P0

A Mean Field Approximation in Data Assimilation for Nonlinear Dynamics

Gregory L. Eyink [a] Juan M. Restrepo [b] Francis J. Alexander [c]

[a] *Mathematical Sciences Department, The Johns Hopkins University, Baltimore, MD 21218, U.S.A.*

[b] *Department of Mathematics and Department of Physics, University of Arizona, Tucson, AZ 85721, U.S.A.*

[c] *C C S 3, Los Alamos National Laboratory, Los Alamos, NM, 87545, U.S.A.*

Abstract

This paper considers the problem of data assimilation into nonlinear stochastic dynamic equations, from the point of view that the optimal solution is provided by the probabilities conditioned upon observations. An implementation of Bayes formula is described to calculate such probabilities. In the context of a simple model with multimodal statistics, it is shown that the conditional statistics succeed in tracking mode transitions where some standard suboptimal estimators fail. However, in complex models the exact conditional probabilities cannot be practically calculated. Instead, approximations to the conditional statistics must be sought. In this paper, attention is focused on approximations to the analysis step arising from the conditioning on observational data. A suboptimal mean-field conditional analysis is obtained from a statistical mechanics of time-histories. It is shown to have a variational formulation, reducing the approximate calculation of the conditional statistics to the minimization of the "effective action," a convex cost function. This mean-field analysis is compared with a standard linear analysis, based on a Kalman gain matrix. In the simple model problem, the mean-field conditional analysis is shown to approximate well the exact conditional statistics.

1 Introduction

It has been appreciated for some time that a probabilistic approach is necessary for data assimilation in the numerical modeling of ocean or climate dynamics or in numerical weather prediction. Leith long ago discussed the limits to theoretical skill in stochastic dynamic forecasting from the point of view of empirical ensembles of sample states of the model, distributed according to a probability law (see Leith, 1974). From that point of view, the proper

goal of any estimation scheme—whether for forecasting or for hindcasting—is to determine the probability distribution of the possible states of the system. Observational data from satellite stations or other measurements on the system change the prior statistical distribution. Lorenc and Hammon (Lorenc and Hammon, 1988) have discussed how such information from observations may be incorporated into the statistical characterization of initial conditions for weather forecasting by Bayesian probability methods. The Kalman filtering method and least-squares variational method, described in Courtier et al. (1993), provide convenient algorithms to calculate such conditional statistics for linear dynamical systems with additive, Gaussian error statistics.

More recently, it has been realized that systems with strongly nonlinear dynamics and/or multiplicative or non-Gaussian noise pose an especially challenging problem for data assimilation. Methods that were derived and validated for linear systems with Gaussian error statistics have been found often to fail there. For example, when the extended Kalman filter was applied to prediction in a two-layer quasigeostrophic model (see Evensen, 1992; Gauthier et al., 1993; Bouttier, 1994), it was found that the matrix Ricatti equation for the error covariance leads to unbounded error variance growth in the presence of an unstable background flow. Limitations on the growth corresponding to error saturation due to nonlinearities had to be put in by hand. Multi-modal, non-Gaussian statistics associated to multiple attractor regimes of the nonlinear dynamics also cause problems for such methods. In Miller et al. (1994) it was shown for a simple model with bimodal statistics (representing, for example, an "ice age" and a "normal climate" state), that the extended Kalman filter and least-squares variational method both fail to detect state-transitions observed in the data, when the measurements are not very accurate or very dense in time. Only the inclusion of high-order moment statistics or of empirically-determined noise statistics were found to solve the problem in those methods.

Such difficulties do not exist at all if the conditional statistics of the system are correctly calculated. In Miller et al. (1999) it was shown that transitions observed in the data are indeed reflected in the conditional probability distributions for the simple model problems studied earlier by Miller et al. (1994), at much lower accuracy and frequency of measurements than for the suboptimal "linearization" methods. The required conditional distributions were calculated in Miller et al. (1999) by solving partial differential equations on the state space of the system (also, see Campillo et al., 1993). Such methods also solve the difficulty with unstable error growth observed in the linearization methods, because the equations for the probability distributions relax at long times to the stationary statistics or climate state of the model. Thus, error statistics will naturally saturate due to nonlinear effects, as discussed long ago by Leith (1974). Unfortunately, while such methods are conceptually correct and efficacious, they cannot be applied to the realistic, spatially-extended sys-

tems of interest in the geosciences. For such large-scale systems the equations for the full probability distributions will not be able to be solved numerically by computers for the foreseeable future.

This fact has occasioned the development of methods to approximate the required conditional statistics. Monte Carlo methods—as originally advocated by Leith (1974)—have in particular been actively developed by G. Evensen and his collaborators, (see Evensen, 1994; van Leeuwen and Evensen, 1996; Burgers et al., 1998; Evensen and van Leeuwen, 2000). More recent developments are reviewed in Tippett et al. (2003). Such methods solve the nonlinear dynamics for an empirical set of N samples, with $N = O(10^2)$, to approximate the statistical distribution of an infinite ensemble. Although it is not easy to assess the size of the errors incurred, such methods provide a rather effective means to approximate the evolution of the probabilities for nonlinear dynamics. It was shown in Evensen (1994) that the saturation of error growth for a quasi-geostrophic model was naturally achieved by such a method. However, problems remain. Miller et al. 1999 have found that such methods still fail to properly track transitions in systems with multimodal statistics. The failure can be traced to the linear interpolation scheme presently employed in such methods for the "analysis" of forecast and observational data (see Evensen and van Leeuwen, 2000). This interpolation scheme is a holdover of the Kalman filtering methods justified for linear dynamics with Gaussian errors and does not represent a correct implement of Bayes formula in general. It is an open research problem to develop effective methods to approximate conditional statistics for large-scale geophysical systems

This paper proposes a method for the practical calculation of conditional probabilities for large-scale, nonlinear dynamical systems. Specifically, this paper focuses on the "analysis step", or the modification in the statistics brought about by the conditioning upon available observational data. In many respects, this is the crucial part of the problem. We propose here a new "mean-field conditional analysis" which performs the conditioning upon observations in an approximate manner. It has a convenient variational formulation, which reduces the calculation of the conditional statistics to the minimization of a certain cost function, the so-called "effective action". It is still not practical to apply this "mean-field analysis" directly to large-scale systems, but it becomes a practical method within various approximate methods for evolving the statistics under the nonlinear dynamics, e.g. moment-closure methods. The mathematical theory underlying this paper is developed at greater length in Eyink (2002) and a brief description of its application has already been given in Eyink and Restrepo (2000).

The detailed contents of this paper are as follows: In Section 2 we carefully outline the problem of state estimation in its statistical formulation. We also discuss there the exact calculation of probability distributions conditioned

upon the full set of observations, both past and future, by means of suitable partial differential equations. The "KSP method" discussed there generalizes that of Miller et al. (1999); Campillo et al. (1993), who conditioned only upon past measurements and not future ones. We use the results of such a calculation as the main basis of comparison for our approximations and we argue that they should be so used more generally. This KSP method thus has some independent interest, apart from the specific approximate methods proposed in this work. The latter are introduced in Section 3, where the relevant statistical mechanics of time-histories is discussed. The "mean-field conditional analysis" which we propose is also compared theoretically in this section with the more standard "linear analysis" which is employed in most existing data assimilation schemes. In Section 4 we test our approximate analysis method on the same simple model previously considered in Miller et al. (1994, 1999). We calculate both the exact conditional statistics and our approximate conditional statistics for several sets of "measurements" on a history of the model, and the results are compared in detail. Section 5 contains our summary discussion and conclusion.

2 Probabilistic Formulation

2.1 General Statement of the Problem

Consider a nonlinear model dynamics for state vector $\mathbf{x}(t)$ given by

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, t) + \mathbf{D}^{1/2}(\mathbf{x}, t)\mathbf{q}(t). \tag{1}$$

The vector $\mathbf{q}(t)$ is a white-noise with zero mean and covariance $\langle q_i(t)q_j(t')\rangle = 2\delta(t - t')$. It represents noise from neglected degrees of freedom, or "model error." $\mathbf{D}$ is a covariance matrix giving the strength of the noise. We include as a special case $\mathbf{D} \equiv \mathbf{0}$, i.e. no model error. The initial conditions $\mathbf{x}_0$ are taken to be random, with a known distribution $P_0(\mathbf{x})$. If the dynamics are given by a partial differential equation, then there will be boundary conditions with possible randomness as well. Observations $\mathbf{y}_m$ are taken of a linear function $\mathbf{h}(\mathbf{x}, t_m) = \mathbf{H}(t_m)\mathbf{x}$, including some measurement errors $\boldsymbol{\rho}_m$ with covariance $\mathbf{R}_m$:

$$\mathbf{y}_m = \mathbf{h}(\mathbf{x}(t_m), t_m) + \boldsymbol{\rho}_m, \quad m = 1, ..., M. \tag{2}$$

It will be assumed that the distribution of the measurement errors is known as well, e.g. Gaussian. The problem is to determine the best estimate of the

state history $\mathbf{x}(t)$ given the measurements, and, as well, to obtain a measure of the uncertainty in this estimate. We have made here several simplifying assumptions which are by no means necessary, such as a linear measurement function $\mathbf{h}(t_m)$ and Gaussian-distributed observation errors. However, these allow us to illustrate our methods in the simplest context.

