldsymbol{\eta}_m(\boldsymbol{\lambda}, \mathbf{\Lambda})] \tag{81}$$

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

$$\mathbf{x}' = \boldsymbol{\mu}_m(\boldsymbol{\lambda}, \mathbf{\Lambda}) + \sum_{a=1}^{p} \xi_a \sqrt{\gamma_m^a}\, \hat{\mathbf{e}}_m^a, \tag{82}$$

where $\xi_a,\ a = 1, ..., p$ are i.i.d. normal random variables and $\gamma_m^a, \hat{\mathbf{e}}_m^a$ are the eigenvalues and eigenvectors of $\mathbf{C}_m$. Note that the eigensystems do not depend on $\mathbf{\Lambda}$ and that the vectors $\boldsymbol{\mu}_m(\boldsymbol{\lambda}, \mathbf{\Lambda})$ can be efficiently calculated from (62). The updates (82) 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 $P_M(\mathbf{x}; \boldsymbol{\lambda}, \mathbf{\Lambda})$ is then as follows: First, set $\mathbf{x}_m = \boldsymbol{\mu}_m(\boldsymbol{\lambda}, \mathbf{\Lambda})$ as an initial guess of the state in the mth component for each $m = 1, ..., M$. Then, successively for $n = 1, ..., N$, choose $\mathbf{x}^{(n)}$ by first selecting a component index $m = 1, ..., M$ with probability $w_m(\boldsymbol{\lambda}, \mathbf{\Lambda})$. For the selected m, generate a new trial state $\mathbf{x}'_m$ via (82) 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 $\mathbf{x}^{(n)},\ n = 1, ..., N$ will be generated, distributed according to $P_M(\mathbf{x}; \boldsymbol{\lambda}, \mathbf{\Lambda})$. 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 $\mathbf{\Lambda}$ is small, but rejection rates will be high if the values of the energy function E in (81) 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,

$$\boldsymbol{\lambda}^+ = \boldsymbol{\lambda}^- + \mathbf{R}^{-1}\mathbf{y} \approx \mathbf{R}^{-1}\mathbf{y}, \quad \mathbf{\Lambda}^+ = \mathbf{\Lambda}^- - \mathbf{R}^{-1} \approx -\mathbf{R}^{-1}, \tag{83}$$

and updated values of parameters, to leading order, are independent of their values $\boldsymbol{\lambda}^-, \mathbf{\Lambda}^-$ before the measurement. In that case, the mixture model $P_M(\mathbf{x}; \boldsymbol{\lambda}^+, \mathbf{\Lambda}^+)$ simplifies considerably. It is easy to show using (58), (62), (63) that, as $\|\mathbf{R}\| \to 0$, the following asymptotic formulas hold for the component weights

$$w_m^+ = \frac{w_m \exp\left\{-\frac{1}{2}[\mathbf{y} - \boldsymbol{\mu}_m^H]^\top \mathbf{\Gamma}_m^H [\mathbf{y} - \boldsymbol{\mu}_m^H]\right\}}{\mathcal{N}\sqrt{\mathrm{Det}\mathbf{C}_m^H}}[1 + O(\|\mathbf{R}\|)] \tag{84}$$

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

$$\boldsymbol{\mu}_m^+ = \boldsymbol{\mu}_m + \mathbf{C}_m \boldsymbol{\mathcal{H}}^\top \mathbf{\Gamma}_m^H (\mathbf{y} - \boldsymbol{\mu}_m^H) + O(\|\mathbf{R}\|) \tag{85}$$

and for component covariances

$$\mathbf{C}_m^+ = \mathbf{C}_m - \mathbf{C}_m \boldsymbol{\mathcal{H}}^\top (\mathbf{\Gamma}_m^H - \mathbf{\Gamma}_m^H \mathbf{R} \mathbf{\Gamma}_m^H) \boldsymbol{\mathcal{H}} \mathbf{C}_m + O(\|\mathbf{R}\|^2). \tag{86}$$

Notice that the exponential factor in (84) 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 (85) and that $\boldsymbol{\mathcal{H}}\mathbf{C}_m^+\boldsymbol{\mathcal{H}}^\top = \mathbf{R} + O(\|\mathbf{R}\|^2)$ from (86), as would be expected for the limit of very accurate measurements. A simple sampling scheme in this limit, therefore, is to choose components $m = 1, ..., M$ with the probabilities w_m^+ in (84) 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 (86) does not depend upon $\boldsymbol{\lambda}^-, \mathbf{\Lambda}^-$.

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.

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, $G_t(\mathbf{y}_t|\mathbf{x}^{(n)})$, $n = 1, ..., N$ for the update (42). 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 (44), for $n = 1, ..., N$. Resampling requires just N independent random numbers.

EnKF: To calculate the mean and the covariance by N-sample averages in the matching step (45) uses $O(Np^2)$ multiplications. In the limit $p \gg q$ which mostly concerns us, the further calculation of $\mathbf{C}_{t^-}^H$ requires $O(p^2q)$ multiplications, dominated by the matrix multiplication $\mathbf{C}_{t^-}\boldsymbol{\mathcal{H}}_t^\top$. Calculation of the inverse and/or Cholesky factors of $\mathbf{C}_{t^-}^H + \mathbf{R}_t$ is $O(q^3)$ multiplications and calculation of $\boldsymbol{\mathcal{K}}_t$ by multiplication of $\mathbf{C}_{t^-}\boldsymbol{\mathcal{H}}_t^\top$ and the inverse matrix requires an additional $O(pq^2)$. Hence, the total cost to calculate the Kalman gain matrix in (48) scales as $O(p^2 \max\{N, q\})$ in the limit $p \gg q$. Update of $\mathbf{x}_{t^-}^{(n)}$ to $\mathbf{x}_{t^+}^{(n)}$ for $n = 1, ..., N$ via (47) is another $O(Npq)$ operations, either by matrix multiplication with $\boldsymbol{\mathcal{K}}_t$ or by backsubstitution using the Cholesky factors of $\mathbf{C}_{t^-}^H + \mathbf{R}_t$ followed by multiplication with $\mathbf{C}_{t^-}\boldsymbol{\mathcal{H}}_t^\top$. Nq random numbers must be generated for the measurement resampling in (46).

MEF: The matching and updating steps, (15) 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 (15) 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 (9) for the Gaussian components of the mixture model (16). As discussed above, calculating $\boldsymbol{\mu}_m(t; \boldsymbol{\lambda}, \boldsymbol{\Lambda})$ and $\mathbf{C}_m(t; \boldsymbol{\Lambda})$ for $m = 1, ..., M$ in (16) requires $O(Mpq)$ and $O(Mp^2q)$ operations, respectively, at each measurement time t. This is just M times the cost to calculate the Kalman gain matrix in EnKF. On the other hand, generating new samples by (9) requires Np random numbers and $O(Np^2)$ multiplications, 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 (9), 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 (8) 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 (8). 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 $\mathbf{C}_m(t; \mathbf{\Lambda})$ in (16) 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 $\mathbf{C}_m,\ m = 1, ..., M$ in (8) 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 (10) 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 $m = 1, ..., M$ requires $O(Mp^2p_{\max})$ 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 $p_{\max}/p$.

MFF: The number of operations to calculate the function inside the brackets in (27) and its gradient in (29) 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 $O(n_{CG}M(q/p)^2/q)$ than the cost to calculate the Kalman gain matrix in EnKF. The resampling step in the mean-field MEF uses $O(Mpq)$ multiplications to calculate the means $\boldsymbol{\mu}_m(t; \boldsymbol{\lambda})$, $m = 1, ..., M$ in (16) [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 $\mathbf{C}_m,\ m = 1, ..., M$. Also, Np random numbers and $O(Np^2)$ multiplications are needed to generate new samples by (9). 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.

References

[1] A. F. Bennett. *Inverse Methods in Physical Oceanography*. Cambridge University Press, Cambridge, 1992.

[2] G. Burgers, P. J. van Leeuwen, and G. Evensen. Analysis scheme in the ensemble kalman filter. *Mon. Wea. Rev.*, 126:1719–1724, 1998.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

[4] D. Crisan and A. Doucet. A survey of convergence results on particle filtering for practitioners. *IEEE Trans. Signal Processing*, 50:736–746, 2002.

[5] L. Devroye. The double kernel method in density estimation. *Ann. Inst. H. Poincar*è, 25:533–580, 1989.

[6] A. Doucet, N.de Freitas, and N. Gordon editors. *Sequential Monte Carlo in Practice*. Springer-Verlag, 2001.

[7] R. Ellis. *Entropy, Large Deviations, and Statistical Mechanics*. Springer-Verlag, New York, 1985.

[8] G. Evensen. Inverse methods and data assimilation in nonlinear ocean modles. *Physica D*, 77:108–129, 1994a.

[9] 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.

[10] G. L. Eyink, J. M. Restrepo, and F. J. Alexander. A mean field approximation in data assimilation for nonlinear dynamics. *Physica D.*, 2003.

[11] D. Fox. Adapting the sample size in particle filters through kld-sampling. *Int. J. Robot. Res.*, 12:985–1003, 2003.

[12] M. Frontini and A. Tagliani. Maximum entropy in the finite stieltjes and hamburger moment problem. *J. Math. Phys.*, 35:6748–6756, 1994.

[13] A. Gelb. *Applied Optimal Estimation*. The MIT Press, Cambridge, MA, 1974.

[14] M. Ghil and K. Ide. Data assimilation in meteorology and oceanography: Theory and practice. *J. Meteor. Soc. Japan*, 75:111–496, 1997.

[15] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings-F*, 140:107–113, 1993.

[16] P. Hänggi, P. Talkner, and M. Borkovec. Reaction-rate theory: fifty years after kramers. *Rev. Mod. Phys*, 62(2):251–341, 1990.

[17] L. F. James, C. E. Priebe, and D. J. Marchette. Consistent estimation of mixture complexity. *Ann. Stat.*, 29(5):1281–1296, 2001.

[18] 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.

[19] G. Kitagawa and W. Gersch. *Smoothness Priors Analysis of Time Series*, volume 116 of *Lecture Notes in Statistics*. Springer-Verlag, 1996.

[20] R. Kleeman. Measuring dynamical prediction utility using relative entropy. *J. Atmos. Sci.*, 59:2057–2072, 2002.

[21] P. E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer Verlag, 1997.

[22] M. Loève. *Probability Theory, 3rd ed.* Van Nostrand, New York, 1963.

[23] E. N. Lorenz. Deterministic nonperiodic flow. *J. Atmos. Sci*, 20:130–141, 1963.

[24] G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, New York, 2000.

[25] P. Del Moral. Nonlinear filtering : interacting particle solution. *Markov Proc. Rel. Fields.*, 2:555–579, 1996.

[26] P. Del Moral. Nonlinear filtering using random particles. *Theor. Probab. Appl.*, 40:690–701, 1996.

[27] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer-Verlag, 1999.

[28] J. Jacod P. Del Moral and Ph. Protter. The monte carlo method for filtering with discrete-time observations. *Probab. Theory. Rel.*, 120:346–368, 2001.

[29] C. E. Priebe. Adaptive mixtures. *J. Amer. Stat. Assoc.*, 89:796–806, 1994.

[30] C. E. Priebe and D. J. Marchette. Alternating kernel and mixture density estimates. *Comput. Stat. Data. An.*, 35:43–65, 2000.

[31] H. Risken. *The Fokker-Planck Equation*. Springer-Verlag, New York, 1984.

[32] F. Schoegl. Fluctuations in thermodynamic non-equilibrium states. *Z. Physik*, 244:199–205, 1971.

