Ensemble Filtering for Nonlinear Dynamics

Sangil Kim [1], Gregory L. Eyink [2,*], Juan M. Restrepo [1],

Francis J. Alexander [3], and Gregory Johnson [3]

[1] *Department of Mathematics*

University of Arizona, Tucson, AZ 85721 U.S.A.

[2] *Department of Mathematical Sciences*

The Johns Hopkins University, Baltimore, MD 21218 U.S.A.

eyink@mts.jhu.edu

[3] *CCS-3, Los Alamos National Laboratory*

Los Alamos, NM 87545 U.S.A.

April 19, 2003

[*] *Corresponding author.*

Abstract

A method for data assimilation currently being developed is the Ensemble Kalman Filter. This method evolves the statistics of the system by computing an empirical ensemble of sample realizations, and incorporates measurements by a linear interpolation between observations and predictions. However, such an interpolation is only justified for linear dynamics and Gaussian statistics, and it is known to produce erroneous results for nonlinear dynamics with far-from-Gaussian statistics. For example, the Ensemble Kalman Filter method, when used in models with multimodal statistics, fails to track state transitions correctly.

We study here alternative ensemble methods for data assimilation into nonlinear dynamical systems, in particular those with a large state space. In these methods we calculate conditional probabilities at measurement times by applying Bayes' rule. Our results show that the new methods accurately track the transitions between likely states in a system with bimodal statistics, where the Ensemble Kalman Filter method does not perform well. The proposed new ensemble methods are conceptually simple and potentially applicable to large-scale problems.

1 Motivation and Model

The problem of data assimilation is to determine the best estimate of the solution history of a dynamical system given some partial and inaccurate measurements. The *filtering problem* is defined as that of estimating the present state, given prior observations. It is generally accepted that the optimal solution is obtained by calculating the conditional statistical distribution $P(\mathbf{x}, t|\mathcal{R}(t))$ of the state vector $\mathbf{x}$ of the system, given the set $\mathcal{R}(t)$ of measurements up to the current time t; see Stratonovich (1960) and Kushner (1962), also Lorenc and Hammon (1988). Between measurements, the conditional probability density solves the forward Kolmogorov equation, a parabolic partial differential equation for the state space. At measurement times, the probability density function is updated by Bayes' rule:

$$P(\mathbf{x}, t_m^+|\mathcal{R}(t_m^+)) = \frac{1}{\mathcal{N}_m} P(\mathbf{y}, t_m|\mathbf{x}) P(\mathbf{x}, t_m^-|\mathcal{R}(t_m^-)) \tag{1}$$

where $\mathcal{R}(t_m^{\mp})$ is the set of measurements before and after the application of Bayes' rule at time t_m, $P(\mathbf{y}, t_m|\mathbf{x})$ is the probability of the outcome of the measurement given the current state of the system, and $\mathcal{N}_m$ is a normalization factor. Such an optimal filtering scheme has been implemented for simple models of stochastic ordinary differential equations in Miller *et al.* (1999).

Solving the forward Kolmogorov equation for multi-dimensional, many-variable systems is not a practical option. An effective way to circumvent this difficulty is to evolve the statistics by computing an N-sample ensemble of realizations $\mathbf{x}_n(t)$ of the system, for $n = 1, \ldots, N$. The initial conditions for these realizations are drawn independently from an assumed prior distribution at the starting time and are assigned equal weights $\frac{1}{N}$ for the calculation of ensemble-averages. In the geosciences, the best known Monte Carlo

method of this type is the *Ensemble Kalman Filter (EnKF)*; see Evensen (1994), also Evensen and van Leeuwen (2000) and references therein. In this approach a "Kalman gain matrix" is calculated from ensemble-average statistics at measurement times and employed to make a linear interpolation between current ensemble states $\mathbf{x}_n(t_m^-)$ and error-contaminated measurements. These generate a new ensemble of realizations $\mathbf{x}_n(t_m^+)$ for $n = 1, \ldots, N$ constructed at each time where measurements are taken. This is traditionally called the "analysis step" in filtering applications. The new set of realizations are then evolved forward under the model dynamics to the next measurement time and the procedure is repeated.

Unfortunately, the EnKF prescription for generating the new samples $\mathbf{x}_n(t_m^+),\ n = 1, \ldots, N$ does not, in general, produce an ensemble representative of the correct conditional distribution (1). Only for linear dynamics with Gaussian error statistics does the interpolation scheme with the "Kalman gain matrix" lead to such a correct ensemble. The EnKF analysis step lacks any fundamental justification for nonlinear dynamics with non-Gaussian statistics.

These problems are more than of merely theoretical concern. In fact, it has been shown in simple models with multi-modal distributions that the EnKF fails to properly track transitions between probable states. In Miller *et al.* (1999) and Evensen and van Leeuwen (2000) the nonlinear stochastic differential equation in one variable $x(t) \in \mathbb{R}^1$ was investigated:

$$\begin{aligned} \frac{dx(t)}{dt} &= f(x) + \kappa\eta(t), \qquad t > 0 \\ x(0) &= x_0 \end{aligned} \tag{2}$$

$\eta(t)$ is white-noise with zero mean and covariance $\langle \eta(t)\eta(t') \rangle = \delta(t - t')$. The

function $f(x) = -U'(x)$, where the potential $U(x) = -2x^2 + x^4$ with minima at $x = \pm 1$. The steady-state probability distribution P_s giving the "climate statistics" of this model is proportional to $\exp(-\frac{2U(x)}{\kappa^2})$. Hence, it is bimodal with peaks at $x = \pm 1$, the two fixed points of the deterministic dynamics. Without the stochastic term in (2), the value of $x(t)$ will tend toward one of the stable steady states for any initial condition. The stochastic forcing term, however, can drive the state of the system from one potential well to the other. An important test for a data assimilation scheme is to see whether it can track such transitions, given measurements y_m of the state of the system at times t_m,

$$y_m = x(t_m) + \rho_m,$$

$m = 1, ..., M$, where ρ_m are random observation errors. For the case where the latter are assumed independent and normally distributed, Miller *et al.* (1999) have calculated the exact conditional statistics from the forward Kolmogorov equation and Bayes' rule. They found that the EnKF lags in tracking the transition from one well to the other in this model compared with the optimal filter result. See also Eyink *et al.* (2002b). Evensen and van Leeuwen (2000) acknowledged this problem with EnKF, and further found that the inability of EnKF to track transitions accurately is rooted in the linear analysis at measurement times.