The optimal solution of this problem is provided by the *conditional statistics*, given the measurements. Thus, the conditional mean

$$\mathbf{x}_S(t) = \mathsf{E}[\mathbf{x}(t)|\mathbf{y}_1, ..., \mathbf{y}_M] \tag{3}$$

is the best estimate of the state, and the conditional covariance matrix (where $\top$ denotes transpose)

$$\mathbf{C}_S(t) = \mathsf{E}[(\mathbf{x}(t) - \mathbf{x}_S(t))(\mathbf{x}(t) - \mathbf{x}_S(t))^\top|\mathbf{y}_1, ..., \mathbf{y}_M] \tag{4}$$

provides a measure of its uncertainty. Of all estimators, the conditional mean $\mathbf{x}_S(t)$ is distinguished as the one which minimizes $\mathrm{tr}\mathbf{C}_S(t) = \mathsf{E}[|(\mathbf{x}(t)-\mathbf{x}_S(t))|^2|\mathbf{y}_1, ..., \mathbf{y}_M]$, i.e. the trace of the conditional covariance matrix. It is the variance-minimizing estimator, or *smoother* estimate. A corresponding set of statistics using only the currently available set of measurements from prior times,

$$\mathbf{x}_F(t) = \mathsf{E}[\mathbf{x}(t)|\mathbf{y}_1, ..., \mathbf{y}_k] \tag{5}$$

and

$$\mathbf{C}_F(t) = \mathsf{E}[(\mathbf{x}(t) - \mathbf{x}_F(t))(\mathbf{x}(t) - \mathbf{x}_F(t))^\top|\mathbf{y}_1, ..., \mathbf{y}_k] \tag{6}$$

with k chosen so that $t_{k+1} > t \geq t_k$, is called the *filter* estimate.

2.2 Exact Bayesian Solution: Evolution & Analysis

The optimal filtering problem in the above general setting has been solved exactly by Stratonovich (1960), and Kushner (1962, 1967b) within a Bayesian formulation. We define the conditional probability density

$$P_F(\mathbf{x}, t) = P(\mathbf{x}, t|\mathbf{y}_1, ..., \mathbf{y}_k),$$

given the current set of measurements $\mathbf{y}_1, ..., \mathbf{y}_k$, with $t_{k+1} > t \geq t_k$. This filter distribution is obtained as follows. Starting from the initial condition $P_0(\mathbf{x})$

at time $t_0 < t_1$ and between measurement times, $P_F(\mathbf{x}, t)$ solves the forward Kolmogorov equation (see Risken, 1984)

$$\partial_t P_F(\mathbf{x}, t) = \hat{L}(t) P_F(\mathbf{x}, t), \tag{7}$$

where $\hat{L}(t) = -\boldsymbol{\nabla}_{\mathbf{x}} \cdot [\mathbf{f}(\mathbf{x}, t)(\cdot)] + \boldsymbol{\nabla}_{\mathbf{x}} \boldsymbol{\nabla}_{\mathbf{x}}^{\top} : [\mathbf{D}(\mathbf{x}, t)(\cdot)]$ is the Fokker-Planck operator. At measurement times t_m, $m = 1, ..., M$, $P_F(\mathbf{x}, t)$ satisfies the forward "jump condition"

$$P_F(\mathbf{x}, t_m+) = \frac{\exp\left[\mathbf{y}_m^{\top} \mathbf{R}_m^{-1} \mathbf{h}(\mathbf{x}, t_m) - \frac{1}{2} \mathbf{h}^{\top}(\mathbf{x}, t_m) \mathbf{R}_m^{-1} \mathbf{h}(\mathbf{x}, t_m)\right]}{W(\mathbf{y}_1, ..., \mathbf{y}_m)} P_F(\mathbf{x}, t_m-), \tag{8}$$

where $-, +$ denote times just before and after the measurement, respectively. $W(\mathbf{y}_1, ..., \mathbf{y}_m)$ is the normalization factor that ensures that $P_F(\mathbf{x}, t_m+)$ integrates to one. Notice that measurements are used sequentially to obtain the filter distribution $P_F(\mathbf{x}, t)$. Once the conditional distribution is known, its first two moments, $\mathbf{x}_F(t) = \int d\mathbf{x}\ \mathbf{x} P_F(\mathbf{x}, t)$ and $\mathbf{C}_F(t) = \int d\mathbf{x}\ (\mathbf{x} - \mathbf{x}_F(t))(\mathbf{x} - \mathbf{x}_F(t))^{\top} P_F(\mathbf{x}, t)$, give the filter mean and covariance.

The optimal smoother distribution $P_S(\mathbf{x}, t)$ is similarly obtained, by an adjoint algorithm due to Pardoux (1982): Starting from a final condition $A_S(\mathbf{x}, t_f) = 1$ at a time $t_f > t_M$ and in reverse in time between measurements, one solves the backward Kolmogorov equation (see Risken, 1984)

$$\partial_t A_S(\mathbf{x}, t) + \hat{L}^*(t) A_S(\mathbf{x}, t) = 0, \tag{9}$$

in which $\hat{L}^*(t) = \mathbf{f}(\mathbf{x}, t) \cdot \boldsymbol{\nabla}_{\mathbf{x}} + \mathbf{D}(\mathbf{x}, t) : \boldsymbol{\nabla}_{\mathbf{x}} \boldsymbol{\nabla}_{\mathbf{x}}^{\top}$ is the adjoint Fokker-Planck operator. At measurement times t_m, $m = 1, ..., M$ the backward "jump condition" is imposed:

$$A_S(\mathbf{x}, t_m-) = A_S(\mathbf{x}, t_m+) \frac{\exp\left[\mathbf{y}_m^{\top} \mathbf{R}_m^{-1} \mathbf{h}(\mathbf{x}, t_m) - \frac{1}{2} \mathbf{h}^{\top}(\mathbf{x}, t_m) \mathbf{R}_m^{-1} \mathbf{h}(\mathbf{x}, t_m)\right]}{W(\mathbf{y}_1, ..., \mathbf{y}_m)}. \tag{10}$$

Here $W(\mathbf{y}_1, ..., \mathbf{y}_m)$ is the same normalization factor as determined for the forward jump condition, which must be stored for use in the backward evolution. The distribution $P_S(\mathbf{x}, t) = P(\mathbf{x}, t | \mathbf{y}_1, ..., \mathbf{y}_M)$ conditioned on the entire set of available measurements is finally obtained from the product

$$P_S(\mathbf{x}, t) = A_S(\mathbf{x}, t) P_F(\mathbf{x}, t). \tag{11}$$

Jump conditions (8),(10) together imply that $P_S(\mathbf{x}, t)$ is continuous in time. The moments $\mathbf{x}_S(t) = \int d\mathbf{x}\ \mathbf{x} P_S(\mathbf{x}, t)$ and $\mathbf{C}_S(t) = \int d\mathbf{x}\ (\mathbf{x} - \mathbf{x}_S(t))(\mathbf{x} -$

$\mathbf{x}_S(t))^\top P_S(\mathbf{x},t)$ give the smoother mean and covariance.

The solution algorithm outlined above will be termed the Kushner-Stratonovich-Pardoux (KSP) method, since those authors first developed it instead for the (more difficult) case of continuous-time measurements. The simpler case of discrete-time measurements, discussed above, was treated in Jazwinski (1970) for the filtering part of the algorithm and in Appendix I of Eyink (2002) for the smoothing part. This method provides not only the conditional mean and covariance, $\mathbf{x}_S(t)$ and $\mathbf{C}_S(t)$, but even the entire conditional distribution $P_S(\mathbf{x},t)$ instantaneously at time t. This is almost all that one could wish. However, this is not so for certain purposes. A conditional mean history $\mathbf{x}_S(t)$ could be very atypical and unrealistic, as was emphasized long ago by Leith (1974). Its properties might be very misleading as an indicator of the behavior of individual realizations. For some purposes, it might be useful instead to have a way of selecting some representatives from an ensemble of histories conditioned on the measurements. Techniques for doing so can be developed using methods related to those in this paper (see Alexander et al., 2002), but this will not be discussed here.

However, the KSP method, while giving the optimal solution of the problem, is computationally intractable when applied to realistic spatially-extended systems with many degrees of freedom. This was already pointed out long ago by Kushner (1967a) himself. When (1) is a partial differential equation (PDE), then the KSP forward and backward equations (7),(9) are *functional* differential equations for a solution which is a distribution on a function space. There are a few exceptional situations where the optimal smoother is "finite-dimensional" and its calculation reduces to solving PDE's of a comparable number of degrees of freedom as (1) itself. For example, when (1) is linear and the model noise is additive, then the conditional distributions are Gaussian and the means and covariances are equivalent to those obtained from the Kalman filter and smoother. The latter require solving equations only of the same dimension as (1) for the means and of the square of the dimension for the covariance. However, in general there is no such exact simplification, and the KSP method is impossible to apply without some simplifying approximation.