[33] F. Schoegl. On stability of steady states. *Z. Physik*, 243:303–310, 1971.

[34] D. W. Scott. *Multivariate Density Estimation*. John Wiley, New York, 1992.

[35] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York, 1986.

[36] W. Tucker. A rigorous ode solver and smale's 14th problem. *Found. Comput. Math.*, 2:53–117, 2002.

[37] S. Ulam and J. von Neumann. The monte carlo method. *J. Amer. Stat. Assoc.*, 44:335–341, 1949.

[38] 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.

[39] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman and Hall, London, 1995.

[40] L.-S. Young. What are srb measures and which dynamical systems have them? *J. Stat. Phys.*, 108:733–754, 2002.

Table Captions

1. Relative Mean Errors in Experiment A.
2. Relative Mean Errors in Experiment B.
3. Relative Mean Errors in Experiment C.

Figure Captions

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

2. Particle filter results for Experiment A with $N = 10^2$ 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 $\pm$ standard deviations of the approximate filters. The dashed lines are the mean and $\pm$ standard deviations from the Fokker-Planck solution of (10).

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

4. 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 (10), and the black dots (joined by solid lines) are the approximate values from the particle filters.

5. 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 Figure 4.

6. 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.

7. Relative entropy for Experiment A with $N = 10^4$ samples: (a) EnKF, (b) MEF, (c) MFF. Symbols as in Figure 6.

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

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

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 Figure 4.

11. Relative entropy for Experiment B with $N = 10^2$ samples: (a) EnKF, (b) MEF, (c) MFF. Symbols as in Figure 6.

12. Relative entropy for Experiment B with $N = 10^4$ samples: (a) EnKF, (b) MEF, (c) MFF. Symbols as in Figure 6.

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

14. WRF results for $\bar{x}(t)$ in Experiment C. (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.

15. WRF results for $\sigma_x(t)$ in Experiment C. (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.

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

17. Approximate filter standard deviations $\sigma_x(t)$ for Experiment C with $N =$

10^2. (a) EnKF, (b) MEF, and (c) MFF. Approximations shown as solid lines and exact filter result (WRF) as dot-dashed line.

18. Log-likelihoods $L_{1:t}$ versus time t for Experiment C with $N = 10^2$ 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.

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

Table 1. Relative Mean Errors in Experiment A

(a) Filter Mean

N	WRF	$EnKF$	MEF	MFF
10^2	1.12297975	1.10542974	0.01507894	0.09369440
10^4	0.01067231	0.52647795	0.00187048	0.06703097

(b) Filter Standard Deviation

N	WRF	$EnKF$	MEF	MFF
10^2	0.03442152	0.03567663	0.03216601	0.57671912
10^4	0.02368712	0.01193383	0.00970724	0.54157417

Table 2. Relative Mean Errors in Experiment B

(a) Filter Mean

N	WRF	$EnKF$	MEF	MFF
10^2	0.04648154	0.07271474	0.04568027	0.13560871
10^4	0.00739388	0.05204111	0.00687383	0.09932946

(b) Filter Standard Deviation

N	WRF	$EnKF$	MEF	MFF
10^2	0.08999746	0.18219459	0.08845087	0.404094241
10^4	0.01668710	0.15788042	0.01384682	0.379310678

Table 3. Relative Mean Errors in Experiment C

(a) Filter Mean

N	$EnKF$	MEF	MFF
10^2	0.65833419	0.44622136	0.64868442
10^4	0.54089056	0.41519496	0.62956885

(b) Filter Standard Deviation

N	$EnKF$	MEF	MFF
10^2	1.48521106	1.52572389	3.28886961
10^4	1.43654565	1.63067845	3.28098684

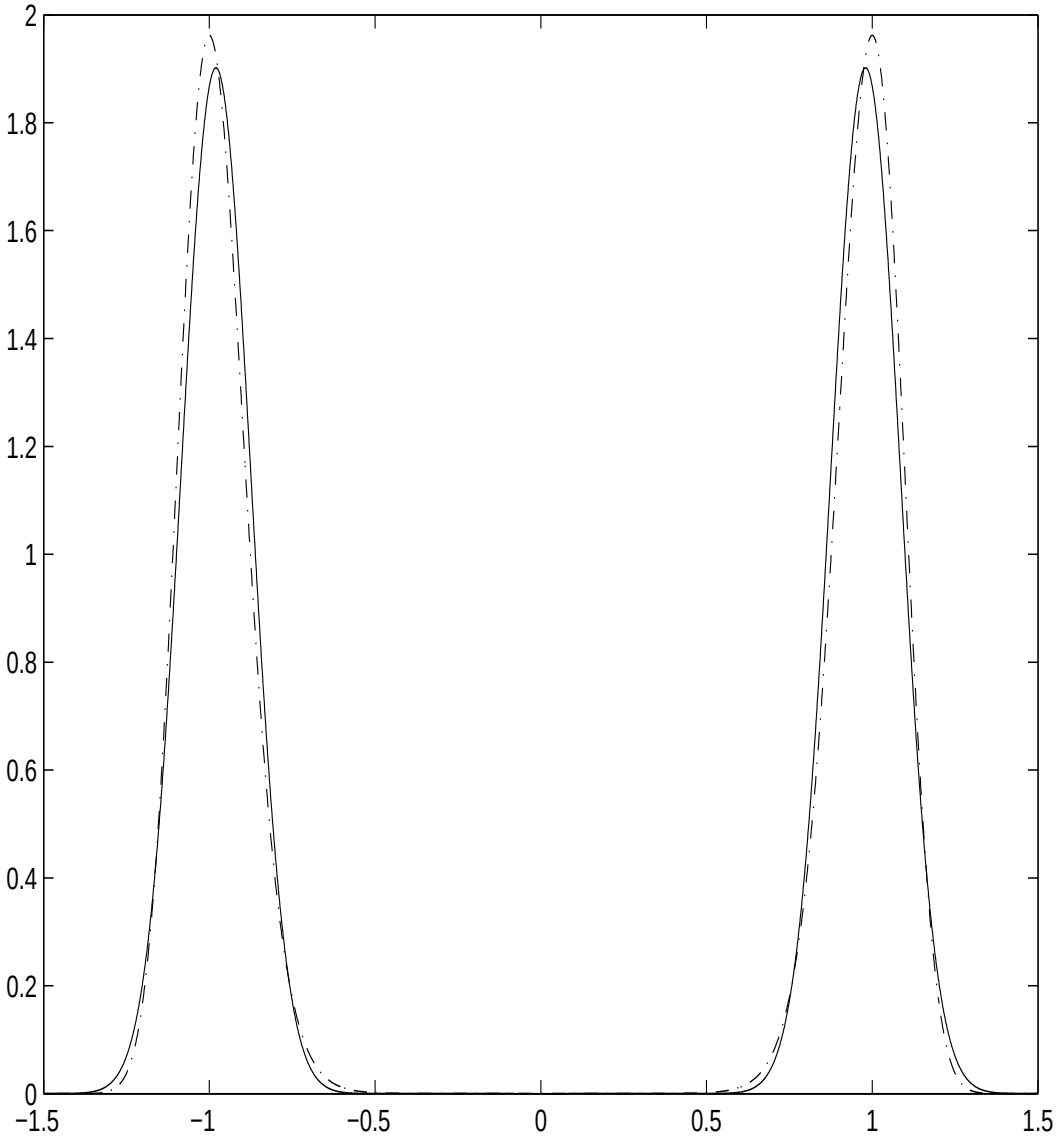

Figure 1: Exact steady state density for DW model, $\kappa = 0.4$ (dashed line) and mixture model with $M = 2$ (solid line).

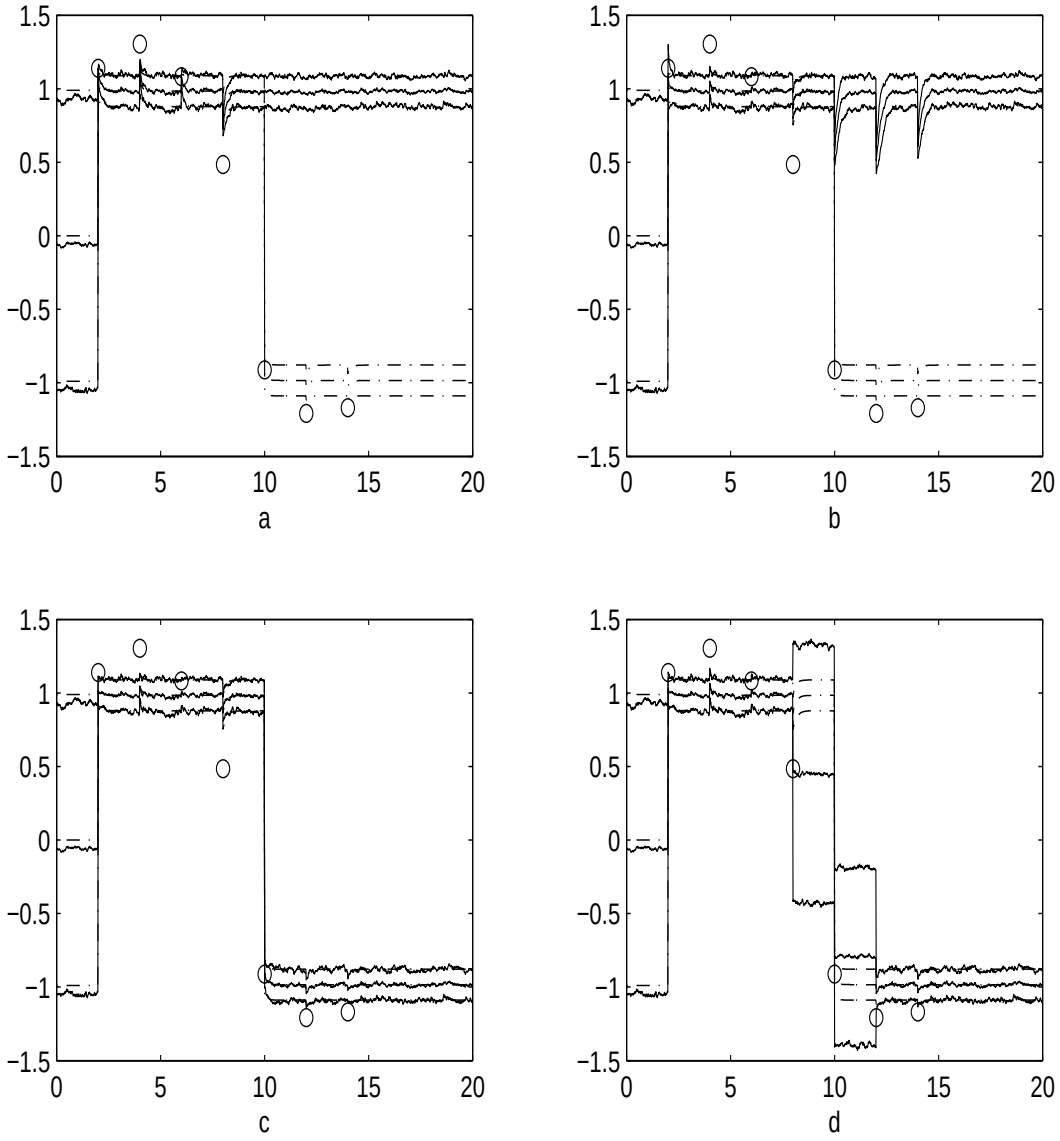

Figure 2: Particle filter results for Experiment A with $N = 10^2$ samples: (a) WRF, (b) EnKF, (c) MEF, (d) MFF. The circles represent measurements taken from one sample path and the solid lines are the mean and $\pm$ standard deviations of the approximate filters. The dashed lines are the mean and $\pm$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