We study in this same model some alternative ensemble methods of data assimilation with a different analysis step than EnKF. As will be shown later, these new methods successfully track the state transitions for equation (2). In what follows it will also be clear that these methods have important appealing characteristics: they are very simple, efficient, and potentially applicable to

large-scale spatially extended systems.

2 Bayesian Ensemble Methods

The idea behind each of these methods is just to apply Bayes' rule (1) at measurement times t_m appropriately. Suppose that an F-dimensional vector $\mathbf{h}_m(\mathbf{x})$ is measured at time t_m, where $\mathbf{h}_m$ an arbitrary function of the D-dimensional state vector $\mathbf{x}$. The outcome of the measurement will be assumed to be of the form

$$\mathbf{y}_m = \mathbf{h}_m(\mathbf{x}(t_m)) + \boldsymbol{\rho}_m,$$

where $\boldsymbol{\rho}_m$ is a normally distributed observation error with mean zero and covariance matrix $\mathbf{R}_m$. Then Bayes' rule (1) at the measurement time t_m becomes

$$P(\mathbf{x}, t_m^+) = \frac{\exp\left[-\frac{1}{2}(\mathbf{y}_m - \mathbf{h}_m(\mathbf{x}))^\top \mathbf{R}_m^{-1}(\mathbf{y}_m - \mathbf{h}_m(\mathbf{x}))\right]}{\mathcal{N}_m} P(\mathbf{x}, t_m^-) \qquad (3)$$

where we abbreviate, for simplicity, $P(\mathbf{x}, t) = P(\mathbf{x}, t|\mathcal{R}(t))$.

We now investigate several schemes for implementing this formula when using ensemble methods to evolve the statistics of a dynamical system.

2.1 Weighted Ensemble Filter (WEF)

The first method that we study shall be referred to here as the Weighted Ensemble Filter (WEF); in the applied probability literature it is more commonly referred to as a "particle filter". It was originally proposed by Ulam and von Neumann (1949) but only became more widely known after the work of Enachescu (1985). In this approach, one assigns to each sample $\mathbf{x}_n(t)$ a variable weight $w_n(t)$, $n = 1, ..., N$, which is altered according to

(3) at measurement times. Each of the realizations is given initial weight $\frac{1}{N}$ at the starting time. At a measurement time t_m, Bayes' rule is applied to recalculate the weight for each realization as

$$w_n(t_m^+) = \frac{\exp\,[\,-\frac{1}{2}(\mathbf{y}_m - \mathbf{h}_m(\mathbf{x}_n(t_m^-)))^\top \mathbf{R}_m^{-1}(\mathbf{y}_m - \mathbf{h}_m(\mathbf{x}_n(t_m^-)))]}{\mathcal{N}_m} w_n(t_m^-),$$

where $\mathcal{N}_m$ is a normalization factor to ensure that $\sum_{n=1}^N w_n(t) = 1$ for all times t. Weights remain constant between measurement times. In this scheme, averages of a moment variable $M(\mathbf{x})$ over the weighted ensemble are given by

$$\langle M(t) \rangle = \sum_{n=1}^{N} w_n(t) M(\mathbf{x}_n(t)).$$

Note that the calculation of the weights is very simple, even when $\mathbf{x}$ has very high dimension D. The most expensive operation, in fact, is the evaluation of the nonlinear function $\mathbf{h}_m(\mathbf{x})$, N times.

This scheme can be shown to be convergent to the optimal filter estimate in the limit as $N \to \infty$. See Del Moral (1996) and also the useful recent review Crisan and Doucet (2002). However, the rate of convergence is rather slow, with the $O(1/\sqrt{N})$ errors typical of all Monte Carlo methods. Furthermore, in practical applications, an operational ensemble prediction can deal with only moderate values of N, generally $N \ll 100$. For a discussion of this point, see Miller and Ehret (2002), where extensive further references to the literature may be found. For these reasons, the convergence properties of the scheme as $N \to \infty$ may be mostly of academic interest. Instead, one must consider the performance of the method with only moderate numbers of samples. Therefore, in our test of this method, and of the other methods introduced below, we shall consider both a large ensemble with $N = 10^4$ and a smaller one with $N = 100$, to see the effect of reduced number of samples.

Our results on the WEF method in the model system (2) are shown in Figure 1. with $N = 100$ samples used in obtaining Figure 1(a) and $N = 10^4$ in obtaining Figure 1(b). The simulated measurements are represented by the open circles. These were generated by adding independent normal random errors of zero mean and standard deviation 0.2 to a particular solution $x(t)$ of (2), sampled at unit time intervals. The solution considered experienced a state transition from $x = +1$ to $x = -1$ sometime around $t = 4 \sim 5$. For reference, we have plotted in dashed lines the optimal filter mean and plus and minus the standard deviation, which were calculated using the forward Kolmogorov equation, as in Miller *et al.* (1999). As may be seen, the optimal filter estimate clearly indicates the transition around $t = 4 \sim 5$. The WEF estimates of the same statistics are represented by solid lines in Figure 1.

As shown in Figure 1(b), the WEF method for $N = 10^4$ also succeeds in tracking the transition at $t = 4 \sim 5$, whereas it was shown in Miller *et al.* (1999) and Eyink *et al.* (2002), that EnKF lags by one time unit in making the transition. Evensen and van Leeuwen (2000) have traced this defect of EnKF to the linear analysis using the Kalman gain matrix. We see here that an ensemble method which employs an analysis scheme based instead on an approximate implementation of Bayes' theorem correctly tracks the transition with no lag. The WEF result for $N = 10^4$ is, in fact, quite close to the optimal filter result, except that after the transition the fluctuation effects in the averages are rather large. Most of the weight is given then to just a few samples that made the transition from $x = 1$ to $x = -1$ at the correct time. Therefore, the effective size of the ensemble is much less than 10^4 and only a few fluctuating samples dominate the averages. Nevertheless, the correspondence of WEF for $N = 10^4$ with the optimal filter averages is

quite acceptable.

However, for $N = 100$, the agreement is much worse. As seen in Figure 1(a), the transition is now missed entirely, even after three consecutive measurements in the well at $x = -1$. The only sign of the transition is a broadening of the variance after the third measurement. The reason is that, with only 100 realizations, no sample made the transition from $x = 1$ to $x = -1$ at the correct time (or even after the subsequent three measurements). Therefore, the transition to $x = -1$ occurred only very slowly due to the gradual growth of the weights, after successive measurements, of samples already residing in the $x = -1$ well. From this example we see that the consequences of small sample size in the WEF method may be severe for realistic applications, where N is at most a few hundreds.