On the other hand, the KSP method gives the correct standard of comparison for approximation methods, in the simple systems where it can be applied. This point has already been made in a recent paper by Miller et al. (1999), where the Kushner-Stratonovich filter was calculated for some simple models: the double-well model, which is also considered in this work, and the 3-mode, chaotic Lorenz model. This is also the point of view adopted here. We shall compare results of all of our approximation schemes with the exact conditional statistics provided by KSP. In fact, we wish to emphasize the importance of this comparison. It is common practice to judge the "success" of a data assimilation scheme based upon its ability to recover a single, particular realization,

after the values of the latter at a few points have been contaminated with random errors and measured. In fact, this is a faulty test of the success of an estimation scheme whose goal is the calculation of conditional statistics. An assimilation method which, for a particular realization, happened to reproduce it better, might actually be inferior to one which gave a result further from that realization but closer to the conditional average. For systems with random noise or with deterministic chaos it is impossible in principle to set the goal of recovering each individual realization from partial and imperfect information about it. It is only meaningful to search for statistical information about the system: a variance-minimizing estimate and a measure of the possible spread in the ensemble. In the case of large, spatially-extended systems where it is impossible to calculate such conditional statistics exactly by KSP, it is only possible to compare the results of different approximation schemes with each other. Only in this way can it be determined if the results of any (or none) of the approximations is likely to be accurate.

The KSP method can also be a guide to constructing suitable approximation schemes, because it provides itself the correct optimal solution to the problem. It therefore gives some idea of the ingredients which must go into any successful approximation. We see that the KSP calculation algorithm divides neatly (for discrete-time observations) into two distinct elements: dynamical evolution provided by the Kolmogorov equations (7),(9) and statistical conditioning provided by the "jump conditions" (8),(10) at the measurement times. In the traditional terminology of data assimilation, the latter is called the "analysis" of the estimate and the observation. In this paper we consider approximation schemes for the analysis step of the estimation problem.

3 Analysis Approximations

3.1 Mean-Field Conditional Analysis & Statistical-Mechanics of Histories

Rather than imposing the observations as exact conditions, one can instead employ them in a mean-field manner. Let $\{\mathbf{x}^{(n)}(t) : t \in [t_0, t_f]\}]$, for $n = 1, ..., N$, be an ensemble of solution histories of the equation (1) for N initial data $\mathbf{x}_0^{(n)}$, $n = 1, ..., N$, chosen independently from the distribution P_0. Let

$$\overline{\mathbf{x}}^N(t) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}^{(n)}(t) \tag{12}$$

be the *empirical ensemble-average* formed from the N independent sample realizations. Likewise, form $\overline{\boldsymbol{\rho}}_m^N = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\rho}_m^{(n)}$, an N-sample average of inde-

pendent measurement errors for each ensemble realization at time t_m. Then

$$\overline{\mathbf{y}}^N(t_m) = \mathbf{H}(t_m)\overline{\mathbf{x}}^N(t_m) + \overline{\boldsymbol{\rho}}_m^N \tag{13}$$

is an N-sample average measurement. We take as a suboptimal smoother estimate the conditional mean

$$\mathbf{x}_*(t) = \lim_{N\to\infty} \mathsf{E}[\mathbf{x}(t)|\overline{\mathbf{y}}^N(t_1) = \mathbf{y}_1, ..., \overline{\mathbf{y}}^N(t_M) = \mathbf{y}_M] \tag{14}$$

and the conditional covariance matrix

$$\begin{array}{l} \mathbf{C}_*(t) = \\ \lim_{N\to\infty} \mathsf{E}[(\mathbf{x}(t) - \mathbf{x}_*(t))(\mathbf{x}(t) - \mathbf{x}_*(t))^\top|\overline{\mathbf{y}}^N(t_1) = \mathbf{y}_1, ..., \overline{\mathbf{y}}^N(t_M) = \mathbf{y}_M]. \end{array} \tag{15}$$

We emphasize that, in our mean-field approximation, $\mathbf{y}_1, ..., \mathbf{y}_M$ are the actual values obtained in a single set of measurements, not in an ensemble of such measurements. N-sample ensembles are only introduced in this approximation scheme for theoretical purposes and are never employed in its practical implementation. An advantage of this approximation is that the conditional mean (14) and covariance (15) can be obtained from a thermodynamic formalism, as the minimizer and inverse Hessian, respectively, of a certain "entropy function" $H_*(\mathbf{x}_1, ..., \mathbf{x}_M)$.

We explain briefly the relevant statistical mechanics on time-histories. Details may be found in Eyink (2002). Let us denote by $P[\{\mathbf{x}(t) : t \in [t_0, t_f]\}]$ the distribution on path-space of the entire history. It is formally given by a path-integral formula (cf. van Leeuwen and Evensen (1996), Eq.(15)):

$$\begin{array}{l} P[\{\mathbf{x}(t) : t \in [t_0, t_f]\}] \propto \\ \exp\left\{-\frac{1}{4}\int_{t_0}^{t_f} dt\ [\dot{\mathbf{x}}(t) - \mathbf{f}(\mathbf{x}, t)]^\top \mathbf{D}^{-1}(\mathbf{x}, t)[\dot{\mathbf{x}}(t) - \mathbf{f}(\mathbf{x}, t)]\right\}. \end{array} \tag{16}$$

For N independent samples of the process, the distribution is given by the product-measure $P^{\otimes N}[\{\mathbf{x}^{(1)}(t), ..., \mathbf{x}^{(N)}(t) : t \in [t_0, t_f]\}] = \prod_{n=1}^N P[\{\mathbf{x}^{(n)}(t) : t \in [t_0, t_f]\}]$. The N-sample distribution conditioned on empirical sample-means at a sequence of times,

$$P^{\otimes N}[\{\mathbf{x}^{(1)}(t), ..., \mathbf{x}^{(N)}(t) : t \in [t_0, t_f]\}|\overline{\mathbf{x}}^N(t_1) = \mathbf{x}_1, ..., \overline{\mathbf{x}}^N(t_M) = \mathbf{x}_M], \tag{17}$$

is analogous to a "microcanonical distribution" in equilibrium statistical mechanics. In the limit $N \to \infty$ it becomes equivalent to a corresponding "canon-

ical distribution" of product-measure form:

$$\prod_{n=1}^{N} P[\{\mathbf{x}^{(n)}(t) : t \in [t_0, t_f]\}; \boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M] \tag{18}$$

where the factors are

$$\begin{aligned} &P[\{\mathbf{x}(t) : t \in [t_0, t_f]\}; \boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M] = \\ &\frac{\exp[\sum_{m=1}^{M} \boldsymbol{\lambda}_m^\top \mathbf{x}(t_m)]}{N_X(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M)} P[\{\mathbf{x}(t) : t \in [t_0, t_f]\}] \end{aligned} \tag{19}$$

and $N_X(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M)$ is the normalization integral to ensure total unit probability. The appropriate values of $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M$ are determined by a thermodynamic argument. Define a convex *cumulant generating function*

$$\begin{aligned} F_X(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M) :&= \log N_X(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M) \\ &= \log\langle \exp[\sum_{m=1}^{M} \boldsymbol{\lambda}_m^\top \mathbf{x}(t_m)]\rangle, \end{aligned} \tag{20}$$

analogous to the "free-energy" in equilibrium statistical mechanics. Its pth-order partial derivatives are the *pth-order cumulants* of the random variables $\mathbf{x}(t_1), ..., \mathbf{x}(t_M)$. In particular, $\mathbf{x}_m = \frac{\partial F_X}{\partial \boldsymbol{\lambda}_m}(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M)$ is the mean of $\mathbf{x}(t_m)$ and $\mathbf{C}(t_m, t_{m'}) = \frac{\partial^2 F_X}{\partial \boldsymbol{\lambda}_m \partial \boldsymbol{\lambda}_{m'}}(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M)$ the covariance of $\mathbf{x}(t_m), \mathbf{x}(t_{m'})$ in the canonical distribution. The Legendre transform

$$H_X(\mathbf{x}_1, ..., \mathbf{x}_M) = \max_{\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M} \left\{ \sum_{m=1}^{M} \mathbf{x}_m^\top \boldsymbol{\lambda}_m - F_X(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M) \right\} \tag{21}$$

is called the *multi-time (relative) entropy*. The values of $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M$ which appear in the "thermodynamic limit" of the microcanonical distribution are those at which the maximum in (21) is achieved. Equivalently,

$$\boldsymbol{\lambda}_m = \frac{\partial H_X}{\partial \mathbf{x}_m}(\mathbf{x}_1, ..., \mathbf{x}_M), \quad m = 1, ..., M. \tag{22}$$

The minimum of the convex entropy H_X occurs for the mean values $\overline{\mathbf{x}}_1, ..., \overline{\mathbf{x}}_M$ of $\mathbf{x}(t_1), ..., \mathbf{x}(t_M)$ in the original distribution (with all $\boldsymbol{\lambda}_1 = \cdots = \boldsymbol{\lambda}_M = \mathbf{0}$). The entropy is also a generating function for so-called *irreducible correlation functions* of $\mathbf{x}(t_1), ..., \mathbf{x}(t_M)$. In particular, $\boldsymbol{\Gamma}(t_m, t_{m'}) = \frac{\partial^2 H_X}{\partial \mathbf{x}_m \partial \mathbf{x}_{m'}}(\overline{\mathbf{x}}_1, ..., \overline{\mathbf{x}}_M)$ is the inverse of the covariance matrix $\mathbf{C}(t_m, t_{m'})$.