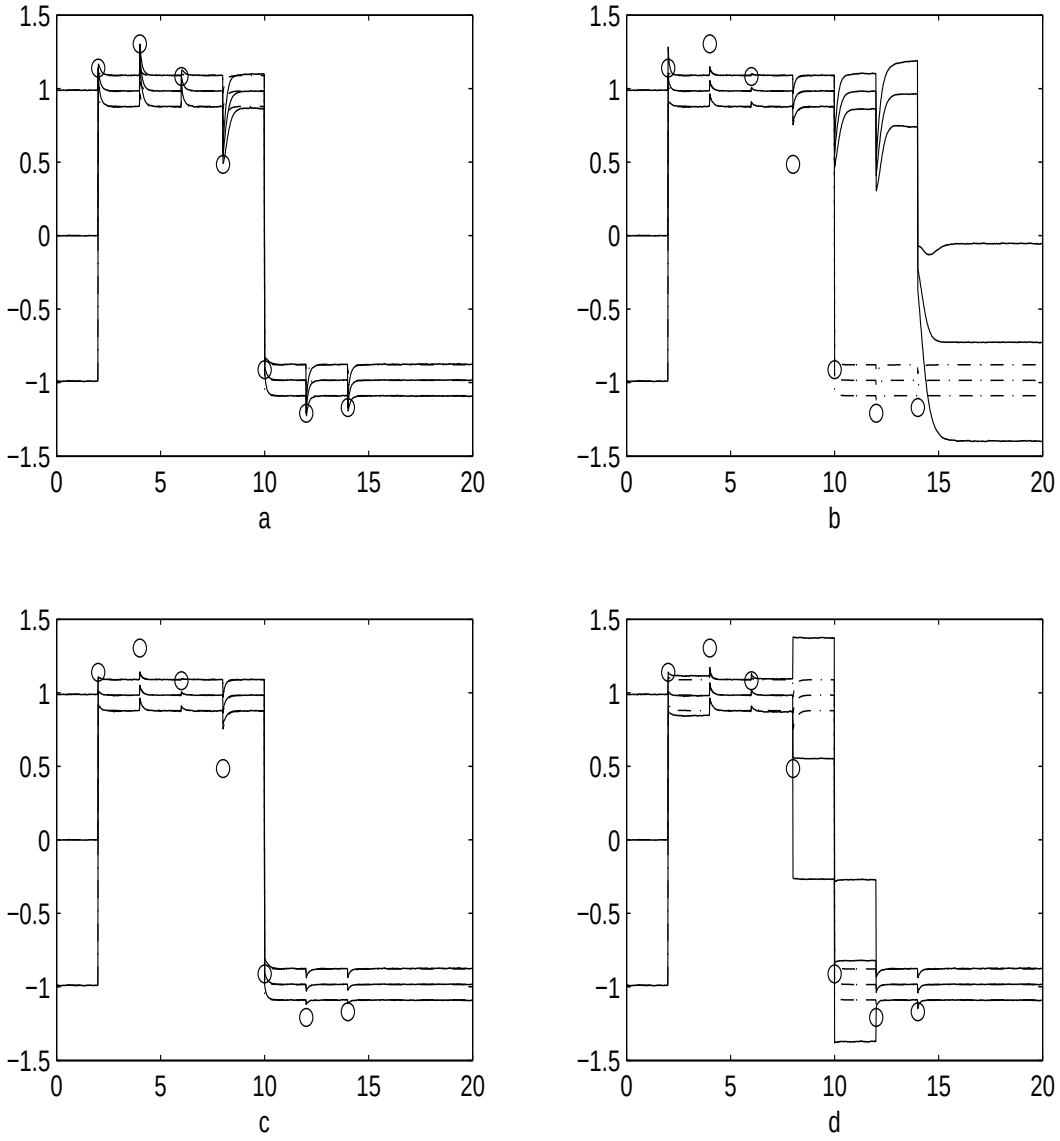

Figure 3: Particle filter results for Experiment A with $N = 10^4$ samples: (a) WRF, (b) EnKF, (c) MEF, (d) MFF. Symbols as in Figure 2.

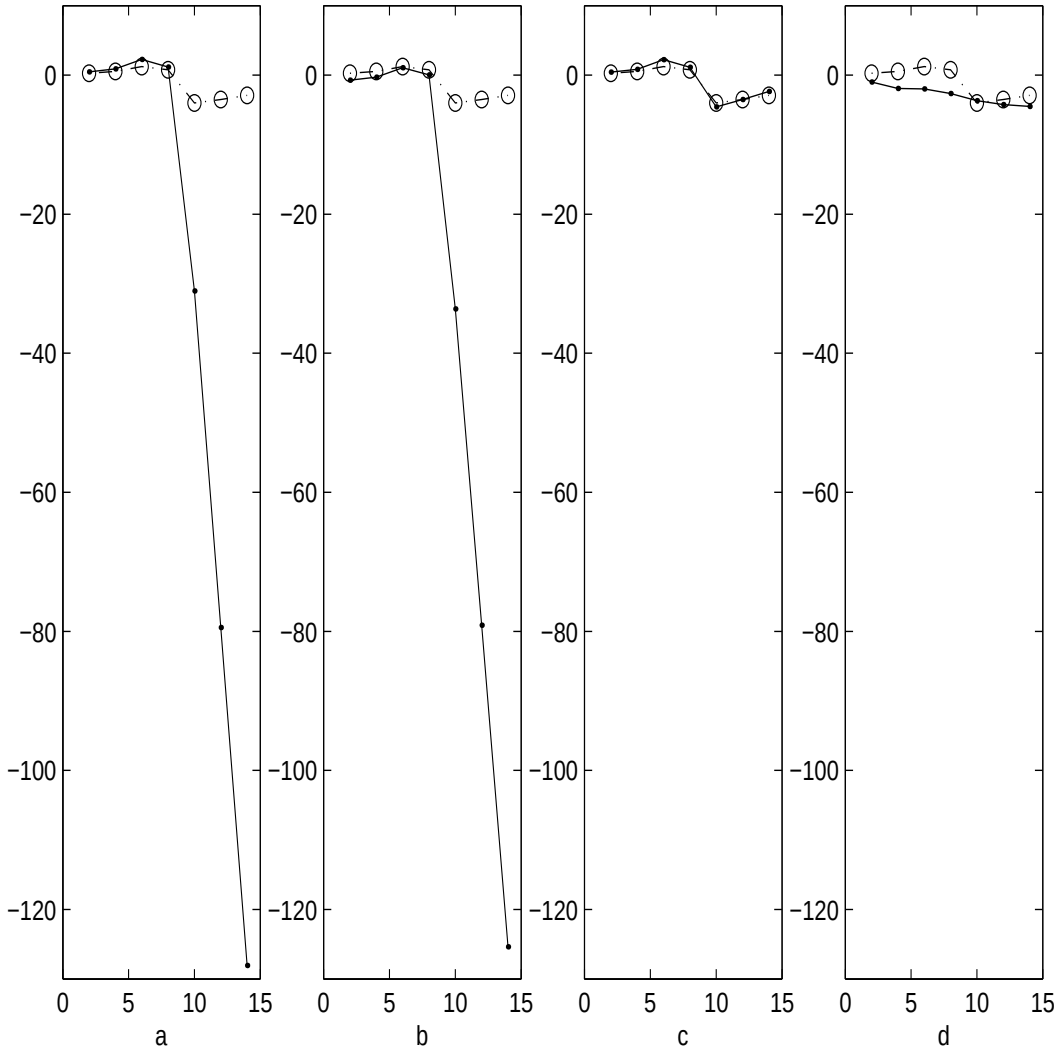

Figure 4: Log-likelihoods $L_{1:t}$ versus time t for Experiment A with $N = 10^2$ samples: (a) WRF , (b) EnKF, (c) MEF, (d) MFF. The circles (joined by dotted lines) are the exact values from the Fokker-Planck solution of (10), and the black dots (joined by solid lines) are the approximate values from the particle filters.

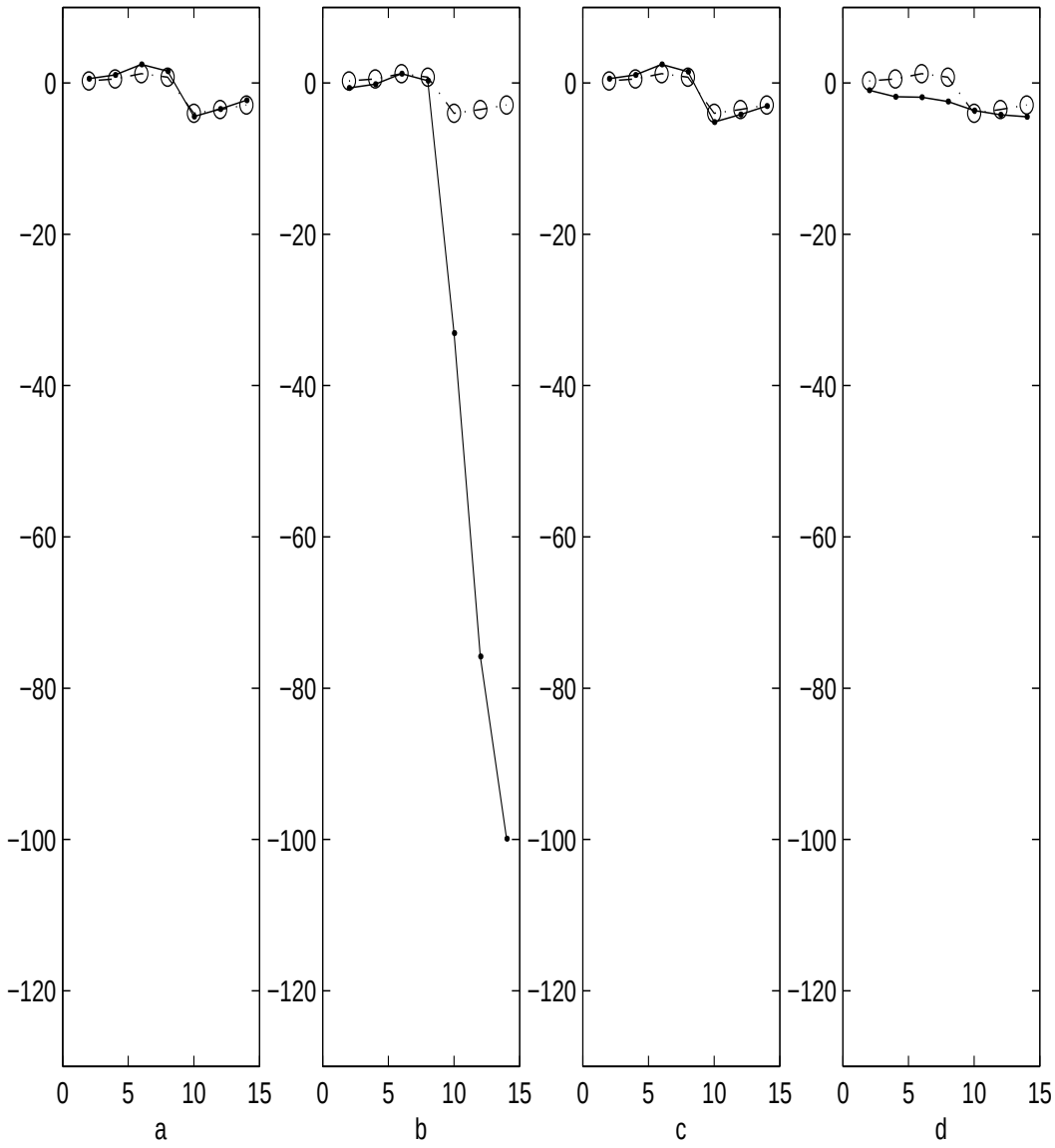

Figure 5: Log-likelihoods $L_{1:t}$ versus time t for Experiment A with $N = 10^4$ samples: (a) WRF , (b) EnKF, (c) MEF, (d) MFF. Symbols as in Figure 4.