2.2 Weight Resampling Filter (WRF)

One idea to overcome the problem with effective reduced sample size in the WEF method is to redistribute the large weights. This leads us to a simple modification of WEF, already considered in other contexts by several authors, see Kitagawa and Gersch (1996) (also Kitagawa, 1996, Kitagawa, 1998, Gordon *et al.*, 1993), which we call here the Weight Resampling Filter (WRF). Its potential application to data assimilation for atmospheric predictability has already been studied by Anderson and Anderson (1999) and Pham (2001) in the context of the chaotic 3-dimensional model of Lorenz (1963). This method is almost the same as WEF save for an additional resampling at measurement times. That is, after we have new weights $w_n(t_m^+)$ the same as in WEF above, we select a new N-sample ensemble of realizations each with

the same weight $\frac{1}{N}$. The new ensemble is generated by selecting from the N values $\mathbf{x}_n(t_m^-),\ n = 1, ...N$ of the old ensemble with probabilities determined by the weights $w_n(t_m^+)$. After this resampling, some of the realizations may disappear and some of them may be replicated several times. The result, however, is that large weights are redistributed over samples of equal weight. Note that for deterministic dynamics, replicated samples will evolve in unison in the future and, hence, shall effectively represent still a single sample with a large weight. This is not a problem in our stochastic system (2), or other systems with random model error, because identical initial conditions which experience different realizations of the system noise have distinct future evolutions. However, for deterministic dynamics, the replicated samples in the WRF method must be variously perturbed to show different future behaviors. Such perturbations may be generated by the singular vector or bred vector methods discussed in Miller and Ehret (2002), by density kernel methods as in Hürzeler and Künsch (1998), Anderson and Anderson (1999), Pham (2001), or by an effective combination of these.

In Figure 2, we plot the results of our experiment with the WRF estimator in the model system (2) for $N = 100$ and for $N = 10^4$. Notations are the same as in Figure 1. Just as the WEF method, the WRF scheme produces an estimate which converges to the optimal filter estimate as $N \to \infty$. See Crisan and Doucet (2002). Indeed, in Figure 2(b) we see that WRF with $N = 10^4$ is nearly indistinguishable from the optimal filter for the statistics considered. As expected, the result is better than that of WEF with the same number of samples and, in particular, the fluctuations due to small effective sample size are reduced. Furthermore, the WRF estimate for $N = 100$ shown in Figure 2(a) now succeeds in tracking the transition, where WEF failed to

do so at all. However, there is still a lag, just as for EnKF. The reason for this is easy to understand. Since a sequence of accurate measurements was taken at times $t = 1$ to 3 in the well at $x = +1$, all realizations "died out" in the other well at $x = -1$. That is, those realizations acquired small weights under Bayes' rule (1) and, consequently, were selected to survive with small probability upon resampling. After several such resamplings, the entire population of $N = 100$ samples resided in the well at $x = +1$. Then, despite the measurements indicating a transition at $t = 4 \sim 5$, there were simply no ensemble members in the opposite well at $x = -1$ to receive the higher weights under the Bayes' rule calculation. In consequence, the WRF method required a few of the samples in the well at $x = +1$ to dynamically make the transition to the opposite side, and this required additional time. The time was shortened compared with WEF, because the samples most likely to make the transition were preferentially resampled. As soon as a few samples moved to the -1 well, they received most of the weight at the next measurement and then repopulated that well. However, a lag-time still resulted in making the transition. This type of problem is likely to be generic when the WRF method is applied for moderate sample size N to nonlinear systems with multi-modal statistics.

2.3 Parametric Resampling Filter (PRF)

It is desirable to have a resampling method which can keep uniform weights, as WRF, but which avoids the difficulty of complete disappearance of realizations with small weights. A solution is to parametrically model the probability density, so that small probability events are preserved, even for

small values of N. In other words, the idea is to employ a *parametric density function* $P(\mathbf{x};\boldsymbol{\lambda})$ to approximate the real probability density function at measurement times t_m. We outline here a particular example of such a method, which we call the Parametric Resampling Filter (PRF), that uses an *exponential family* of probability density functions (PDF's). Full details of the method are outlined here in an Appendix. In this section we shall simply describe the method in its application to the model equation (2).

A convenient form for the parametric representation of the PDF in this case is

$$P(x;\boldsymbol{\lambda}) = \exp[\lambda_1 x + \lambda_2 x^2] \cdot Q(x),$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$. $Q(x)$ is an appropriate reference PDF, which is here chosen to be an equally weighted mixture of two Gaussians:

$$Q(x) = \frac{1}{2}[\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x+1)^2) + \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x-1)^2)],$$

with $\sigma^2 = \kappa^2/16$. This already gives a reasonable representation of the bimodal conditional PDF's of the model. Eyink *et al.* (2002a, b, c) use such a representation to carry out data assimilation for the model by moment-closure methods. In principle, convergence to the exact PDF's can be obtained by taking more Gaussians in the weighted sum for $Q(x)$, approaching the invariant measure $P_s(x)$. The basic idea is then: (i) to fit such a distribution $P(x;\boldsymbol{\lambda}_m^-)$ to the empirical ensemble of samples $x_n(t_m^-),\ n = 1, \cdots, N$ before the measurement, (ii) to apply Bayes' rule to generate a new parametric model distribution $P(x;\boldsymbol{\lambda}_m^+)$ subsequent to the measurement, and (iii) to sample this distribution to generate a new ensemble of samples $x_n(t_m^+),\ n = 1, ...N$ after the measurement.

Step (i) is achieved by determining the moments $\mu_i = \langle x^i \rangle$, $i = 1, 2$ empirically from the N-sample ensemble

$$\mu_i(t_m^-) = \frac{1}{N} \sum_{n=1}^{N} [x_n(t_m^-)]^i, \ i = 1, 2,$$

To fit the statistics by the parametric model, we must determine corresponding values of λ_1, λ_2. This is accomplished by carrying out the following maximization:

$$\boldsymbol{\lambda}_m^- = \text{argmax}_{\boldsymbol{\lambda}} \{\boldsymbol{\lambda}^\top \boldsymbol{\mu}_m^- - F(\boldsymbol{\lambda})\}$$

where $\boldsymbol{\mu}_m^- = (\mu_1(t_m^-), \mu_2(t_m^-))$ and $F(\boldsymbol{\lambda})$ is the cumulant generating function $F(\boldsymbol{\lambda}) = \ln \left[\int \exp(\lambda_1 x + \lambda_2 x^2) Q(x) dx\right]$, given explicitly below.