The conditional mean $\mathbf{x}_*(t_m)$ and conditional covariance matrix $\mathbf{C}_*(t_m, t_{m'})$ from (14),(15) can be obtained similarly as the minimizer and inverse Hessian,

respectively, of a joint entropy $H_*(\mathbf{x}_1, ..., \mathbf{x}_M) = H_{X,Y}(\mathbf{x}_1, ..., \mathbf{x}_M; \mathbf{y}_1, ..., \mathbf{y}_M)$ for both state variables and observations:

$$\begin{aligned} &H_*(\mathbf{x}_1, ..., \mathbf{x}_M) = H_X(\mathbf{x}_1, ..., \mathbf{x}_M) + \\ &\tfrac{1}{2} \textstyle\sum_{m=1}^{M} [\mathbf{y}_m - \mathbf{H}(t_m)\mathbf{x}_m]^\top \mathbf{R}_m^{-1} [\mathbf{y}_m - \mathbf{H}(t_m)\mathbf{x}_m]. \end{aligned} \tag{23}$$

The additional contribution to the cost function arising from the measurements is quadratic because of the assumptions of Gaussian-distributed observation errors and of linear measurements. Our goal is to calculate H_*, to minimize it, and to calculate its Hessian.

We must first calculate F_X and then H_X. It turns out that there is an algorithm to do this, based upon the forward and backward Kolmogorov equations (7),(9). To calculate $F_X(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M)$ we need only to solve the forward equation (7). At measurement times, the solution satisfies jump conditions

$$P(\mathbf{x}, t_m+) = \frac{e^{\boldsymbol{\lambda}_m^\top \mathbf{x}}}{W(t_m-)} P(\mathbf{x}, t_m-), \quad m = 1, ...M, \tag{24}$$

where $W(t_m-)$ is the normalization integral. From it we form

$$\begin{aligned} (\Delta F)_m(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_m) &:= \log W(t_m-) \\ &= \log \left(\int d\mathbf{x}\ e^{\boldsymbol{\lambda}_m^\top \mathbf{x}} P(\mathbf{x}, t_m-) \right). \end{aligned} \tag{25}$$

Whereas the dependence upon $\boldsymbol{\lambda}_m$ is explicit, note that the dependence upon the remaining variables $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_{m-1}$ is only implicit through $P(\mathbf{x}, t_m-)$. Finally, we obtain

$$F_X(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M) = \sum_{m=1}^{M} (\Delta F)_m(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_m) \tag{26}$$

by summing up the contributions from each of the measurement times. Having determined $F_X(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M)$, the entropy $H_X(\mathbf{x}_1, ..., \mathbf{x}_M)$ can then be obtained by carrying out the maximization in the Legendre transform formula (21). To do so by a descent algorithm requires having also the derivatives

$$\mathbf{x}_m = \frac{\partial F_X}{\partial \boldsymbol{\lambda}_m}(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M), \quad m = 1, ..., M. \tag{27}$$

This may be calculated by an adjoint algorithm using the backward Kolmogorov equation (9). At measurement times, the solution now satisfies jump

conditions

$$A(\mathbf{x}, t_m-) = \frac{e^{\boldsymbol{\lambda}_m^\top \mathbf{x}}}{W(t_m-)} A(\mathbf{x}, t_m+). \tag{28}$$

We obtain finally $\mathbf{x}_m := \mathbf{x}(t_m; \boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M),\ m = 1, ..., M$, with the latter given by

$$\mathbf{x}(t; \boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M) = \int d\mathbf{x}\ \mathbf{x} A(\mathbf{x}, t) P(\mathbf{x}, t). \tag{29}$$

for all times $t \in [t_i, t_f]$. For the values $\boldsymbol{\lambda}_1^*, ..., \boldsymbol{\lambda}_M^*$ corresponding to the minimizer $\mathbf{x}_1^*, ..., \mathbf{x}_M^*$ of $H_*(\mathbf{x}_1, ..., \mathbf{x}_M)$, then $\mathbf{x}_*(t) = \mathbf{x}(t; \boldsymbol{\lambda}_1^*, ..., \boldsymbol{\lambda}_M^*)$ gives the smoother estimate at all times.

It is sometimes advantageous to formulate the problem in terms of a cost function which depends upon the entire time-history $\{\mathbf{x}(t) : t \in [t_0, t_f]\}]$. Such a functional $\Gamma_X[\mathbf{x}]$ exists and is called the *effective action*. (See Eyink, 2002). The minimizer of the effective action is the average history $\{\overline{\mathbf{x}}(t) : t \in [t_0, t_f]\}]$ in the absence of measurements. Its Hessian $\frac{\delta^2 \Gamma_X}{\delta \mathbf{x}(t) \delta \mathbf{x}(t')}[\overline{\mathbf{x}}]$ at the minimum is the inverse of the 2-time covariance matrix $\mathbf{C}(t, t') = \langle [\mathbf{x}(t) - \overline{\mathbf{x}}(t)][\mathbf{x}(t') - \overline{\mathbf{x}}(t')]^\top \rangle$. The effective action is related to the M-time entropy by the formula

$$H_X(\mathbf{x}_1, ..., \mathbf{x}_M) = \min_{\{\mathbf{x}: \mathbf{x}(t_m) = \mathbf{x}_m, m = 1, ..., M\}} \Gamma_X[\mathbf{x}], \tag{30}$$

where the minimum is over all histories that satisfy the constraints. This formula is a particular case of a general result in large deviations theory, called the Contraction Principle (Varadhan, 1984). To obtain the entire optimal history $\{\mathbf{x}_*(t) : t \in [t_0, t_f]\}]$, including times intermediate to measurements, one can minimize a joint effective action $\Gamma_*[\mathbf{x}] := \Gamma_{X,Y}[\mathbf{x}, \mathbf{y}]$ for both the state and the measurements. Analogous to (23), this is given by

$$\Gamma_*[\mathbf{x}] = \Gamma_X[\mathbf{x}] + \frac{1}{2} \sum_{m=1}^{M} [\mathbf{y}_m - \mathbf{H}(t_m)\mathbf{x}(t_m)]^\top \mathbf{R}_m^{-1} [\mathbf{y}_m - \mathbf{H}(t_m)\mathbf{x}(t_m)] \tag{31}$$

in the case of normally distributed observation errors and linear measurements. In this situation of discrete-time measurements, the continuous-time optimal history $\{\mathbf{x}_*(t) : t \in [t_0, t_f]\}]$ obtained as minimizer coincides with that given by equation (29). The conditional covariance matrix $\mathbf{C}_*(t, t')$ at all continuous times may also be obtained from the inverse Hessian of $\Gamma_*[\mathbf{x}]$. The resulting formalism appears very similar to the so-called "4D-VAR" data assimilation scheme which is currently practiced at many operational forecast centers around the world (Courtier et al., 1993). However, its statistical

basis and interpretation is entirely different. The "4D-VAR" approach is a maximum-likelihood estimation scheme, based upon minimizing the "bare" action in the exponent of (16) (along with suitable terms for measurements, as in (31)). Thus, it seeks the conditional mode, rather than the conditional mean. The effective action employed in our "mean-field" approach bears the same relation to the "bare" action used in 4D-VAR as does a macroscopic entropy or Gibbs free-energy to a microscopic Hamiltonian, in the analogy to Gibbsian equibrium statistical mechanics.

If we compare the computational algorithm which results from our "mean-field" approximation to the exact conditional analysis in KSP, we arrive at the somewhat paradoxical conclusion that it is even more difficult to apply than the exact method. Just as for KSP, the forward and backward Kolmogorov equations (7),(9) must be integrated. However, in the scheme discussed here, one such integration yields just $F_X(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M)$ and $\mathbf{x}(t; \boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M)$ at the current values of the "thermodynamic fields" $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_M$. One must evaluate these many times in order to apply any minimization algorithm to determine the optimal fields $\boldsymbol{\lambda}_1^*, ..., \boldsymbol{\lambda}_M^*$ required for the final estimate. The simplification achieved by the mean-field approximation lies in the jump conditions (24),(28). Compared with the exact jump conditions (8), (10), these involve only *linear* functions of the measured variable $\mathbf{h}(\mathbf{x}, t)$ in the exponent, whereas KSP—for Gaussian observation errors—requires quadratic functions of $\mathbf{h}(\mathbf{x}, t)$ in the exponent. In fact, the mean-field conditional analysis requires only linear functions of the observation variable in the exponent even if the observations are not Gaussian-distributed. This simpler form of the jump conditions at measurements will allow for simplified closure approximations to the evolution.