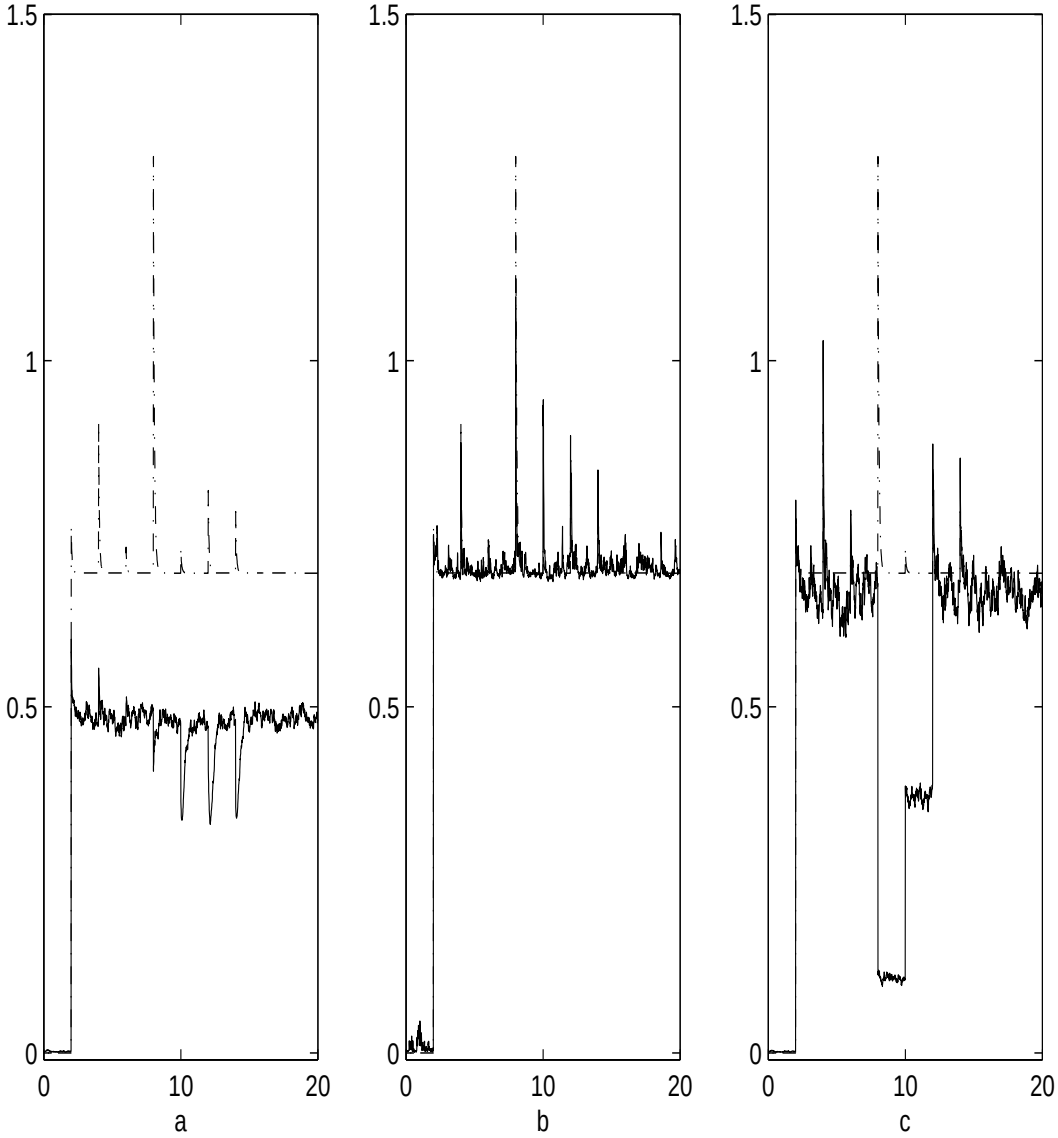

Figure 6: Relative entropy for Experiment A with $N = 10^2$ samples: (a) EnKF, (b) MEF, (c) MFF. The solid line is the approximation and the dashed line is the exact result from the Fokker-Planck solution.

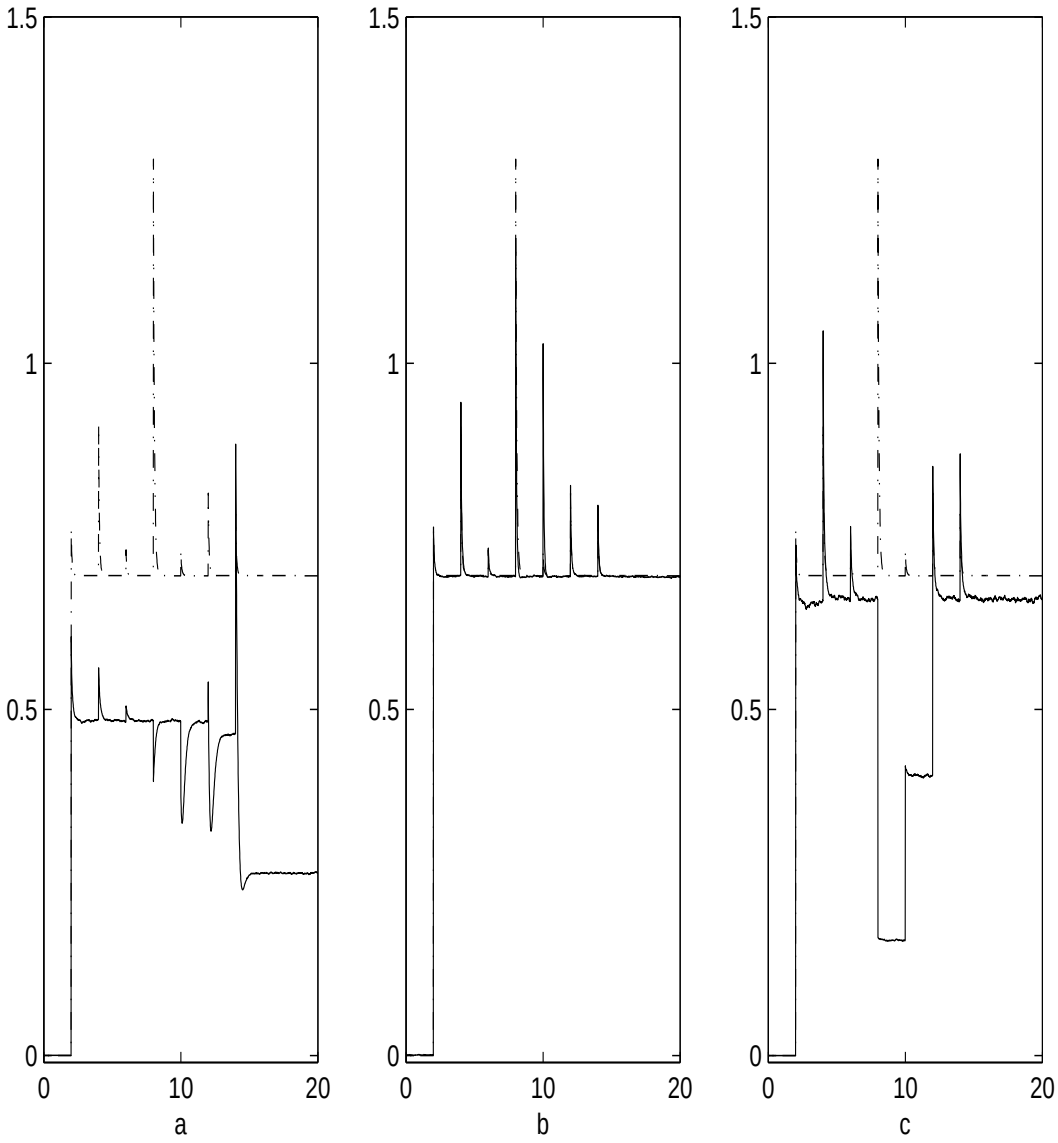

Figure 7: Relative entropy for Experiment A with $N = 10^4$ samples: (a) EnKF, (b) MEF, (c) MFF. Symbols as in Figure 6.

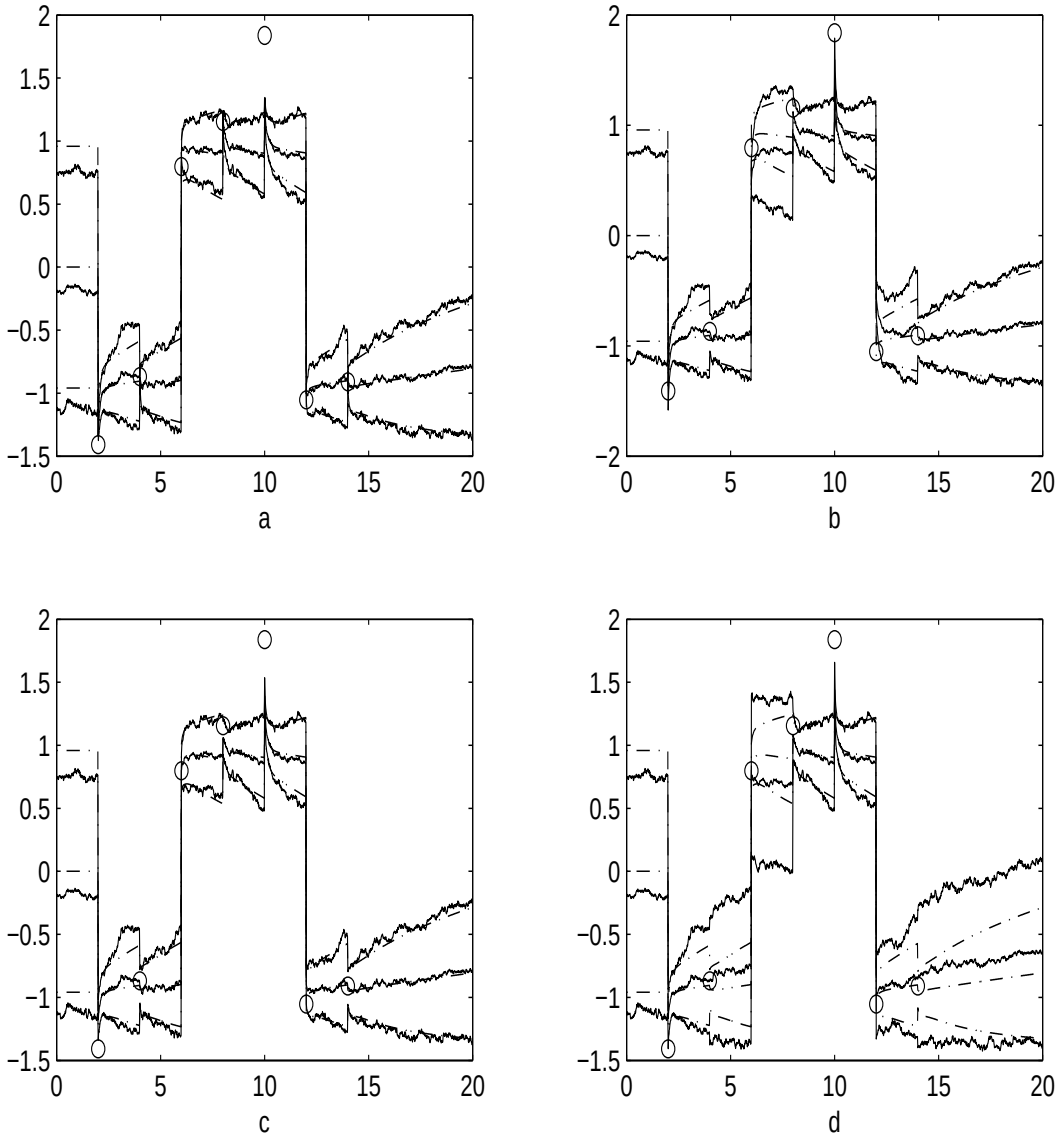

Figure 8: Particle filter results for Experiment B with $N = 10^2$ samples: (a) WRF, (b) EnKF, (c) MEF, (d) MFF. Symbols as in Figure 2.