Step (ii) is carried out simply by setting $\lambda_1(t_m^+) = \lambda_1(t_m^-) + y_m/R_m$, $\lambda_2(t_m^+) = \lambda_2(t_m^-) - 1/(2R_m)$, because of the exponential form of the parametric representation.

Step (iii) is made simpler by completing squares in the parametric model to write

$$\begin{aligned} P(x; \boldsymbol{\lambda}) &= \frac{1}{\sqrt{2\pi\sigma^2(\boldsymbol{\lambda})}} [w_-(\boldsymbol{\lambda}) \exp(-\frac{1}{2\sigma^2(\boldsymbol{\lambda})}(x - \xi_-(\boldsymbol{\lambda}))^2) \\ &+ w_+(\boldsymbol{\lambda}) \exp(-\frac{1}{2\sigma^2(\boldsymbol{\lambda})}(x - \xi_+(\boldsymbol{\lambda}))^2)] \end{aligned} \tag{4}$$

where $w_\pm(\boldsymbol{\lambda}), \xi_\pm(\boldsymbol{\lambda}), \sigma(\boldsymbol{\lambda})$ are elementary functions of the parameters $\boldsymbol{\lambda}$:

$$w_\pm(\boldsymbol{\lambda}) = \frac{e^{\pm\lambda_1 \sigma^2(\boldsymbol{\lambda})/\sigma^2}}{2\cosh(\lambda_1 \sigma^2(\boldsymbol{\lambda})/\sigma^2)},$$

$$\xi_\pm(\boldsymbol{\lambda}) = \sigma^2(\boldsymbol{\lambda}) \left(\lambda_1 \pm \frac{1}{\sigma^2}\right),$$

and

$$\sigma^2(\boldsymbol{\lambda}) = \frac{\sigma^2}{1 - 2\lambda_2 \sigma^2}.$$

The convex "free energy" $F(\boldsymbol{\lambda})$ is found from the same calculation to be

$$F(\boldsymbol{\lambda}) = \frac{\sigma^2(\boldsymbol{\lambda})}{2\sigma^2}(\sigma^2\lambda_1^2 + 2\lambda_2) + \ln\left(\frac{\sigma(\boldsymbol{\lambda})}{\sigma}\right) + \ln\cosh\left(\frac{\lambda_1\sigma^2(\boldsymbol{\lambda})}{\sigma^2}\right).$$

Now the N samples can be generated for $n = 1, ..., N$ by first making a random selection of one of the Gaussians with probabilities $w_\pm(\boldsymbol{\lambda})$ and then, for the $\pm$ term selected, taking $\mathbf{x}_n = \xi_\pm(\boldsymbol{\lambda}) + \sigma(\boldsymbol{\lambda})\nu_n$. Here $\nu_n,\ n = 1, ..., N$ is a sequence of independent, normally distributed random variables with mean 0 and variance 1.

These formulas all have relatively simple generalizations to multidimensional systems, when the reference PDF is taken to be a weighted sum of Gaussians. See the Appendix for those formulas and computational details in the general case.

In Figure 3 we plot the results of the PRF method for our model system (2), with the double-Gaussian parametric PDF for $N = 100$ and $N = 10^4$. As for the WRF method, fluctuations are reduced by the resampling, relative to the WEF statistics at the same value of N. The results are not as good as those for WRF in the large ensemble with $N = 10^4$, but they could be improved by a parametric model with a mixture of more than two Gaussians. Even for the double-Gaussian PDF, the agreement is quite satisfactory. Furthermore, for $N = 100$ the results are the best of all the approximation methods considered, successfully tracking the transition, with relatively small fluctuations. The key to the success in following the measurements without a time lag is the presence of both Gaussians in the model, even when the weights $w_\pm(\boldsymbol{\lambda})$ of one or the other become small. As soon as measurements indicate a transition back to that side, Bayes' rule makes the corresponding weight large and resampling repopulates that well with many

samples. Note in this one-dimensional model that EOF's, SV's, bred vectors, etc. all coincide, and that simply generating samples in the wrong well by a superposition of "optimal growth" modes would not solve the problem. In that case, as for WRF, there would be a time-lag in tracking the transition, required for some samples to make the excursion to the correct well on the opposite side.

There is an additional advantage of the PRF method involving the quantity called *relative entropy* of a probability density P. With respect to the long-time "climate state" P_s, relative entropy is defined by

$$H(P|P_s) = \int d\mathbf{x}\ P(\mathbf{x}) \ln\left(\frac{P(\mathbf{x})}{P_s(\mathbf{x})}\right). \tag{5}$$

For example, see Kleeman (2002). Consider the entropy of a time sequence of PDF's $P(t)$ obtained by exactly solving the filtering problem for measurements at a discrete set of times $t_m,\ m = 1, ..., M$. Assuming climate statistics hold initially, then $H(t) = 0$ at $t = 0$ and it does not change until there is a difference of $P(t)$ from P_s due to information acquired from observations. When a measurement occurs, there is an upward jump in entropy whose size is a numerical measure of the information gain from or "utility" of the observations (see Kleeman, 2002). Between measurements $H(t)$ decreases back to zero because the PDF's $P(t)$ at large $t > 0$ converge again to P_s. This is an exact H-theorem for Markov stochastic processes like our model (2). Parametric models give an approximation to the entropy, as $H(\tilde{P}(t)|\tilde{P}_s)$, where $\tilde{P}(t) = P(\boldsymbol{\lambda}(t))$ and $\tilde{P}_s = P(\boldsymbol{\lambda}_s)$. Here, the parameters $\boldsymbol{\lambda}_s$ should be chosen so that moments of $P(\boldsymbol{\lambda}_s)$ match those of the exact P_s. It is straightforward to show for exponential families of parametric PDF's, as considered in the

PRF method, that

$$H(\tilde{P}(t)|\tilde{P}_s) = (\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_s)^\top \boldsymbol{\mu}(t) - (F(\boldsymbol{\lambda}(t)) - F(\boldsymbol{\lambda}_s)).$$

Hence, it is easy to calculate this approximation $\tilde{H}(t)$ in the PRF method, along with statistics such as the conditional mean and variance.