3.2 Variance-Minimizing Linear Analysis

The most common analysis currently employed in practical assimilation methods is a variance-minimizing linear analysis. In such schemes, the analysis state $\mathbf{x}_a(t)$ is taken to be a linear combination of the forecast state $\mathbf{x}_f(t)$ and the measurements $\mathbf{y}_1, ..., \mathbf{y}_M$, represented schematically as:

$$\mathbf{x}_a = \mathbf{x}_f + \mathbf{K}[\mathbf{y} - \mathbf{H}\mathbf{x}_f], \tag{32}$$

where $\mathbf{K}$ is a suitable matrix. For example, in a Kalman filtering scheme, $\mathbf{x}_f(t_m)$ is the forecast obtained by integrating the model equations or an ensemble average of such model solutions, and, for each m, $\mathbf{K}$ is a $d \times s$ matrix called the *Kalman gain*. d is the dimension of $\mathbf{x}(t_m)$ and s the dimension of $\mathbf{y}_m$. The matrix $\mathbf{K}$ is then chosen, sequentially for each time t_m, so that the variance $\langle |\mathbf{x}_a(t_m) - \mathbf{x}(t_m)|^2 \rangle$ is minimized, within the ansatz (32). This leads to well-known formulas and algorithms for calculating the Kalman gain. In

a Kalman smoothing scheme, the analysis such as (32) might hold for $\mathbf{x}_a(t)$ at all times t together and for the entire set of M measurements $\mathbf{y}_1, ..., \mathbf{y}_M$ simultaneously. In that case $\mathbf{K}$ will be a $(Td) \times (Ms)$ matrix, where T is the total number of time points considered in the calculation. The forecast state $\mathbf{x}_f$ might be, for example, the unconditioned ensemble average. Again, the variance-minimizing condition yields a formula for $\mathbf{K}$. This is embodied in the so-called "representer method" for calculating the least-squares estimator or Kalman smoother (e.g. see van Leeuwen and Evensen (1996), Section 3(b)).

The variance-minimizing linear analysis is known to reproduce the exact conditional statistics for the case of linear dynamics and additive, Gaussian noise. However, for nonlinear dynamics and for non-Gaussian statistics and/or multiplicative noise, the linear analysis is only an approximation. It may be a very poor one for highly non-Gaussian distributions, e.g. multimodal ones, as discussed in a recent paper of Evensen and van Leeuwen (2000). Our mean-field method leads to such a linear analysis, if the effective action is Taylor-expanded to quadratic order. Indeed, recall that the effective action is the generating function of so-called "irreducible correlation functions", and, in particular, its Hessian $\mathbf{\Gamma}(t,t') = \mathbf{C}^{-1}(t,t')$, the (operator) inverse of the 2-time covariance (Eyink, 2002). In that case, the cost function to be minimized in quadratic approximation is just

$$\begin{aligned}\Gamma_*^{(2)}[\mathbf{x}] &= \tfrac{1}{2}\int_{t_i}^{t_f} dt \int_{t_i}^{t_f} dt' \; [\mathbf{x}(t) - \overline{\mathbf{x}}(t)]^\top \mathbf{\Gamma}(t,t')[\mathbf{x}(t') - \overline{\mathbf{x}}(t')] \\ &+ \tfrac{1}{2}\sum_{m=1}^{M} [\mathbf{y}_m - \mathbf{H}(t_m)\mathbf{x}(t_m)]^\top \mathbf{R}_m^{-1}[\mathbf{y}_m - \mathbf{H}(t_m)\mathbf{x}(t_m)]\end{aligned} \tag{33}$$

and this coincides exactly with the expression in equation (19) of van Leeuwen and Evensen (1996), from which the representer solution to the Kalman smoother is derived. This observation should give some insight into the limitations of the linear analysis: it neglects the contributions of the terms of greater than degree two in the action, corresponding to higher-order cumulants of the distribution. This is one reason to expect our mean-field analysis will likely work better for strongly non-Gaussian statistics than will the linear analysis. In addition, the mean-field analysis is still a conditional analysis, although a suboptimal one. If transitions between states in the dynamical system (1) occur typically by some characteristic routes or "optimal paths", then one should not expect the mean-field conditioning to differ much from the exact conditioning. In general, it will be a little "weaker" than the true conditional analysis, but still more effective than the linear analysis. This point is made in greater detail in Eyink (2002), where it is explained how the mean-field conditioning corresponds in general to a larger subensemble than the exact conditioning.

4 Results for a Model Problem

Here we compare our mean-field conditional analysis directly with an exact conditional analysis. We consider as our example of data assimilation in a strongly nonlinear system the stochastically forced double-well system (see Miller et al. (1994), also Eyink and Restrepo (2000)):

$$\dot{x}(t) = f(x(t)) + \kappa\eta(t) \tag{34}$$

where

$$f(x) = 4x(1 - x^2) \tag{35}$$

and $\eta(t)$ is white-noise, with zero mean and covariance $\langle \eta(t)\eta(t') \rangle = \delta(t - t')$. As in Miller et al. (1994), we take $\kappa = 0.5$. Note that $f(x) = -U'(x)$, where $U(x)$ is the double-well potential

$$U(x) = -2x^2 + x^4 \tag{36}$$

with minima at $x = \pm 1$. The solution $x(t)$ of (34) executes small fluctuations about the minima in one of the wells with rather long residence times and, then, more rarely, experiences large fluctuations leading to a transition into the other well. The steady-state probability distribution of the model, $P_s(x) \propto \exp\left(-\frac{2U(x)}{\kappa^2}\right)$, is bimodal with peaks at $x = \pm 1$, the two fixed points of the deterministic dynamics.

We shall now consider results of data assimilation experiments for this model. As a sample history we shall use the same one that appeared in Figure 1 of Miller et al. (Miller et al., 1994), which is plotted in our own Figure 1. This

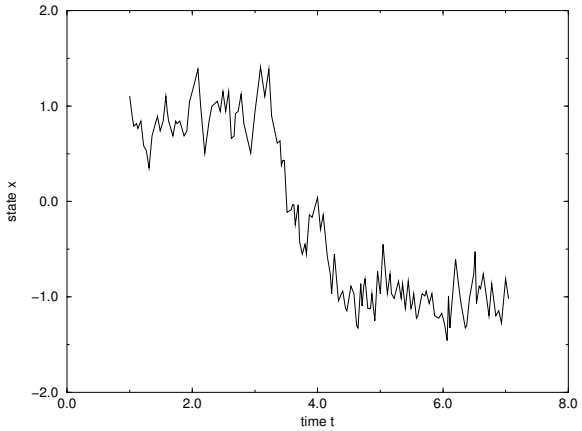

Fig. 1. Sample history.

choice shall allow us to compare the performance of our methods with those studied in Miller et al. (1994). The history adopted from that work is a solution of our model equation (34) for a particular initial condition and realization of the random noise. In terms of our experiment, this history represents "reality", i.e. the actual course of the system over time. However, this history is only imperfectly known from observations. We have generated sets of "measurements" by sampling the history at unit time intervals and adding to those values Gaussian random variables, which represent "observation errors". The observation errors are chosen independently at each measurement time, with mean zero and variance R. We consider several choices of the latter. Note that the standard deviation $\sqrt{R} \times 100\%$ represents the rms error expressed as a percentage of the size of the equilibrium states at ± 1.

4.1 Exact Conditional Analysis

We first describe the exact conditional statistics for the various sets of measurements. We obtain these statistics by solving the KSP equations, (7),(9), with jump conditions (8),(10). We solved the forward equation (7) numerically by the algorithm in Larson et al. (1985), which guarantees positive solutions and long-time convergence to the correct equilibrium solution $P_s(x)$. The algorithm was implemented on the x-interval $[-3, 3]$ with probability-conserving, zero-flux boundary conditions. We solved the backward equation (9) by the corresponding adjoint algorithm. The space grid-spacing $\Delta x = 0.09375$ and time-step $\Delta t = 0.01$ were employed in integrating both equations. For further details, see Alexander et al. (2002).

The first set of "measurements", which we shall call dataset A, were generated by adding to the reference history at unit intervals a particular set of realizations of Gaussian errors with $R = 0.04$, i.e. 20% rms errors. These simulated measurements are plotted in Figure 2. In the same figure we plot as a solid line the mean history conditioned on the measurements and, as a pair of dashed lines, the mean history plus or minus the standard deviation in the ensemble conditioned on the measurements. The mean history gives the "expected" history conditioned upon the observations. It should be kept in mind that it may represent very atypical behavior for actual realizations and only gives the average effect. Realizations which vary from the mean by only two or three standard deviations will have a reasonable probability of occurrence. Thus, the range of variations of typical realizations in the conditional ensemble is roughly indicated by the dashed lines. It is clear from the Figure 2 that these conditional statistics capture very well the transition that occurred around time $t =$ 3-5 in the history of Miller et al. (1994). On the contrary, the suboptimal methods discussed in Miller et al. (1994)—the Extended Kalman Filter and a least-squares variational or Maximum Likelihood Estimator—failed to

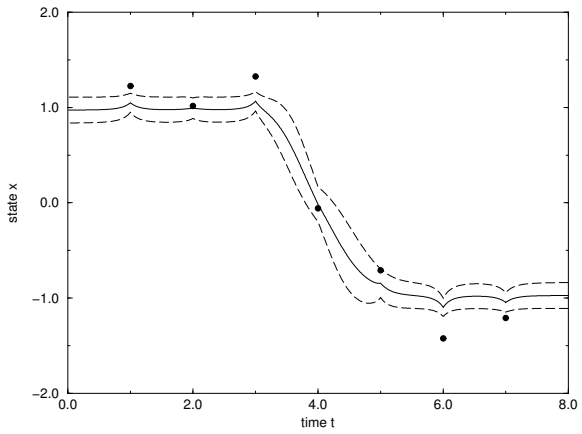

Fig. 2. Exact conditional analysis, using dataset A (20 % rms error) represented by filled circles. Solid line: mean, dashed line: mean ± standard deviation.

track the transition with similar measurements of 20% accuracy spaced at unit intervals.