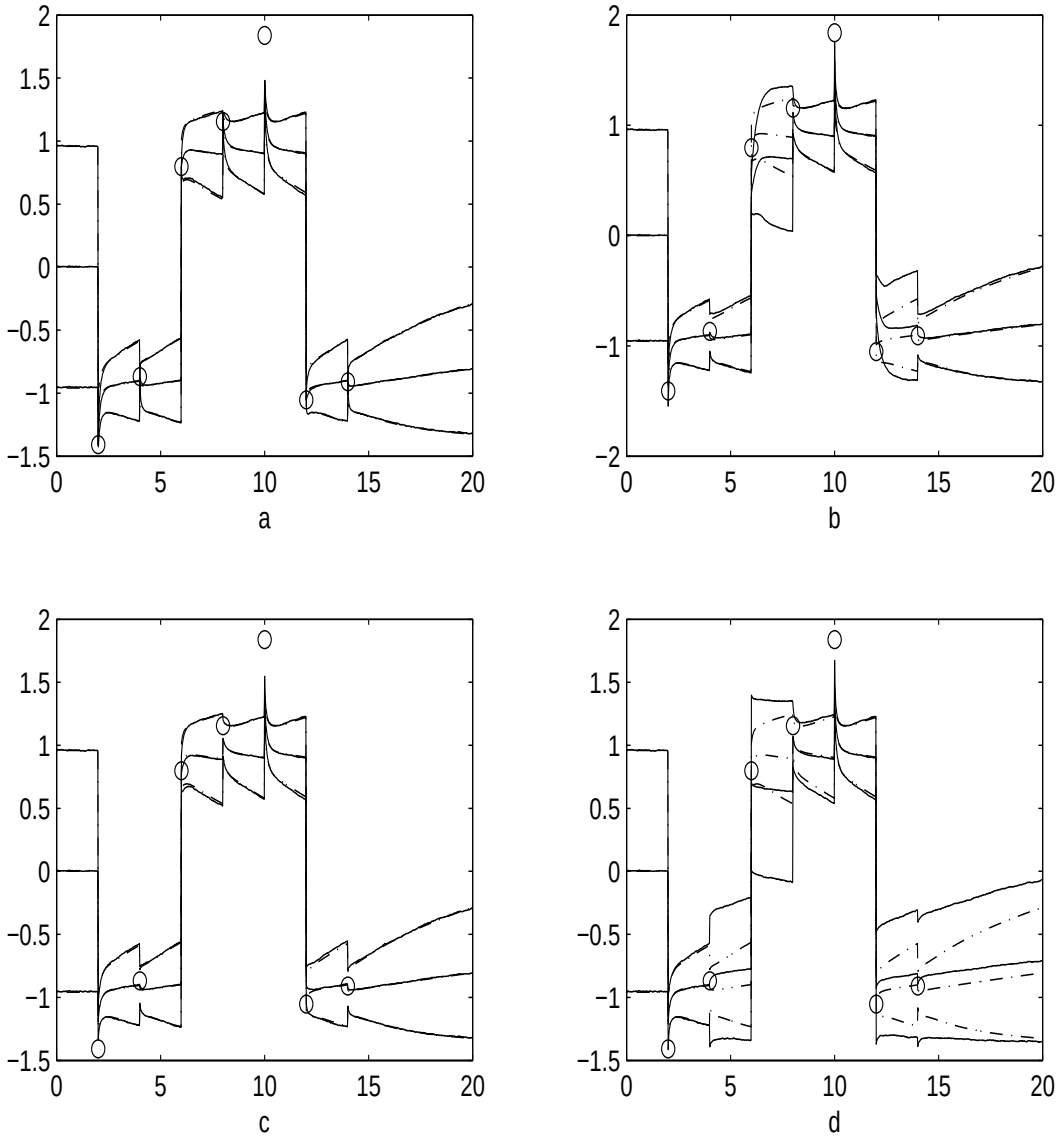

Figure 9: Particle filter results for Experiment B with $N = 10^4$ samples: (a) WRF, (b) EnKF, (c) MEF, (d) MFF. Symbols as in Figure 2.

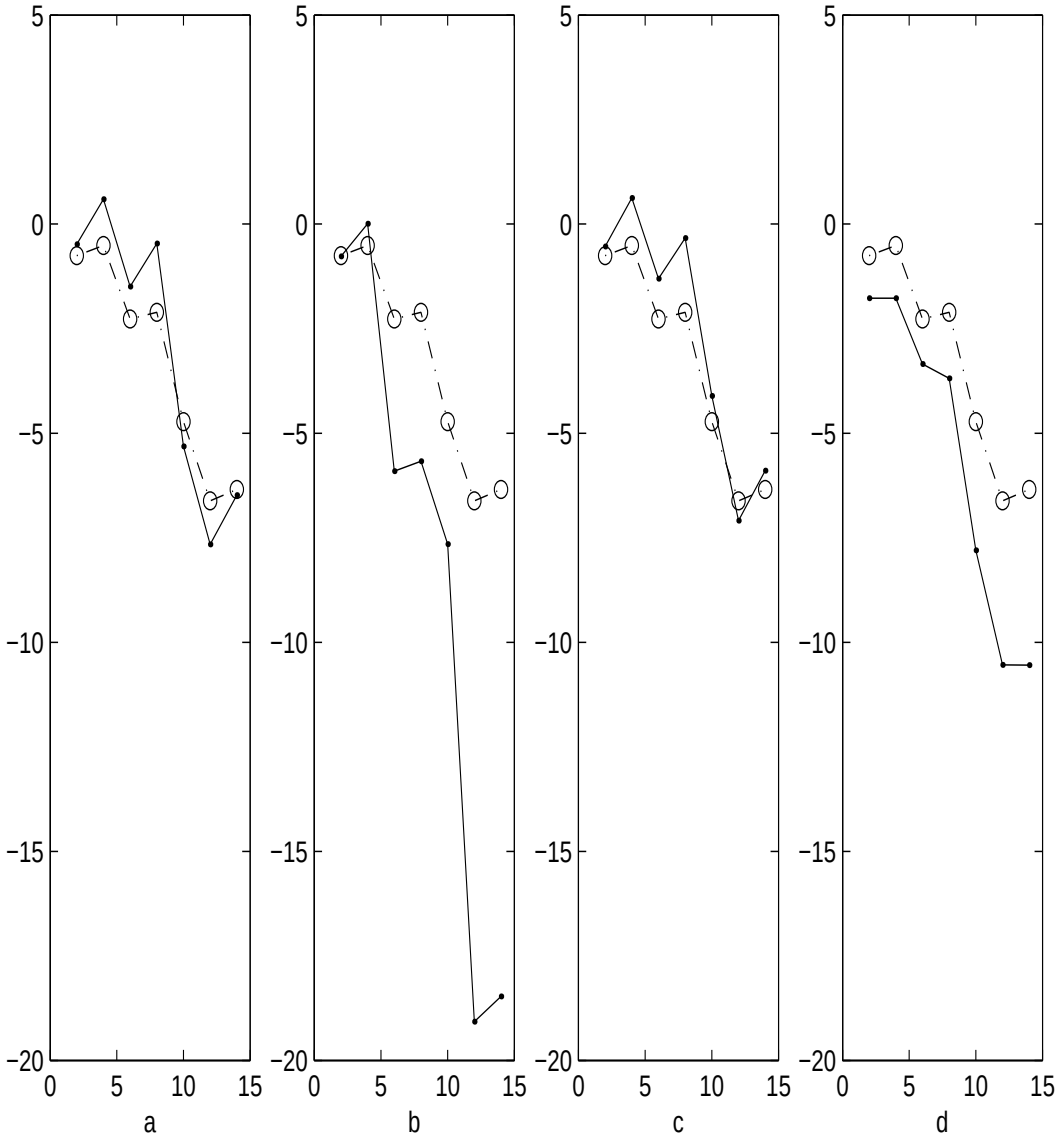

Figure 10: Log-likelihoods $L_{1:t}$ versus time t for Experiment B with $N = 10^2$ samples: (a) WRF , (b) EnKF, (c) MEF, (d) MFF. Symbols as in Figure 4.