In Figure 4 we show results for the relative entropy in our model (2). Figure 4(a) depicts the exact entropy $H(t)$ calculated by formula (5) using the optimal filter PDF's $P(t)$ from the forward Kolmogorov equation and Bayes' rule. We see that there is a large information gain from the measurement at time $t = 4$, with a rapid loss immediately thereafter. In fact, if there were no subsequent measurements, then the entropy would drop close to zero by about time $t = 6$. Thus, with just one measurement indicating a transition from $+1$ to -1, one would lose predictability very quickly. However, with the additional measurements at times $t = 5, 6, 7$ indicating continuing residence in the -1 well, a different behavior is observed. In this case, there is a very slow relaxation back to zero at long times after the last measurement at time $t = 7$. The value of those final measurements is high, and the loss of prediction utility is very slow afterward. Figures 4(b) and 4(c) show the approximate entropy $\tilde{H}(t)$ calculated from the PRF method, with $N = 100$ in 4(b) and $N = 10^4$ in 4(c). They are quite similar to each other, except that the $N = 100$ result is rougher, as expected. Their magnitudes are also similar to the true entropy $H(t)$, although not exactly equal numerically. Most importantly, we see that $\tilde{H}(t)$ captures the large gain of information at time $t = 4$, followed by a rapid loss, and the slow decay at long times after the last measurement at time $t = 7$. These results indicate that the PRF method even for moderate values of N may yield a useful approximation to

the relative entropy, in cases where the statistics are far from Gaussian.

3 Conclusion and Remarks

In this paper we have discussed three ensemble filtering schemes, the Weighted Ensemble Filter (WEF), the Weight Resampling Filter (WRF), and the Parametric Resampling Filter (PRF). The latter, in particular, appears to be new. They differ from the well-known Ensemble Kalman Filter (EnKF) in that the analysis step at the measurement times is carried out by an approximate implementation of Bayes' theorem, not by a linear interpolation via a Kalman gain matrix. For a large enough number of samples, the Bayesian ensemble methods all can successfully track transitions in multimodal systems, where EnKF performs less well. For smaller numbers of samples, PRF performs the best of all the Bayesian methods in a simple test problem with bimodal statistics. This is an important consideration in practical applications, since the number of samples that can be employed at operational forecast centers is restricted. The PRF method also provides a convenient approximation of relative entropy, which measures the information made available from observations on the system.

All three of these Bayesian methods are potentially applicable to large-scale, spatially-extended systems. In addition to filtering and prediction, it is possible to perform smoothing as well using the same methods. Applied to a system with dynamics close to linear and statistics near-Gaussian, these new methods will yield results equivalent to those obtained with the Ensemble Kalman Filter. They should be most useful applied instead to geophysical systems with multiple attractors and multimodal statistics, or with other-

wise highly non-Gaussian distributions. Miller and Ehret (2002) give many examples of geophysical systems with multi-modal statistical distributions, for instance the Kuroshio current south of Japan. Other examples, such as the El Niño-Southern Oscillation (see Burgers and Stephenson, 1999), may have unimodal statistics but still exhibit other symptoms of non-Gaussianity, such as non-zero skewness or kurtosis. Non-normal distributions are to be expected for the nonlinear dynamical systems that govern most geophysical systems. Therefore, we believe that there will be important applications of the methods developed here to prediction of atmospheric and oceanographic systems.

Acknowledgments This work was carried out in part at Los Alamos National Laboratory under the auspices of the Department of Energy and supported by LDRD-ER 2000047. We also received support from NSF/ITR, Grant DMS-0113649 (SK, GLE, JMR), as well as from NASA, Goddard Space Flight Center, Grant NAG5-11163 (JMR).

4 Appendix

In this appendix we briefly outline the Parametric Resampling Method in the general case. Complete details appear in a forthcoming article by Eyink and Kim.

The method is based upon modeling the system statistics by an *exponential family* of probability density functions (PDF's):

$$P(\mathbf{x}, t_m; \boldsymbol{\lambda}) = \exp[\boldsymbol{\lambda}_1^\top \mathbf{h}_m(\mathbf{x}) + \mathbf{h}_m^\top(\mathbf{x}) \boldsymbol{\lambda}_2 \mathbf{h}_m(\mathbf{x})] \cdot Q(\mathbf{x}),$$

where $\mathbf{h}_m(\mathbf{x})$ is the measured F-vector at time t_m, and $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$, with $\boldsymbol{\lambda}_1$ an F-vector and $\boldsymbol{\lambda}_2$ a symmetric $F \times F$ matrix. $Q(\mathbf{x})$ is an appropriate reference PDF, which may be chosen, for example, to be a weighted sum of Gaussians. Most of the method does not depend upon any particular choice of $Q(\mathbf{x})$. As discussed in the text, the analysis at the measurement times $t_m,\ m = 1, ..., M$ is carried out in three steps:

Step (i): In order to determine $\boldsymbol{\lambda}_m^-$, we choose $P(\mathbf{x}; \boldsymbol{\lambda})$ to match the empirical ensemble $\mathbf{x}_n(t_m^-),\ n = 1, \cdots, N$ for the average of a moment-vector $\mathbf{M} = (\mathbf{M}_1, \mathbf{M}_2)$:

$$\int d\mathbf{x}\ \mathbf{M}(\mathbf{x}) P(\mathbf{x}; \boldsymbol{\lambda}_m^-) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{M}(\mathbf{x}_n(t_m^-))$$

where $\mathbf{M}_1(\mathbf{x}) = \mathbf{h}_m(\mathbf{x})$ and $\mathbf{M}_2(\mathbf{x}) = \mathbf{h}_m(\mathbf{x})\mathbf{h}_m^\top(\mathbf{x})$. Given the value on the RHS—call it $\boldsymbol{\mu}_m^-$—we can determine the corresponding value $\boldsymbol{\lambda}_m^-$ by the maximization

$$\boldsymbol{\lambda}_m^- = \text{argmax}_{\boldsymbol{\lambda}} \{\boldsymbol{\lambda}^\top \boldsymbol{\mu}_m^- - F(\boldsymbol{\lambda})\} \tag{6}$$

with $F(\boldsymbol{\lambda}) = \ln \left[\int d\mathbf{x}\ \exp(\boldsymbol{\lambda}^\top \mathbf{M}(\mathbf{x})) Q(\mathbf{x})\right]$ and $\boldsymbol{\lambda}^\top \boldsymbol{\mu} = \boldsymbol{\lambda}_1 \cdot \boldsymbol{\mu}_1 + \boldsymbol{\lambda}_2 : \boldsymbol{\mu}_2$.

