Performance Bounds for Particle Filters Using the Optimal Proposal

CHRIS SNYDER

National Center for Atmospheric Research, Boulder, Colorado*

THOMAS BENGTSSON

Genentech, San Francisco, California

MATHIAS MORZFELD

Department of Mathematics, University of California, Berkeley, Berkeley, California

(Manuscript received 10 April 2015, in final form 5 August 2015)

ABSTRACT

Particle filters may suffer from degeneracy of the particle weights. For the simplest ''bootstrap'' filter, it is known that avoiding degeneracy in large systems requires that the ensemble size must increase exponentially with the variance of the observation log-likelihood. The present article shows first that a similar result applies to particle filters using sequential importance sampling and the optimal proposal distribution and, second, that the optimal proposal yields minimal degeneracy when compared to any other proposal distribution that depends only on the previous state and the most recent observations. Thus, the optimal proposal provides performance bounds for filters using sequential importance sampling and any such proposal. An example with independent and identically distributed degrees of freedom illustrates both the need for exponentially large ensemble size with the optimal proposal as the system dimension increases and the potentially dramatic advantages of the optimal proposal relative to simpler proposals. Those advantages depend crucially on the magnitude of the system noise.

1. Introduction

Particle filters are ensemble-based algorithms for data assimilation that, unlike many schemes, make no assumptions about the probability distributions for the prior or the observation errors. One difficulty is that particle filters may suffer from degeneracy, in which the weight assigned to one ensemble member (or particle) converges to one while those assigned to all other members approach zero. Bengtsson et al. (2008), Bickel et al. (2008), and Snyder et al. (2008) (hereafter BBS08) showed that avoiding degeneracy in the most elementary particle filter [essentially, the bootstrap filter of Gordon et al. (1993)] requires an ensemble size that increases exponentially with the variance of the log-likelihood of the observations given each member, which in simple examples is proportional to the system dimension.

More general particle filters employ sequential importance sampling, in which the ensemble members at each step are drawn from a proposal distribution. The choice of proposal distribution strongly influences the performance of these filters (e.g., Liu and Chen 1998; Doucet et al. 2000; Arulampalam et al. 2002). Here, we consider the proposal given by the distribution of the present state given the state at the previous step and the most recent observations, which is known as the ''optimal'' proposal (Doucet et al. 2000). We extend the asymptotic results of BBS08 to the optimal proposal for the case of linear, Gaussian systems and thus demonstrate that, even with that proposal, the required ensemble size will still grow exponentially with an appropriate measure of the problem size.

Several particle filters that use sequential importance sampling have recently been developed for geophysical

* The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Corresponding author address: Chris Snyder, NCAR, P.O. Box 3000, Boulder, CO 80307.
E-mail: chriss@ucar.edu

DOI: 10.1175/MWR-D-15-0144.1

applications. The implicit particle filter (Chorin and Tu 2009; Morzfeld et al. 2012; Chorin et al. 2013) generates the ith new particle by solving an algebraic equation related to the likelihood of the most recent observations given the ith particle at the previous time. It is equivalent to the optimal proposal for linear, Gaussian systems. The equivalent-weights particle filter (van Leeuwen 2010; Ades and Van Leeuwen 2015) also uses the most recent observations in generating the ith new particle, by nudging the trajectory beginning from the ith particle at the previous time toward the new observations. It includes a further step that depends on the new observations, in which most particles are adjusted toward locations with nearly equal importance ratios. Papadakis et al. (2010) also present a particle filter in which the proposal uses the new observations.

For a given system and a given ensemble size, these more sophisticated particle filters often perform much better than the bootstrap filter, but their behavior as the system size increases has not yet been established. In principle, the analysis of BBS08 could be extended to each new filter. We avoid this nontrivial task by demonstrating that, out of the class of particle filters that generate new particles based only on the new observations and the particles generated at previous step, the optimal proposal minimizes the variance of the (unnormalized) weights over draws of both the previous and new particles. This result extends the usual optimality statement for the optimal proposal, namely, that it minimizes the variance of weights over draws of the new particles. The extended optimality means the particle filter employing the optimal proposal provides a lower bound for the ensemble size necessary to avoid degeneracy of the weights, a bound which applies to all single-step particle filters that use sequential importance sampling, including the implicit particle filter and the equivalent-weights particle filter.

The outline of the paper is as follows. The next section provides further background on sequential importance sampling and the optimal proposal distribution. In section 3, we show that the asymptotic results of BBS08 also hold for the optimal proposal in the case of linear, Gaussian systems and we examine the behavior of the optimal proposal in a simple test problem with independent and identically distributed (i.i.d.) degrees of freedom. Some of the basic results given here can also be found in Snyder (2012). Section 4 demonstrates that the optimal proposal is optimal in the extended sense described above, while section 5 outlines how system dimension, the system dynamics, and details of the observing network influence the required ensemble size, leading to a back-of-the-envelope assessment of particle filtering for global numerical weather prediction. We conclude with a summary and discussion of our results, including a suggestion that effective particle filters for high-dimensional systems will need to include some form of spatial localization, such as is employed in ensemble Kalman filters.

2. Background

This section briefly reviews sequential importance sampling and the optimal proposal distribution, together with the degeneracy of the particle weights and the asymptotic results of BBS08 related to degeneracy. Readers familiar with these topics can proceed to section 3.