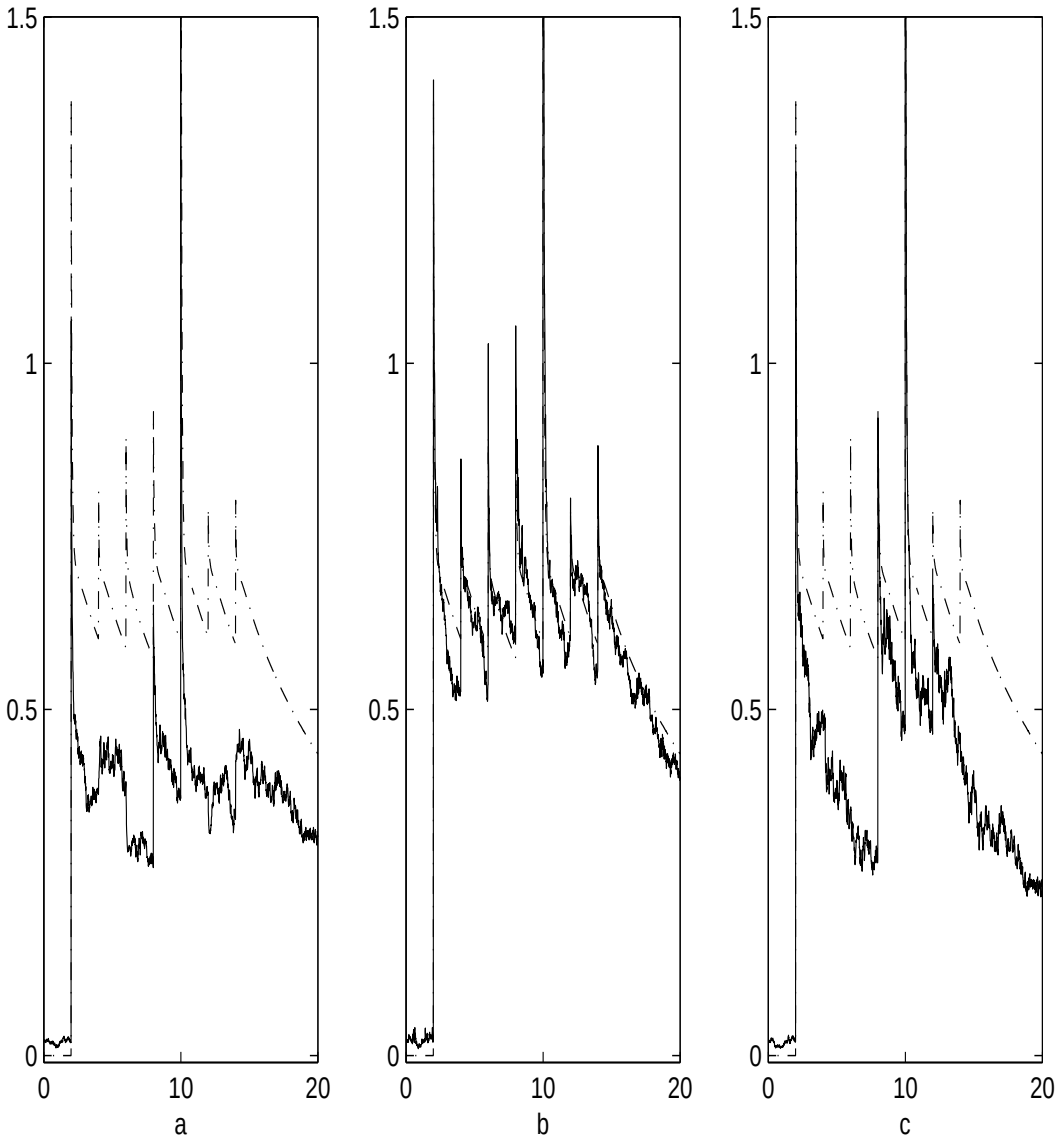

Figure 11: Relative entropy for Experiment B with $N = 10^2$ samples: (a) EnKF, (b) MEF, (c) MFF. Symbols as in Figure 6.

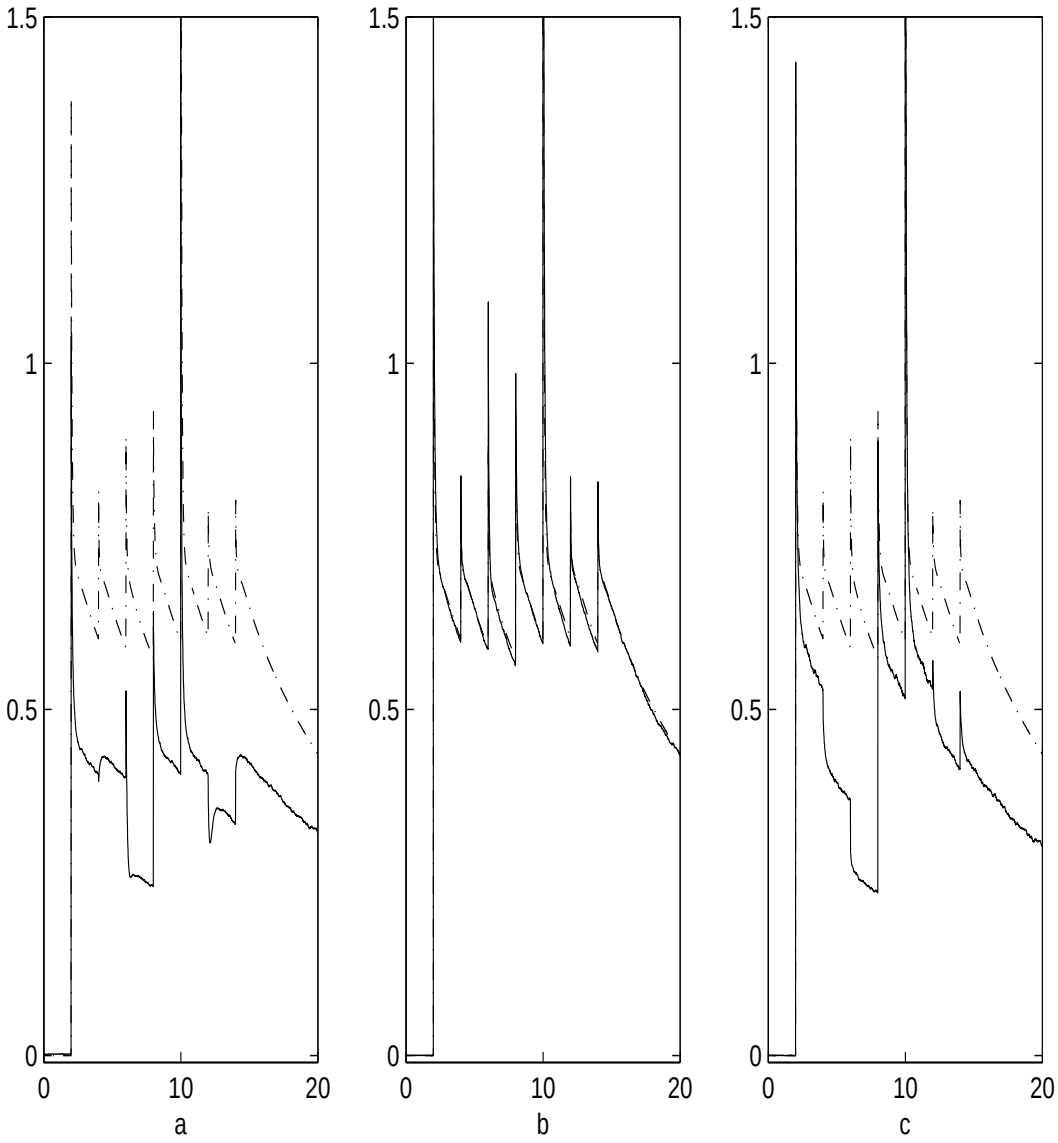

Figure 12: Relative entropy for Experiment B with $N = 10^4$ samples: (a) EnKF, (b) MEF, (c) MFF. Symbols as in Figure 6.

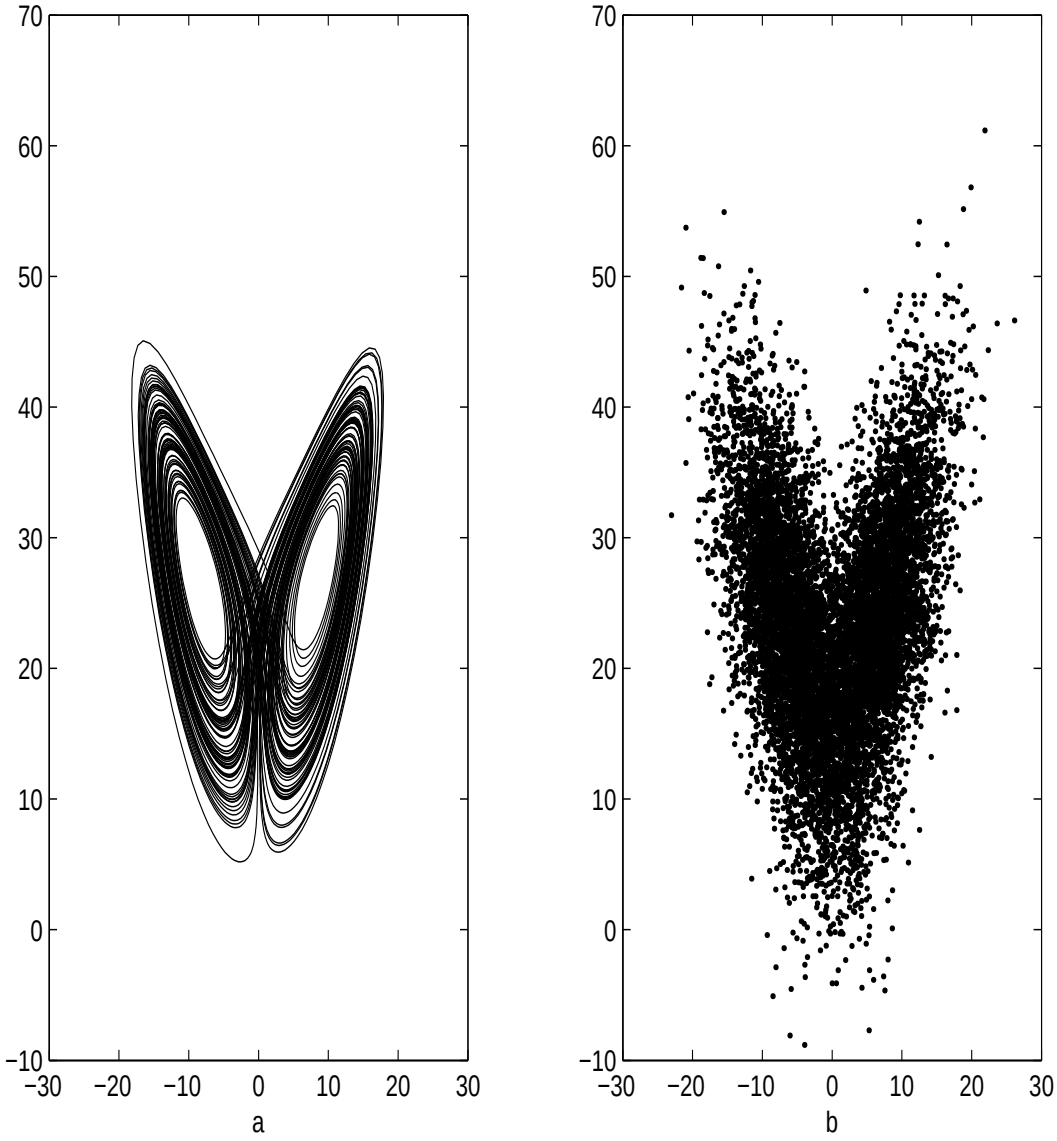

Figure 13: (a) The Lorenz attractor and (b) scatterplot of samples from the mixture model ($M = 2$), both projected to the xy-plane.