Step (ii): After choosing $\boldsymbol{\lambda}^-$, we apply Bayes' rule to get updated $\boldsymbol{\lambda}^+ =$

$(\boldsymbol{\lambda}_1^+, \boldsymbol{\lambda}_2^+)$, viz:

$$\boldsymbol{\lambda}_1^+ = \boldsymbol{\lambda}_1^- + \mathbf{R}^{-1}\mathbf{y}$$
$$\boldsymbol{\lambda}_2^+ = \boldsymbol{\lambda}_2^- - \frac{1}{2}\mathbf{R}^{-1}$$

Step (iii): Draw N independent samples $\mathbf{x}_n(t_m^+)$, $n = 1, ...N$ from the new distribution $P(\mathbf{x}; \boldsymbol{\lambda}_m^+)$. This is the only step which depends upon a particular choice for the reference PDF. Various possibilities are discussed in a forthcoming article. The simplest case arises when the measured variables are linear in the state vector:

$$\mathbf{h}_m(\mathbf{x}) = \mathbf{d}_m + \mathbf{H}_m\mathbf{x},$$

where $\mathbf{d}_m$ is an F-vector and $\mathbf{H}_m$ is an $F \times D$ matrix. It is then very convenient to take Q as a sum of Gaussians:

$$Q(\mathbf{x}) = \sum_{s=1}^{S} w_s \frac{\exp\left[-\frac{1}{2}\mathbf{C}_s^{-1}\mathbf{:}(\mathbf{x}-\boldsymbol{\xi}_s)(\mathbf{x}-\boldsymbol{\xi}_s)\right]}{\sqrt{(2\pi)^D \mathrm{Det}\,(\mathbf{C}_s)}}$$

where $\sum_{s=1}^{S} w_s = 1$ and $\boldsymbol{\xi}_s, \mathbf{C}_s$ are the mean and covariance, respectively, of the sth Gaussian. The terms in the sum should be selected so that $Q(\mathbf{x})$ is a reasonable approximation to the steady-state measure $P_s(\mathbf{x})$, at least in applications to climate forecasting. Such a representation can be obtained empirically from conditional averaging over a long time-series of the dynamics, given some identified set of "statistical states" or "regimes" $s = 1, ..., S$ of the system. Otherwise, for more short-range forecasting, the reference PDF should be chosen as $Q(\mathbf{x}, t)$ to be explicitly time-dependent and to reflect the phase-space divergence of sample solutions on those times scales. We shall not further consider that problem here.

To resample from the PDF after conditioning upon measurements, we note that (4) can be generalized, under our stated assumptions, to:

$$P(\mathbf{x};\boldsymbol{\lambda}) = \sum_{s=1}^{S} w_s(\boldsymbol{\lambda}) \frac{\exp\left[-\frac{1}{2}\mathbf{C}_s^{-1}(\boldsymbol{\lambda}):(\mathbf{x}-\boldsymbol{\xi}_s(\boldsymbol{\lambda}))(\mathbf{x}-\boldsymbol{\xi}_s(\boldsymbol{\lambda}))\right]}{\sqrt{(2\pi)^D \mathrm{Det}\left(\mathbf{C}_s(\boldsymbol{\lambda})\right)}}$$

where $w_s(\boldsymbol{\lambda}), \boldsymbol{\xi}_s(\boldsymbol{\lambda}), \mathbf{C}_s(\boldsymbol{\lambda})$ will be given explicitly in the Eyink and Kim article. The N samples can be generated, for example, by first making a random selection of one of the Gaussians with probabilities $w_s(\boldsymbol{\lambda})$ and then, for the "regime" s_n selected, taking

$$\mathbf{x}_n = \boldsymbol{\xi}_{s_n}(\boldsymbol{\lambda}) + \sum_{d=1}^{D} \nu_{n,d}\sigma_{s_n,d}(\boldsymbol{\lambda})\mathbf{e}_{s_n,d}(\boldsymbol{\lambda})$$

for $n = 1, ..., N$. Here $\sigma^2_{s,d}(\boldsymbol{\lambda})$ are the eigenvalues of $\mathbf{C}_s(\boldsymbol{\lambda})$ and $\mathbf{e}_{s,d}(\boldsymbol{\lambda})$ the corresponding eigenvectors, for $s = 1, ..., S$, $d = 1, ..., D$, and $\nu_{n,d}$, $n = 1, ..., N$, $d = 1, ..., D$ is a sequence of independent, normally distributed random variables with mean 0 and variance 1.

In general, the resampling step (iii) is the most computationally intensive part of the analysis. Carried out as above, it requires diagonalizing the $D \times D$ matrices $\mathbf{C}_s(\boldsymbol{\lambda})$ for any required $\boldsymbol{\lambda}$. This will be too expensive to carry out in general. Instead, the forthcoming article will detail alternative procedures based upon Monte Carlo Markov Chain sampling methods, which require only diagonalizing the fixed set of $\boldsymbol{\lambda}$-independent matrices $\mathbf{C}_s$, $s = 1, ...S$. The corresponding eigenvectors $\mathbf{e}_{s,d}$, $s = 1, ..., S$, $d = 1, ..., D$ are the Conditional Empirical Orthogonal Functions (CEOF's) of the system, corresponding to each of its "statistical states" or "regimes". Alternatively, knowing $\mathbf{C}_s$, the corresponding Singular Vectors (SV's) may be constructed by the procedure discussed in Ehrendorfer and Tribbia (1997), (see also Miller and Ehret, 2002). For small sample sizes N, the generated ensemble may

better represent the PDF a finite time later if the expansion is given in terms of SV's rather than EOF's.

Except for step (iii) the entire procedure is carried out in a space of lower dimension than D. The minimization in step (i) is over $F + F(F + 1)/2 = F(F + 3)/2$ parameters. The Bayesian update in step (ii) and the calculation of $w_s(\boldsymbol{\lambda}), \boldsymbol{\xi}_s(\boldsymbol{\lambda}), \mathbf{C}_s(\boldsymbol{\lambda})$ in step (iii) are accomplished by linear algebra operations in the space of dimension F. Of course, these steps may also be nontrivial when also the number of measurements F is large.