In practice, it is difficult to know how large the errors in observations may be. Although the measurements marked by filled circles in Figure 2 were, in fact, generated by adding Gaussian random errors with $R = 0.04$, in a real assimilation experiment one might not have a good idea of the size of the observation errors. Therefore, we consider in Figure 3 the results of the KSP equations for a dataset B, with the same observations as before but with assumed error variance $R = 0.16$, i.e. 40% observation error. Clearly, the conditional statis-

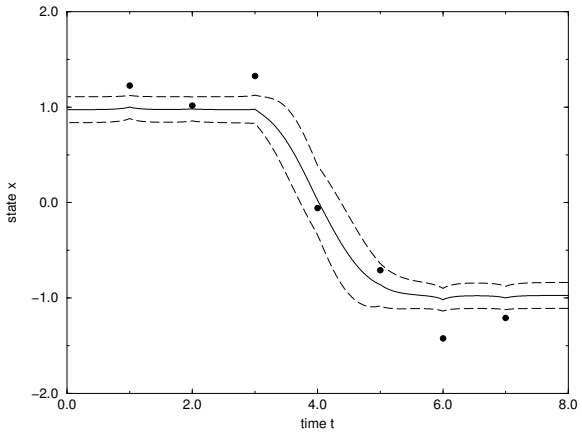

Fig. 3. Exact conditional analysis, using dataset B (40 % rms error) represented by filled circles. Solid line: mean, dashed line: mean ± standard deviation.

tics have changed very little by changing R. Only the conditional variance has increased slightly, as would be expected with measurements assumed to be less accurate. In fact, with the given set of measurements, the conditional statistics show very little change even up to $R = 1.00$, or 100% observation error. The conditional variance steadily increases as R increases, but the conditional mean continues to track well the transition (see Eyink and Restrepo, 2000). Only for $R > 1.00$ is it possible to begin to confuse a measurement in

one well for a value in the well on the other side. This stability of the conditional statistics against changes in the presumed size of observation errors is an important virtue as an estimation tool.

In Figure 4 we show more evidence of this stability. Plotted as filled circles

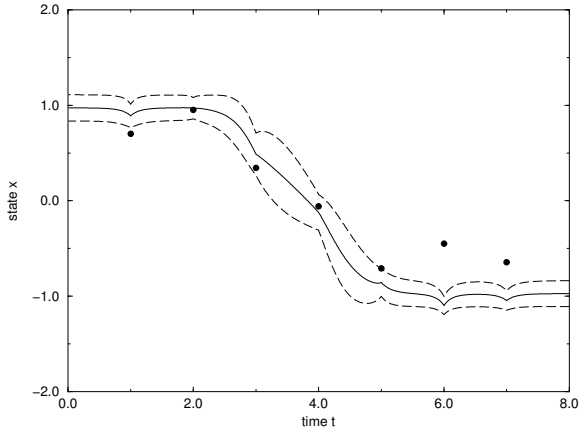

Fig. 4. Exact conditional analysis, using dataset C (20 % rms error) represented by filled circles. Solid line: mean, dashed line: mean ± standard deviation.

there is another independent set of observations, generated as for dataset A by adding Gaussian errors with $R = 0.04$, or 20% rms errors. We call this dataset C. In this case, we see that the errors tended to reduce the magnitude of all the measured values of the history toward zero. In general, with every different set of measurements on the same history, corresponding to different realizations of the observation error, there will be a different set of conditional statistics. However, as shown in Figure 4, the conditional statistics for this new set of observations are not very different from those shown in Figures 2 and 3. Thus we see that the conditional statistics possess remarkable stability to small changes in the measured values or one's assessment of the size of observation errors. Clearly, this *statistical stability* is a very desirable feature for any approximate data assimilation method to preserve. Unfortunately, the standard techniques reviewed in Miller et al. (1994) did not show such stability and might either follow or not follow the transition, depending sensitively upon the presumed value of R.

In Figure 5 we plot a dataset D in which measurements are generated by adding Gaussian errors with $R = 0.09$, or 30% rms errors. In this case, for the particular set of measurements indicated, the conditional statistics track the transition that occurred in the actual history through the first subsequent measurement but lose it thereafter. The conditional mean then becomes close to zero and even slightly positive, whereas the actual history after the transition was in the well at $x = -1$. Another example of this is shown in Figure 6, for a dataset E, which was generated from the history by adding errors with $R = 0.36$, or 60% rms errors. In this case, the conditional mean does not indicate a transition at all, except for a slight lowering of its value from near

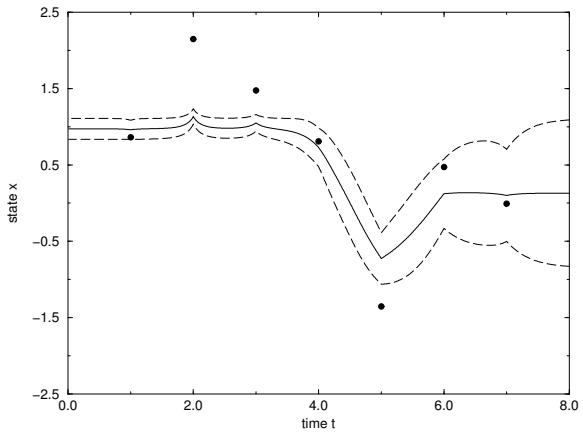

Fig. 5. Exact conditional analysis, using dataset D (30 % rms error) represented by filled circles. Solid line: mean, dashed line: mean ± standard deviation.

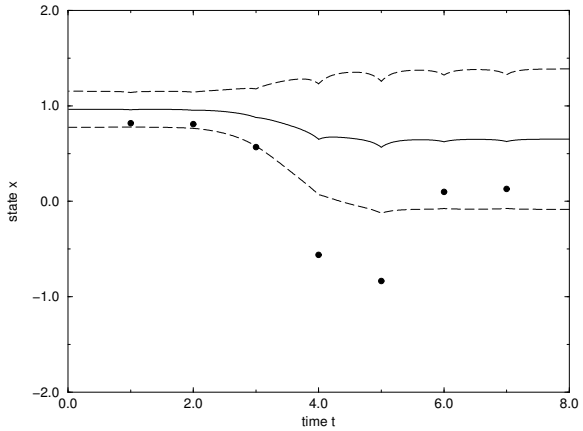

Fig. 6. Exact conditional analysis, using dataset E (60 % rms error) represented by filled circles. Solid line: mean, dashed line: mean ± standard deviation.

1.0 before the transition to about 0.6 afterward. It must be emphasized: this is not a failure of the KSP assimilation algorithm. In fact, using exact dynamics and exact conditioning upon the measurements produces the optimal results. No other information can be assumed to be given about the history other than the measurements indicated by the filled circles. Therefore, the transition that occurred in the actual history, with such poor measurements as given in datasets D and E, is irrecoverably lost. The conditional statistics shown in Figures 5 and 6 indicate that, for the ensembles of histories and observation errors that produce the given set of measurements, there are a majority of members in which the transition was either followed by a switch back to the other well (dataset D) or in which no transition occurred at all (dataset E). In the cases discussed here, the sample history in which the transition occurred is somewhat atypical of a general member of the conditional ensemble. In both Figures 5 and 6, a large increase in the conditional variance does occur after the transition in the sample history. Thus, the conditional statistics indicate a virtually complete loss of predictability after a time about $t = 6$ in Figure 5 and $t = 4$ in Figure 6. The sample history, although somewhat atypical,

is well within the range of allowed variation of members of the conditional ensemble (plus or minus a few standard deviations from the mean). The conditional statistics, for this set of poor measurements, indicate an intrinsic and irremediable uncertainty about the future state of the system.

The goal of any approximate data assimilation technique is not to recover this or that particular history, from which measurements are generated by adding observation errors. In fact, the only correct and achievable goal is to recover the conditional statistics—such as the conditional mean and variance—given the particular set of observations. Any assimilation method which yielded an estimate following the transition in our sample history using only datasets D or E would, in fact, be inferior to one which followed instead the conditional mean and failed to show a transition. The only way that any assimilation method could track the transition is by sheer chance or by surreptitiously employing more information from the sample history than that given in datasets D or E. Needless to say, in a real assimilation experiment, one *does* want to recover the actual history, but only by virtue of calculating correctly the conditional statistics. Such statistics as shown in Figures 5 or 6 would indicate the need for acquiring either more accurate measurements or additional measurement data, in order to recover the actual history. This is the only information one can hope to gain from a successful data assimilation method in these cases.