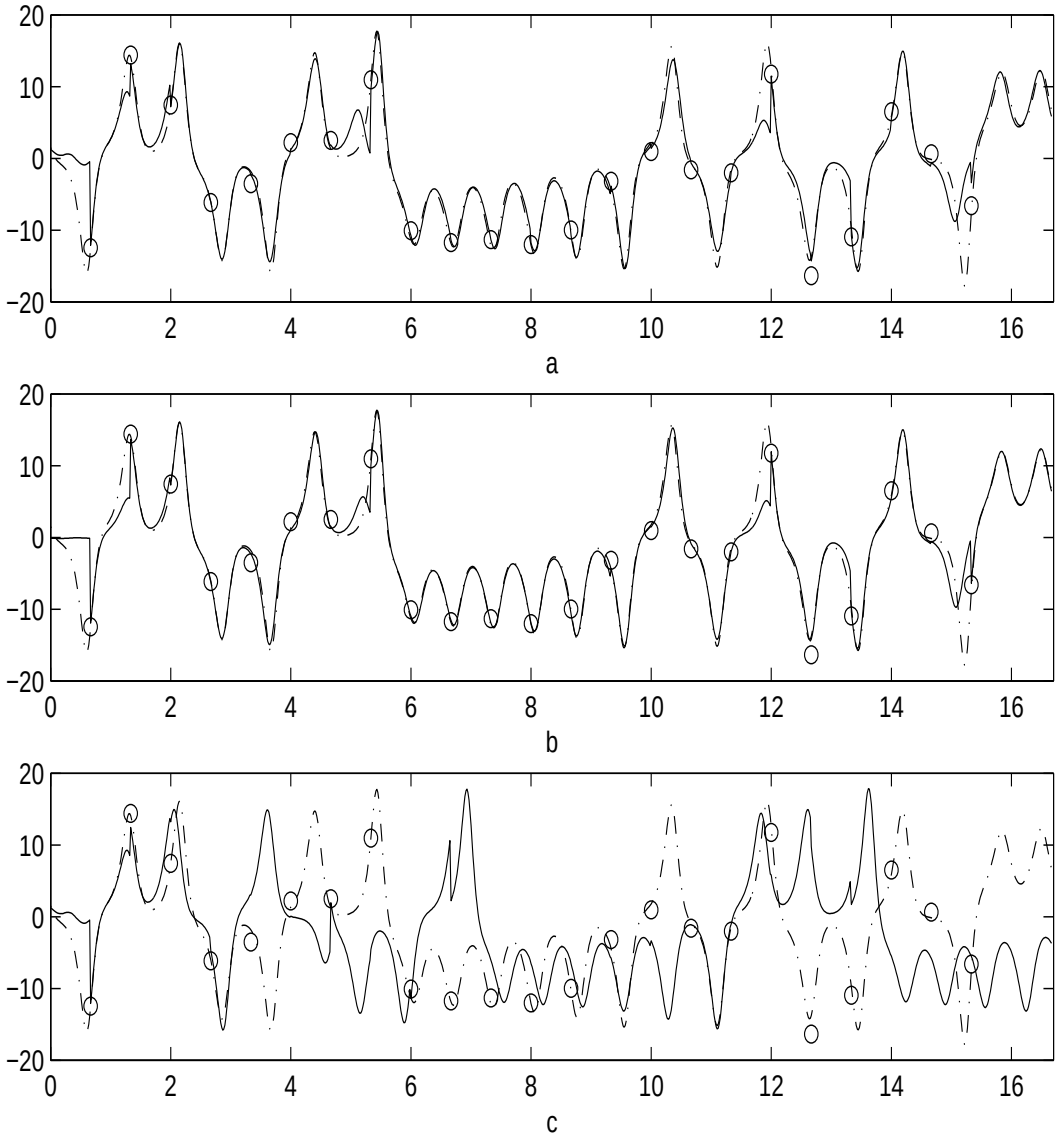

Figure 14: WRF results for $\bar{x}(t)$ in Experiment C. (a)$N = 10^2$, $\delta_N = 0.6$, (b) $N = 10^4$, $\delta_N = 0.1$, and (c) $N = 10^2$, $\delta_N = 0.1$. Measurement data shown as circles, mean as solid line, original solution trajectory as dot-dashed line.

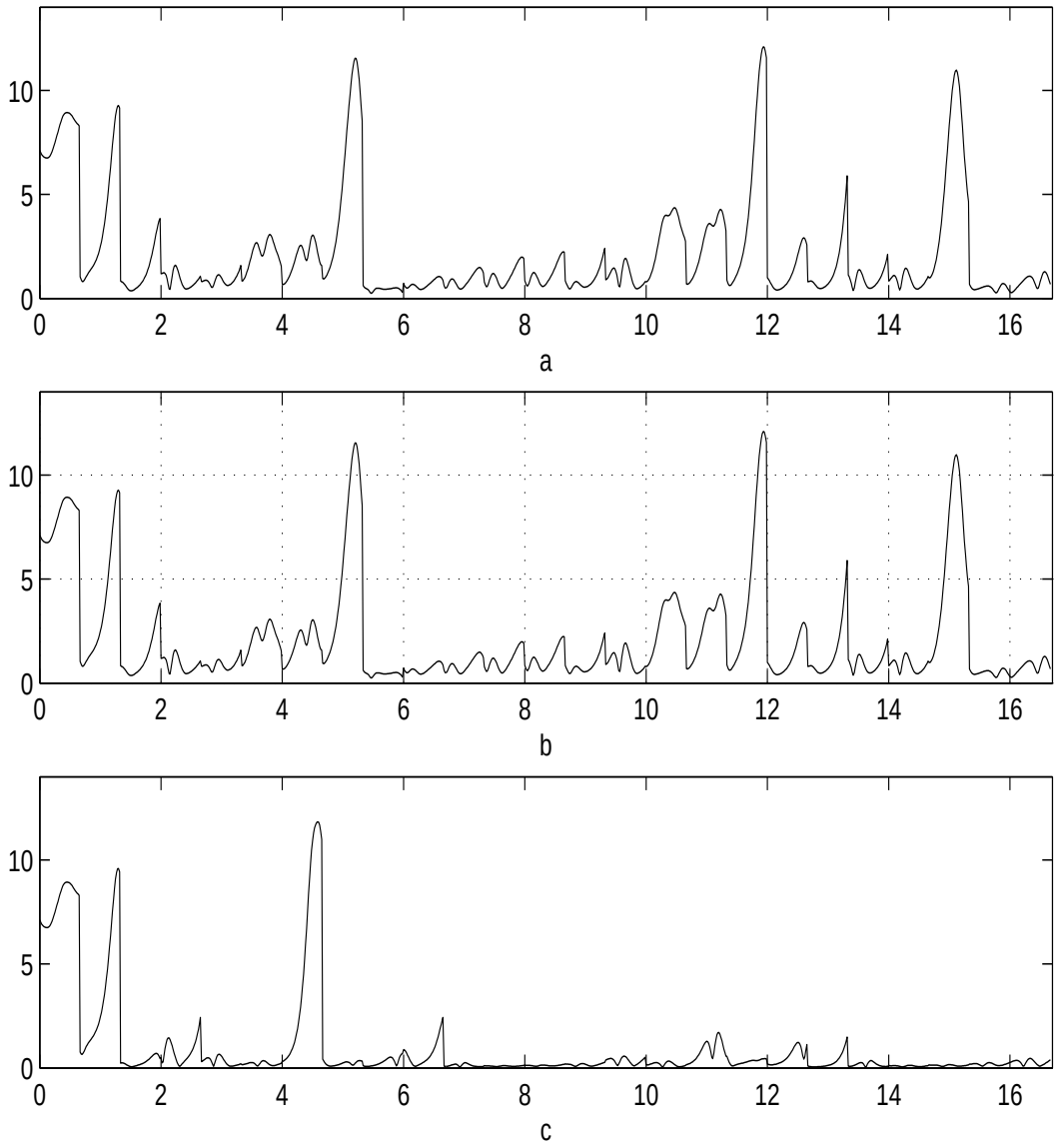

Figure 15: WRF results for $\sigma_x(t)$ in Experiment C. (a) $N = 10^2$, $\delta_N = 0.6$, (b) $N = 10^4$, $\delta_N = 0.1$, and (c) $N = 10^2$, $\delta_N = 0.1$. The conditional standard deviation is plotted versus time as a solid line.

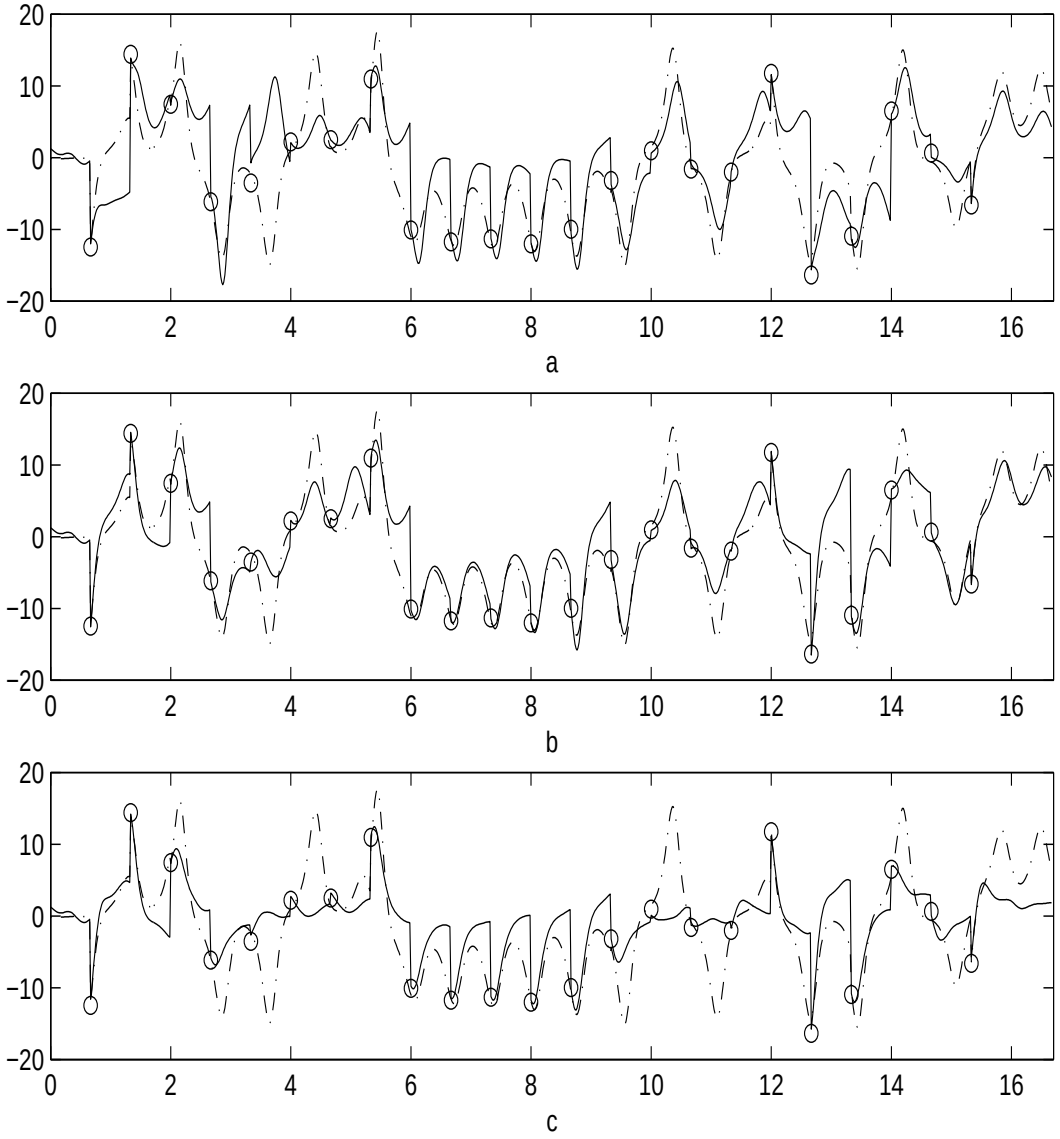

Figure 16: Approximate filter means $\bar{x}(t)$ for Experiment C with $N = 10^2$. (a) EnKF, (b) MEF, and (c) MFF. Measurement data shown as circles, approximations as solid lines, exact filter result (WRF) as dot-dashed line.