References

Anderson, J.L., and S.L. Anderson, 1999: Monte Carlo implementation of the nonlinear filtering problem to produce ensemble asimilations and forecasts, *Mon. Wea. Rev*, **127**, pp. 2741-2758.

Burgers, G., and D. B. Stephenson, 1999: *The 'normality' of El Niño*, **26**, pp. 1027-1030.

Crisan, D., and A. Doucet, 2002: A survey of convergence results on particle filtering methods for practitioners, *IEEE Trans. Sig. Proc.*, **50**, pp. 736-746.

Ehrendorfer, M., and J. J. Tribbia, 1997: Optimal prediction of forecast error covariances through singular vectors, *J. Atmos. Sci.*, **54**, pp. 286-313.

Enachescu, D.M., 1985: Monte Carlo method for nonlinear filtering, in *Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics*, Mathematical Statistics and its Applications, Kluwer.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *J. Geophys. Res.*, **99**, (C5), pp. 10 143-10 162.

_____, and P. J. van Leeuwen, 2000: An ensemble Kalman smoother for nonlinear dynamics, *Mon. Weather Rev.*, **128**, pp. 1852-1867.

Eyink, G.L., J. M. Restrepo, and F. J. Alexander, 2002: A statistical-mechanical approach to data assimilation for nonlinear dynamics. I. Analysis approximations, `http://www.physics.arizona.edu/~restrepo/myweb/downloads/DOUB31-rev.ps.gz`

_____, _____, and _____, 2002: A statistical-mechanical approach to data assimilation for nonlinear dynamics. II. Evolution Approximations, `http://www.physics.arizona.edu/`

∼restrepo/myweb/downloads/DOUB32-rev.ps.gz

_____, _____, and _____, 2002: A statistical-mechanical approach to data assimilation for nonlinear dynamics. III. Numerical algorithms, http://www.physics.arizona.edu/

∼restrepo/myweb/downloads/DOUB33-rev.ps.gz

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

Hürzeler, M., and H.R. Künsch, 1998: Monte Carlo approximations for general state space models, *J. Comp. Graph. Stat.*, **7**, pp. 175-193.

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

_____, 1998: A self-organizing state-space model, *J. Am. Stat. Assoc.*, **93**, pp. 1203-1215.

_____, and W. Gersch, 1996: Smoothness Priors Analysis of Time Series, *Lecture Notes in Statistics*, **116**, Springer-Verlag, p. 261.

Kleeman, R., 2002: Measuring dynamical prediction utility using relative entropy, *J. Atmos. Sci.*, **59**, pp. 2057-2072.

Kushner, H.J., 1962: On the differential equations satisfied by conditional probability densities of Markov processes, with applications, *J. SIAM Control, Ser.A*, **2**, pp. 106-119.

Lorenc, A.C., and O. Hammon, 114: Objective quality control of observations using Bayesian methods. Theory, and a practical implementation, *Q. J. R. Meteorol. Soc.*, **114**, pp. 515-543.

Lorenz, E.N., 1963: Deterministic nonperiodic flow, *J. Atmos. Sci*, **20**, pp. 130-141.

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

_____, and L.L. Ehret, 2002: Ensemble generation for models of multimodal systems, *Mon. Wea. Rev.*, **130**, pp. 2313-2333.

Del Moral, P., 1996: Nonlinear filtering using random particles, *Theor. Prob. Appl.*, **40**, pp. 690-701.

Pham, D.T., 2001: Stochastic methods for sequential data assimilation in strongly nonlinear systems, *Mon. Wea. Rev.*, **129**, pp. 1194-1207.

Stratonovich, R.L., 1960: Conditional Markov processes, *Theor. Prob. Appl.*, **5**, pp. 156-178.

Ulam, S. and J. von Neumann, 1949: The Monte Carlo method, *J. Amer. Stat. Assoc.*, **44**, pp. 335-341.

Figure Captions

1. WEF method; (a) $N = 100$, (b) $N = 10^4$. Data shown as circles, WEF (solid lines), optimal solution (dashed lines).

2. WRF method; (a) $N = 100$, (b) $N = 10^4$. Data shown as circles, WRF (solid lines), optimal solution (dashed lines).

3. PRF method; (a) $N = 100$, (b) $N = 10^4$. Data shown as circles, PRF (solid lines), optimal solution (dashed lines).

4. Relative entropy, (a) for the optimal solution, and computed from the PRF method outcome with (c) $N = 100$, and (c) $N = 10^4$.

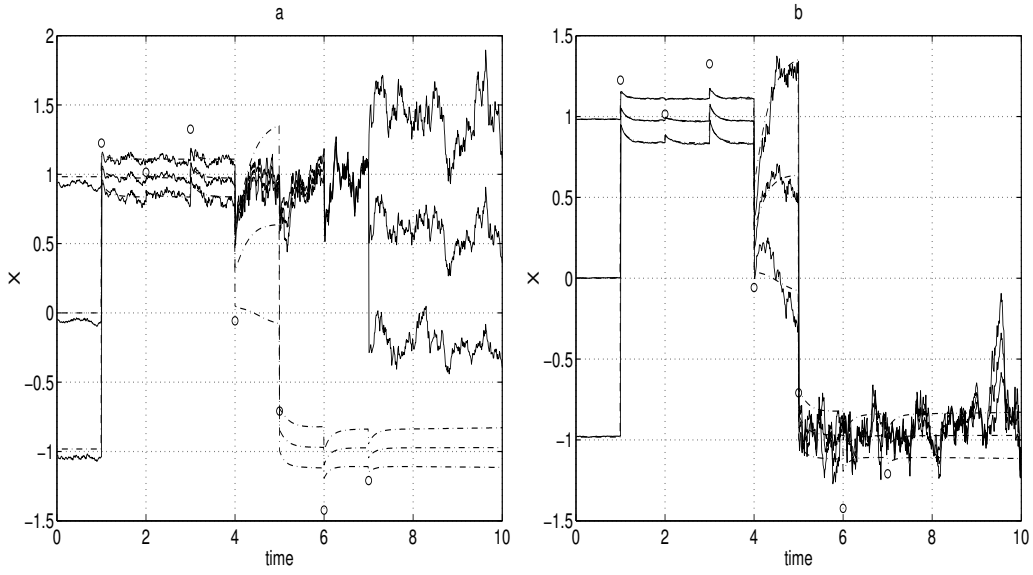

Figure 1: **1.** WEF method; (a) $N = 100$, (b) $N = 10^4$. Data shown as circles, WEF (solid lines), optimal solution (dashed lines).