4.2 Mean-Field Conditional Analysis

We now describe the results of our assimilation experiments using the exact dynamics, but the mean-field conditional analysis. As discussed in subsection 3.1, the mean-field conditional statistics are obtained by solving the exact Kolmogorov forward and backward equations, (7),(9), but with the jump conditions (24),(28) at measurement times, involving the "thermodynamic fields", λ_m $m = 1, ..., M$. From one forward and backward integration of these equations, one obtains the "multi-time free energy" function $F_X(\lambda_1, ..., \lambda_M)$ and its gradient. To obtain the conditional mean history $x_*(t)$, one must solve the following minimax problem

$$\begin{aligned} &H_*(x_1^*, ..., x_M^*) = \min_{x_1,...,x_M} \max_{\lambda_1,...,\lambda_M} \\ &\left\{ \sum_{m=1}^M x_m \lambda_m - F_X(\lambda_1, ..., \lambda_M) + \sum_{m=1}^M \frac{(x_m - y_m)^2}{2R_m} \right\}, \end{aligned} \tag{37}$$

where y_m, $m = 1, ..., M$ are the measured values and R_m, $m = 1, ..., M$ are the error variances of those measurements. Indeed, carrying out the inside maximization over $\lambda_1, ..., \lambda_M$ gives $H_*(x_1, ..., x_M) = H_X(x_1, ..., x_M) + \sum_{m=1}^M \frac{(x_m - y_m)^2}{2R_m}$ [see (21)] and carrying out the outside minimization then determines $x_1^*, ..., x_M^*$. At measurement times $x_*(t_m) = x_m^*$, $m = 1, ..., M$, while

values at intermediate times are determined from (29). We also wish to have the conditional variance $\sigma_*^2(t)$. For this, we calculate the Hessian $(\mathbf{\Gamma}_*)_{mm'} = \partial^2 H_*/\partial x_m \partial x_{m'}(x_1^*, ..., x_M^*)$ and then obtain the variance from the diagonal elements of the inverse Hessian matrix $\sigma_*^2(t_m) = (\mathbf{\Gamma}_*^{-1})_{mm}$. To obtain $\sigma_*^2(t)$ at selected times t intermediate to the measurement times, we insert a set of "pseudo-measurements" at those times with infinite variance. Note that the gradient $\partial H_*/\partial x_m(x_1, ..., x_M)$ is just $\lambda_m(x_1, ..., x_M)$, which is provided by the inside maximization step in (37). Hence it is straighforward to calculate the Hessian by a finite-difference approximation of the first-derivatives $\partial \lambda_m/\partial x_{m'}(x_1^*, ..., x_M^*)$.

It is noted that it is easy to calculate the inverse for our simple 1-variable model, in which case the Hessian is an $M \times M$ matrix. In realistic applications of our method to large-scale systems we would instead use a method based upon the Contraction Principle which would eliminate the need to calculate such inverses. See Eyink et al. (2002b).

We have carried out these calculations numerically, using the same algorithm as in section 4.1 (see Larson et al., 1985) to integrate the Kolmogorov equations. The nested optimizations in (37) were each performed by a conjugate-gradient algorithm. To obtain the Hessian we used a 2nd-order accurate finite-difference approximation to the first derivatives $\partial \lambda_m/\partial x_{m'}$ and then found the inverse Hessian matrix by LU-decomposition.

The results of these calculations are as follows:

In Figure 7 we plot the dataset A and, for it, the mean-field conditional statistics: the mean, $x_*(t)$, and the mean plus or minus the standard devation, $x_*(t) \pm \sigma_*(t)$. If we compare with the exact conditional statistics in Figure 2, we see almost perfect agreement. The only discrepancy is around the time of

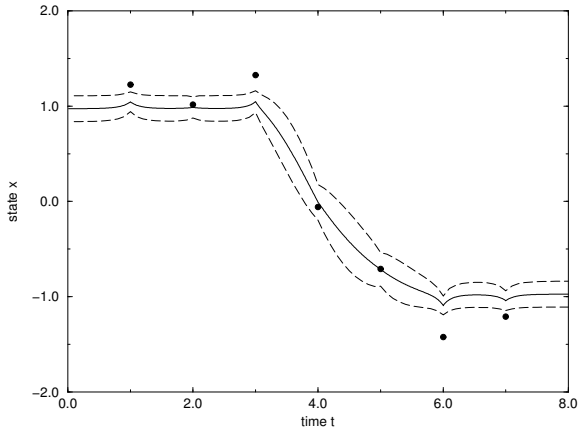

Fig. 7. Mean-field analysis, using dataset A (20 % rms error) represented by filled circles. Solid line: mean, dashed line: mean ± standard deviation.

the 5th measurement, where the measured value is about -0.7. In fact, this

good agreement should be expected. Except for the 4th and 5th measurements, all of the measured values are either bigger than 1 or less than -1. But when $|y_m| > 1$, the mean-field conditions $\overline{x}^N(t_m) = y_m$ will only be satisfied if, in almost every member of the N-sample ensemble, $x^{(n)}(t_m) \approx y_m,\ n = 1, ..., N$. Indeed, the dynamics does not allow a history $x(t)$ to achieve frequently any values much greater in absolute value than 1. Therefore, if $x^{(n)}(t_m)$ differed much from y_m at all, it would very likely be that $x^{(n)}(t_m)$ is either smaller in magnitude than y_m or even of the opposite sign. But if that were true for a finite fraction of ensemble members n, then it could not be true that $\overline{x}^N(t_m) = y_m$. Therefore, when $|y_m| > 1$, the mean-field conditioning is almost the same as the exact conditioning.

In Figure 8 we plot dataset B and its mean-field conditional statistics. Now the measurements are assumed to have 40% rms error, so that the conditions $\overline{x}^N(t_m) = y_m,\ m = 1, ..., M$ are less strictly enforced. Here we see a larger

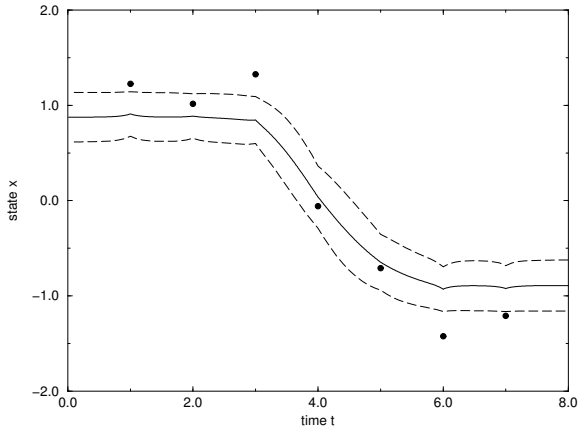

Fig. 8. Mean-field analysis, using dataset B (40 % rms error) represented by filled circles. Solid line: mean, dashed line: mean ± standard deviation.

departure from the exact conditional statistics in Figure 3. The mean history at the initial and final times takes on values slightly smaller in magnitude for the mean-field conditioning than for the exact conditioning. Also, the variance increases in dataset B for both conditional analyses compared to the variances for dataset A, but the increase is greater for the mean-field conditioning. Still, the crucial point is that the transition which is observed in the exact conditional statistics is well-preserved in the mean-field conditional statistics also. This remains true up to larger values of measurement error variance near 1.00 (see Eyink and Restrepo, 2000).

For dataset C in Figure 9 we see larger departures of the mean-field conditional statistics from the exact conditional statistics in Figure 4. In fact, for the exact conditional mean the starting value is near 1.0 and the final value near -1.0, but the mean-field conditional average has starting value near 0.75 and final value near -0.50. This discrepancy is easy to understand. The difference from the dataset A is that now $|y_m| < 1$ for all m. For the exact conditioning,

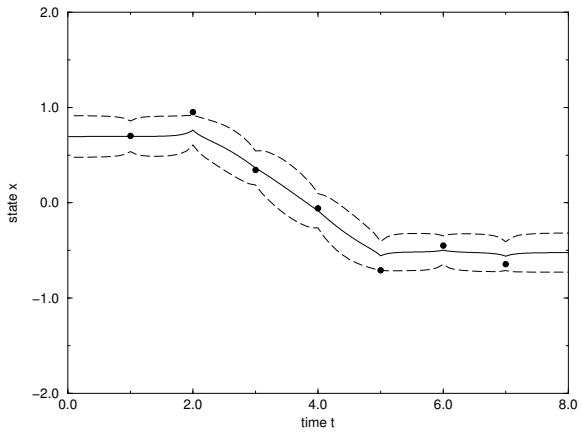

Fig. 9. Mean-field analysis, using dataset C (20 % rms error) represented by filled circles. Solid line: mean, dashed line: mean ± standard deviation.

$x(t_m) + \rho_m = y_m$ at each m^{th} measurement, where ρ_m is the error there. However, for a single history $x(t)$ to achieve a value with magnitude much less than 1 is unlikely. Hence, even though the rms error variance is small, 0.04, the exact conditioning assumes that $x(t_m)$ is near ± 1 for each m and that there is a large error ρ_m. However, the mean-field condition is that $\overline{x}^N(t_m) + \overline{\rho}^N_m = y_m$ for each m. In contrast to a single history, the N-sample ensemble average $\overline{x}^N(t)$ can easily become small, because values of the samples $x^{(n)}(t),\ n = 1, ..., N$ with values all near ± 1 can cancel in the sum. Thus, there is no high cost in the mean-field conditioning for $\overline{x}^N(t_m) \approx y_m$ and instead the mean errors $\overline{\rho}^N_m$ are assumed small. The consequence is that the average history with the mean-field conditioning stays very near the measurements, but the average history with the exact conditioning prefers to stay near the stable values ± 1 at the minima of the double well potential. Despite this discrepancy, the mean-field conditional statistics are still successful in tracking the transition which is observed in the exact conditional statistics. The mean-field conditioning is therefore qualitatively successful here, although not as accurate quantitatively as before.