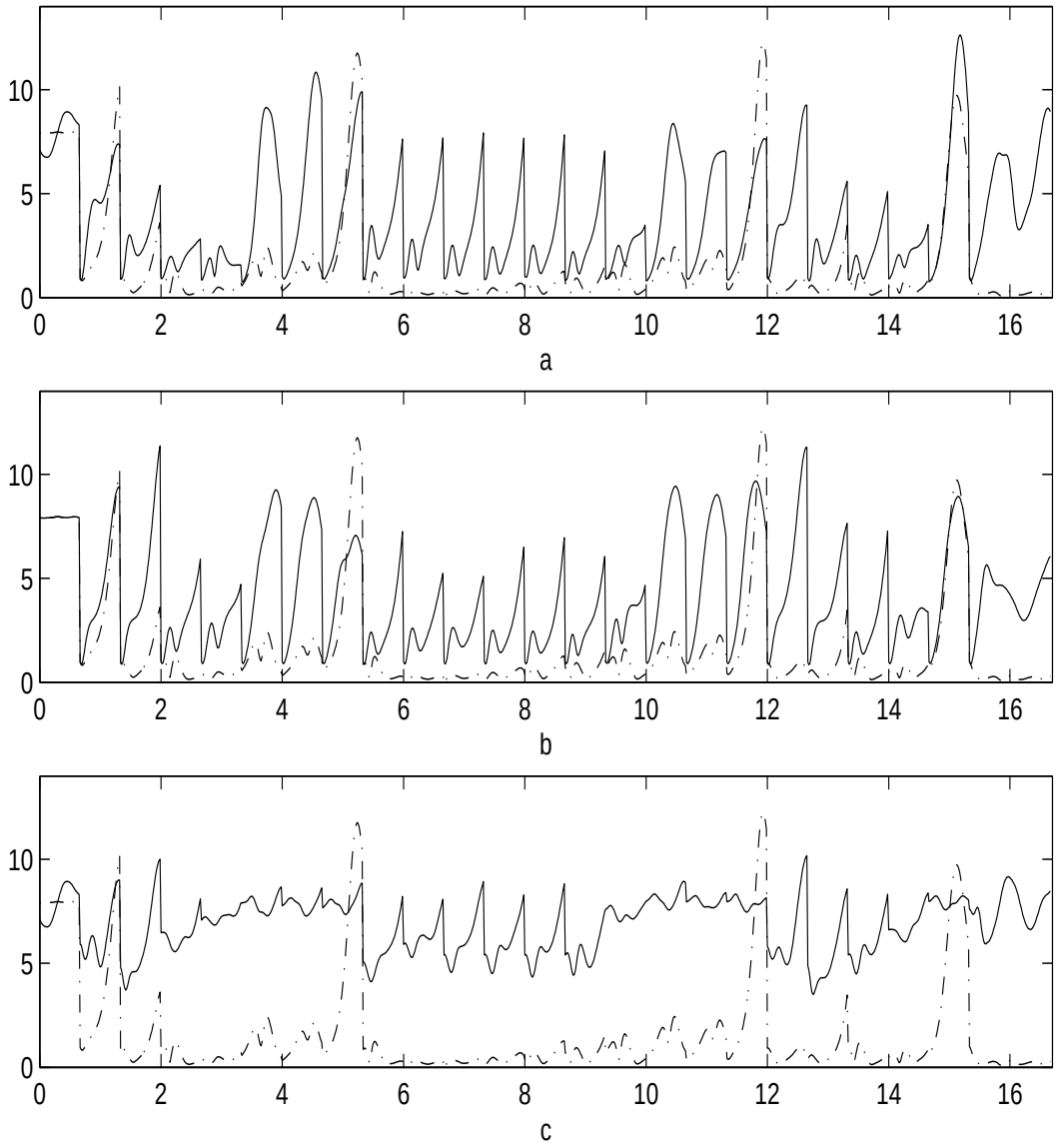

Figure 17: Approximate filter standard deviations $\sigma_x(t)$ for Experiment C with $N = 10^2$. (a) EnKF, (b) MEF, and (c) MFF. Approximations shown as solid lines and exact filter result (WRF) as dot-dashed line.

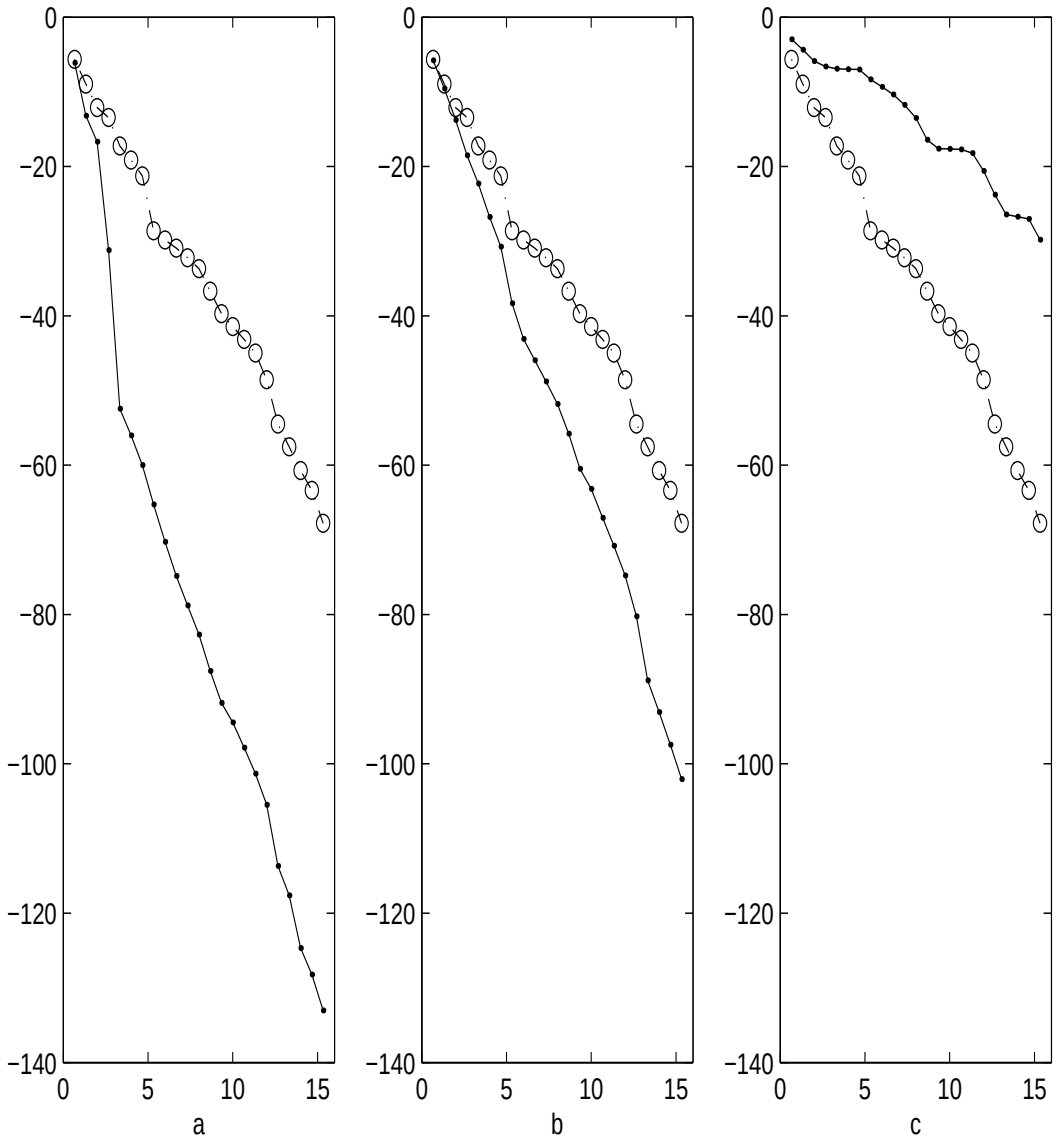

Figure 18: Log-likelihoods $L_{1:t}$ versus time t for Experiment C with $N = 10^2$ samples: (a) EnKF, (b) MEF, (c) MFF. The circles (joined by dotted lines) are the exact values from the WRF method with $N = 10^4$, and the black dots (joined by solid lines) are approximations from the other particle filters.

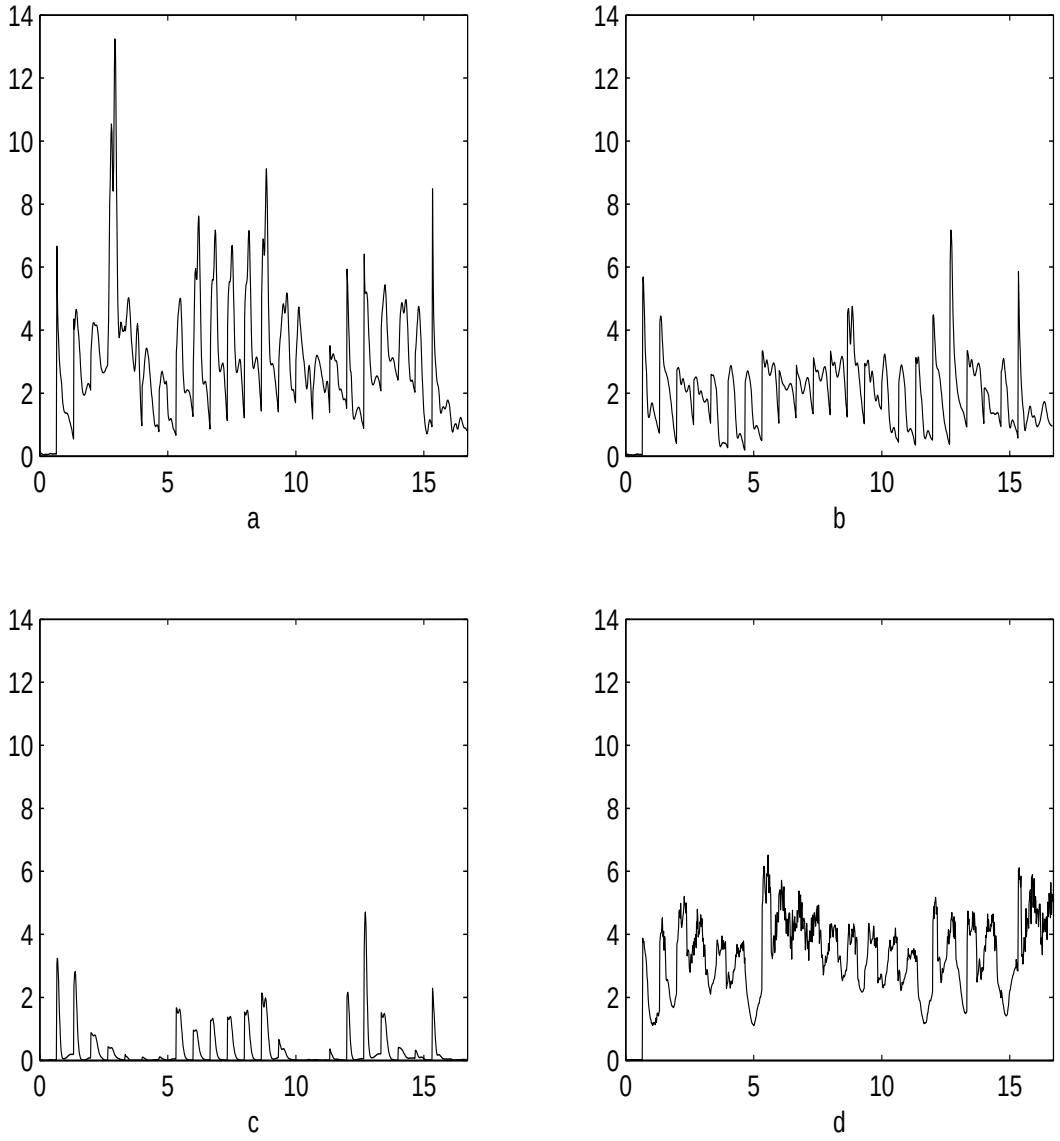

Figure 19: Entropy for Experiment C with $N = 10^2$. (a) EnKF; (b) MEF; and (c) MFF; (d) Relative entropy of xy-marginals using WRF with $N = 10^4$.