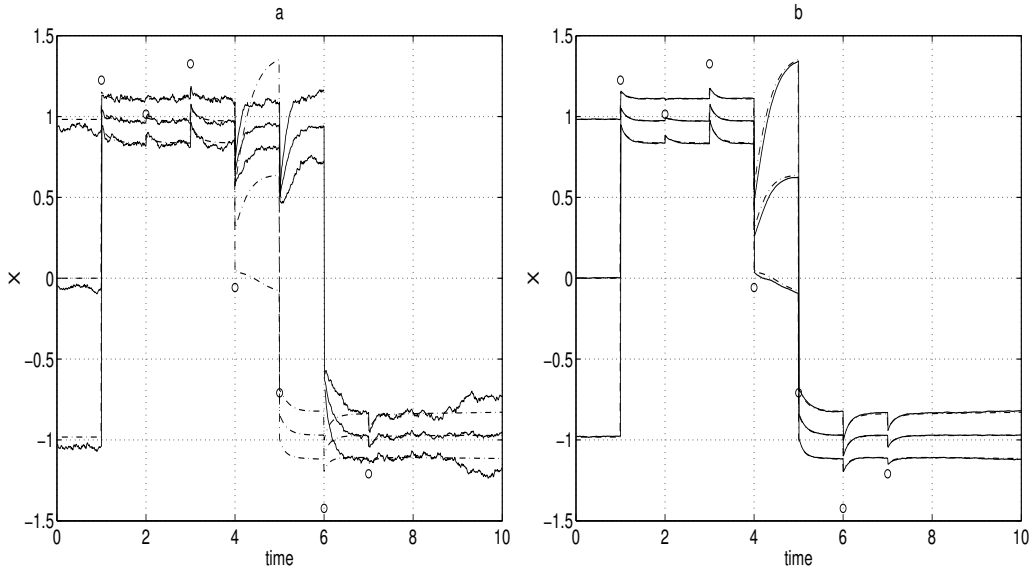

Figure 2: **2.** WRF method; (a) $N = 100$, (b) $N = 10^4$. Data shown as circles, WRF (solid lines), optimal solution (dashed lines).

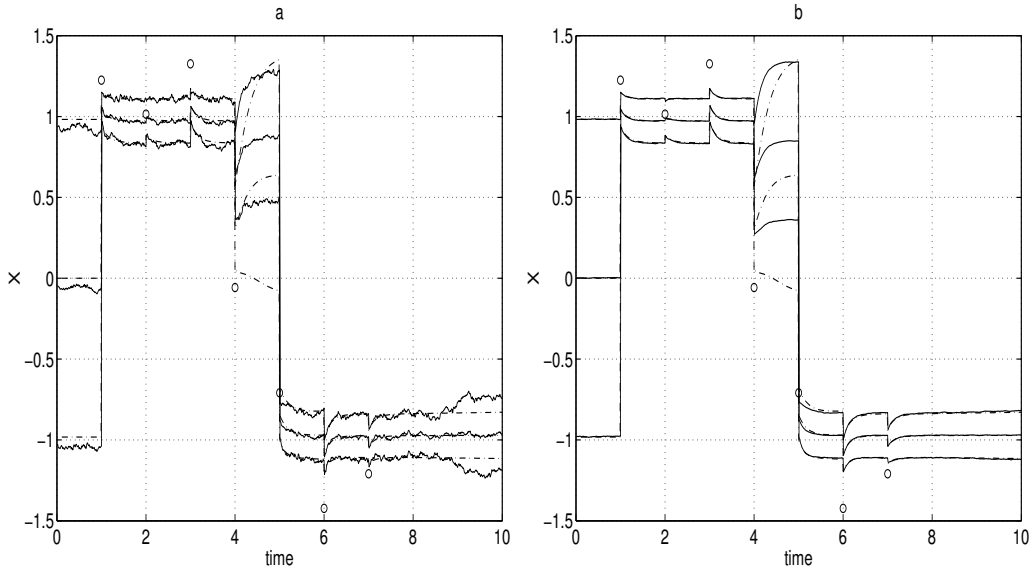

Figure 3: **3.** PRF method; (a) $N = 100$, (b) $N = 10^4$. Data shown as circles, PRF (solid lines), optimal solution (dashed lines).

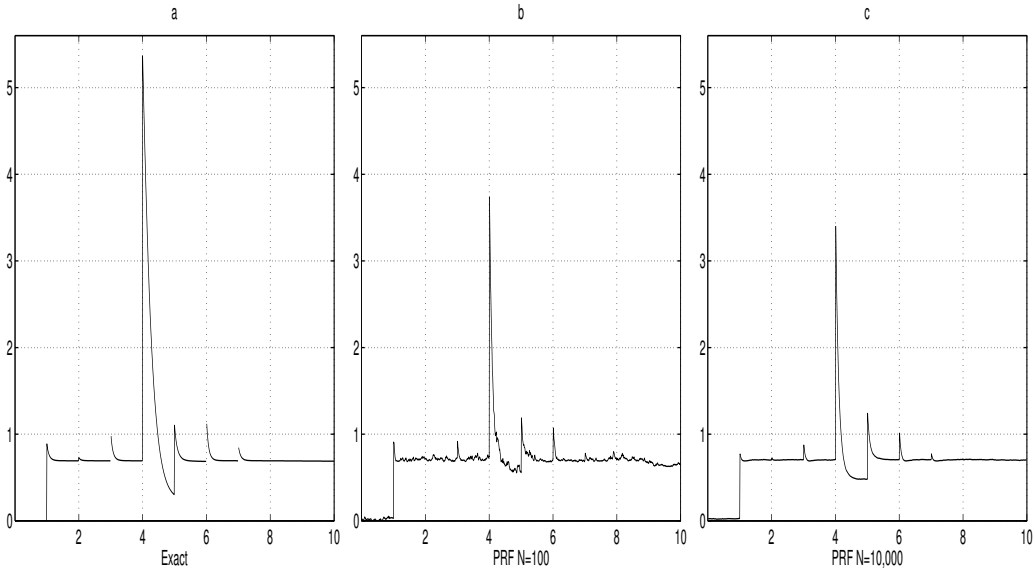

Figure 4: **4.** Relative entropy, (a) for the optimal solution, and computed from the PRF method outcome with (c) $N = 100$, and (c) $N = 10^4$.