For dataset D in Figure 10 we see a different phenomenon. In this case, the expected history with the mean-field conditioning is quite close to that for the exact conditioning, shown in Figure 5. The only discrepancy of reasonable magnitude is near the fifth measurement. However, there is a much greater difference in the variances for the two conditional analyses, after that measurement. The exact conditional statistics show then a large increase in the variance, while the mean-field conditional statistics show a much smaller increase. This can be understood technically from the remark that, for the mean-field analysis, the covariance $(\mathbf{\Gamma}_*)_{mm'} = \Gamma_{mm'} + R_m^{-1}\delta_{mm'} \geq R_m^{-1}\delta_{mm'}$ in the matrix sense. Thus, $\sigma_*^2(t_m) = (\mathbf{\Gamma}_*^{-1})_{mm} \leq R_m$ for each m. It follows that at each measurement time with dataset D, $\sigma_*(t_m) \leq 0.3,\ m = 1, ..., M$ and the value can only slowly increase between measurements. Just as for dataset C, we see that the mean-field conditioning assumes that measurement errors are

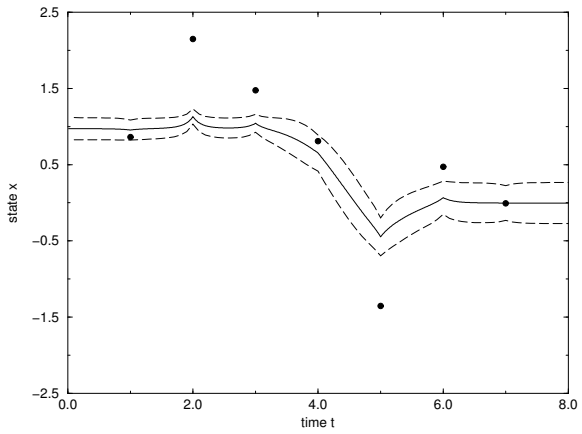

Fig. 10. Mean-field analysis, using dataset D (30 % rms error) represented by filled circles. Solid line: mean, dashed line: mean ± standard deviation.

small and the ensemble histories are closer to the measured values, whereas the exact conditioning prefers to assume that measurement errors are large and thus ensemble members are dispersed more widely about the measured value. Nevertheless, the mean-field conditional statistics do preserve the main features observed in the exact conditional statistics for dataset D: the transition is tracked through the first subsequent measurement and then lost. In both cases, typical ensemble members show a transition from 1 to -1 and then a switch back to the well at 1.

Figure 11 for dataset E combines the features of the two previous examples: the mean-field conditioning produces both a smaller magnitude of the mean history (due to measured values with magnitudes < 1) and also smaller variances at late times (due to assumed smaller observation errors) than the exact conditioning in Figure 6. Still, even in this worst case, the expected history

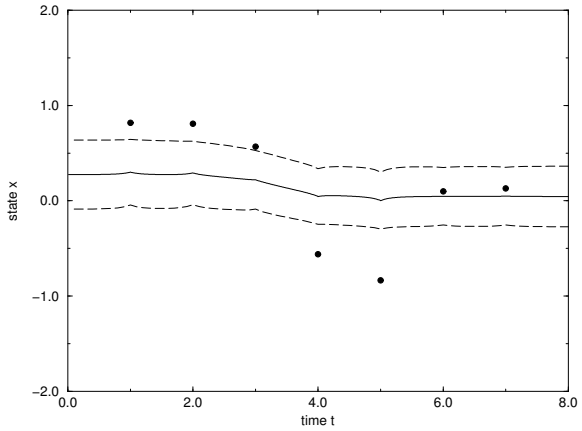

Fig. 11. Mean-field analysis, using dataset E (60 % rms error) represented by filled circles. Solid line: mean, dashed line: mean ± standard deviation.

with the mean-field conditional analysis shows a similar trend as does the expected history with the exact conditional analysis. One should keep in mind that the observation errors are quite large here, 60%, larger than one would

hope to have in practice. Nevertheless, we believe that it is important for data assimilation schemes to work, i.e. give accurate results for conditional statistics, even when presented with very poor measurements. In the first place, it is not possible in many applications to know really how good measurements may be, and then a conservative assessment of their accuracy is prudent. In the second place, relatively poor measurements may be all that one has in certain cases, and yet one would still like to glean what information they contain about the past or future behavior of the system.

5 Conclusion and Discussion

In this paper, we have studied conditional statistics of nonlinear dynamical systems, in particular the mean and the covariance. These are argued to represent the logically correct solution to the problem of data assimilation or inverse modeling in such systems. We have shown how to calculate conditional probability distributions efficiently for systems with a small number of degrees-of-freedom by solving the forward and backward Kolmogorov equations. The observational data is incorporated in that case by "jump conditions" at measurement times, which implement Bayes formula from probability theory. We have, as well, proposed a "mean-field approximation" to this analysis step, which can be formulated as a minimax variational problem. The calculation of the cost function and its gradient requires also the solution of the forward-backward Kolmogorov equations. Therefore, the mean-field approximation to the conditional analysis is not, by itself, a practical method for application to realistic, spatially-extended, nonlinear dynamical systems. Our purpose in this paper has not been to test the mean-field approximation for practical applications on its own, but instead to gain some understanding of the accuracy of the approximation. We have shown in the context of a simple model problem with bimodal statistics, that the mean-field conditional analysis gives a satisfactory approximation to the conditional statistics and is successful in tracking mode transitions where more traditional methods fail. For practical purposes, it would be realistic to apply the mean-field conditional analysis in conjunction with moment-closure approximations to the dynamical evolution. That is the subject of follow-up papers Eyink et al. (2002a), Eyink et al. (2002b).

Acknowledgements: We wish to thank C. D. Levermore and M. Ghil in particular for much useful advice on the subject of this paper and, also, the anonymous reviewers of a previous version of the paper. This work, LAUR 02-3058, 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 (GLE,JMR),

as well as from NASA, Goddard Space Flight Center, Grant NAG5-11163 (JMR).

References

Alexander, F. J., Eyink, G. L., Restrepo, J., 2002. Path integral monte carlo for nonlinear estimation, *in preparation*.

Bouttier, F., 1994. A dynamical estimation of forecast error covariances in an assimilation system. Mon. Wea. Rev. 122, 2376–2390.

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

Campillo, F., Cerou, F., LeGland, F., Rakotozafy, R., 1993. Algorithmes parallèles pour le filtrage non linéare et les équations aux dérivées partielles stochastiques. Bull. Liaison Rech. Info. Auto 141, 21–24.

Courtier, P., Derber, J., Errico, R., Louis, J.-F., Vukicevic, T., 1993. Important literature on the use of adjoint, variational methods and the kalman filter in meteorology. Tellus 45A, 342–357.

Evensen, G., 1992. Using the extended kalman filter with a multilayer quasi-geostrophic ocean model. J. Geophys. Res. 97, 17905–17924.

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), 10 143–10 162.

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

Eyink, G. L., 2002. A variational formulation of optimal nonlinear estimation, *submitted*.

Eyink, G. L., Restrepo, J. M., Alexander, F. J., 2002a. A statistical-mechanical approach to data assimilation for nonlinear dynamics. ii. analysis approximations, *submitted*.

Eyink, G. L., Restrepo, J. M., Alexander, F. J., 2002b. A statistical-mechanical approach to data assimilation for nonlinear dynamics. iii. analysis approximations, *submitted*.

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

Gauthier, P., Courtier, P., Moll, P., 1993. Assimilation of simulated wind lidar data with a kalman filter. Mon. Wea. Rev. 121, 1803–1820.

Jazwinski, A. H., 1970. Stochastic Processes and Filtering Theory. Academic Press, New York.

Kushner, H. J., 1962. On the differential equations satisfied by conditional probability densities of markov processes, with applications. J. SIAM Control, Ser.A 2, 106–119.

Kushner, H. J., 1967a. Approximation to optimal nonlinear filters. IEEE Trans. Auto. Contr. 12, 546–556.

Kushner, H. J., 1967b. Dynamical equations for optimal nonlinear filtering. J. Diff. Eq. 3, 179–190.

Larson, E. W., Levermore, C. D., Pomerang, G. C., Sanderson, J. G., 1985. Discretization methods for one-dimensional fokker-planck operators. J. Comp. Phys. 61, 359–390.

Leith, C. E., 1974. Theoretical skill of monte carlo forecasts. Mon Wea. Rev. 102, 409–418.

Lorenc, A. C., Hammon, O., 1988. Objective quality control of observations using bayesian methods. theory, and a practical implementation. Q. J. R. Meteorol. Soc. 114, 515–543.

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

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

Pardoux, E., 1982. Équations du filtrage non linéaire de la prédiction et du lissage. Stochastics 6, 193–231.

Risken, H., 1984. The Fokker-Planck Equation. Springer-Verlag, New York.

Stratonovich, R. L., 1960. Conditional markov processes. Theor. Prob. Appl. 5, 156–178.

Tippett, M. K., Anderson, J. L., Bishop, C. H., Hamill, T. H., Whitaker, J. S., 2003. Ensemble square-root filters. Monthly Weather Review, to appear.

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

Varadhan, S. R. S., 1984. Large Deviations and Applications. SIAM, Philadelphia.