Consider a discrete-time system with state $\mathbf{x}$ of dimension N_x and noisy observations $\mathbf{y}$ of dimension N_y that are related to the state. The system is determined by the transition density $p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1})$ for the state dynamics and the conditional density for the observations $p(\mathbf{y}_k \,|\, \mathbf{x}_k)$, where the subscript k indicates evaluation at the kth time, t_k.

Our goal is to estimate the filtering density $p(\mathbf{x}_k \,|\, \mathbf{y}_k, \mathbf{Y}_{k-1})$. Here, $\mathbf{Y}_{k-1} = \{\mathbf{y}_1, \ldots, \mathbf{y}_{k-1}\}$ is the set of all observations before t_k. Since all pdfs in what follows will be conditioned on it, $\mathbf{Y}_{k-1}$ will be omitted in subsequent expressions.

Particle filters are sequential Monte Carlo techniques that represent $p(\mathbf{x}_k \,|\, \mathbf{y}_k)$ using a weighted ensemble of states, $\{\mathbf{x}_k^i, w_k^i : i = 1, \ldots, N_e\}$, where N_e is the ensemble size and the weights must sum to 1 over the ensemble. The ensemble members $\mathbf{x}_k^i$ are also called *particles*.

We will be interested in particle filters that employ sequential importance sampling. Given particles and weights $\{\mathbf{x}_{k-1}^j, w_{k-1}^j; j = 1, \ldots, N_e\}$ at t_{k-1}, sequential importance sampling proceeds by drawing a new particle $\mathbf{x}_k^i$ valid at t_k for each $\mathbf{x}_{k-1}^i$ from a specified distribution $\pi(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}^i, \mathbf{y}_k)$. The distribution $\pi(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k)$ is known as the *proposal*. The particle $\mathbf{x}_k^i$ must be weighted according to

$$w_k^i \propto \tilde{w}_k^i w_{k-1}^i, \tag{1}$$

with the unnormalized, incremental weights $\tilde{w}_k^i$ given by

$$\tilde{w}_k^i = \frac{p(\mathbf{x}_k^i \,|\, \mathbf{x}_{k-1}^i)\, p(\mathbf{y}_k \,|\, \mathbf{x}_k^i)}{\pi(\mathbf{x}_k^i \,|\, \mathbf{x}_{k-1}^i, \mathbf{y}_k)}. \tag{2}$$

The proposal $\pi(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k)$ may be chosen as desired, as long as its support includes the support of the target distribution $p(\mathbf{x}_k \,|\, \mathbf{y}_k)$. Further background on sequential importance sampling, using the same notation, appears in Snyder (2012); Doucet et al. (2000) and van Leeuwen (2009) also provide extensive reviews.

It is also possible to develop particle filters that do not employ sequential importance sampling (Klaas et al. 2005; Nakano 2014). While we expect that high-dimensional problems will also be difficult for that class

of particles filters, the results we present in sections 3 and 4 do not carry over directly.

An important subtlety in sequential importance sampling is that, for each i, $\mathbf{x}_k^i$ is drawn from the proposal conditioned at the previous time on the specific realization $\mathbf{x}_{k-1}^i$ of $\mathbf{x}_{k-1}$. This means that if at each step $j<k$ we retain $\mathbf{x}_j^i$, then we have constructed a weighted sample $\{\mathbf{x}_0^i, \mathbf{x}_1^i, \ldots, \mathbf{x}_k^i, w_k^i\}$ from the joint conditional distribution, $p(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_k \,|\, \mathbf{y}_k)$. A weighted sample $\{\mathbf{x}_k^i, w_k^i\}$ may then be obtained by marginalization—simply omitting the samples at $t_0, \ldots, t_{k-1}$. Nevertheless, as will be discussed further in section 4, the sequential sampling causes $\mathbf{x}_k^i$ and w_k^i to depend on the samples at previous times, despite the marginalization.

A common choice for the proposal is the transition distribution for the dynamics,

$$\pi(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k) = p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}), \tag{3}$$

for which the incremental weights are

$$\tilde{w}_k^i = p(\mathbf{y}_k \,|\, \mathbf{x}_k^i). \tag{4}$$

We will term this the *standard* proposal. Another possible choice, which is known as the optimal proposal in the particle-filtering literature (e.g., Doucet et al. (2000)), is

$$\pi(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k) = p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k), \tag{5}$$

with the incremental weights

$$\tilde{w}_k^i = p(\mathbf{y}_k \,|\, \mathbf{x}_{k-1}^i). \tag{6}$$

A key difficulty for particle filters is that $\tilde{w}_k^i$ may vary greatly across particles, so that many particles receive small weights. In the extreme situation, which is termed *degeneracy* in the particle-filtering literature and *collapse* in Snyder et al. (2008), $w_k^i \approx 1$ for a single i, all other particles have weights close to zero and the conditional distribution will be poorly approximated. The asymptotic results of BBS08 concern degeneracy. They define[1]

$$\tau^2 = \mathrm{var}(-\log \tilde{w}_k^i), \tag{7}$$

where the variance is taken over $\pi(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k)$. They then show that, if τ^2 and N_e are large, $(\log N_e)/\tau^2$ is small, and the distribution of $-\log \tilde{w}_k^i$ is sufficiently close to Gaussian, then the maximum weight $w^{(N_e)}$ behaves as

$$E(1/w^{(N_e)}) \sim 1 + \frac{\sqrt{2\log N_e}}{\tau}. \tag{8}$$

Thus, the maximum weight approaches 1 when τ^2 is large unless N_e is comparable to or larger than $\exp(\tau^2/2)$.

Since (6) does not depend on $\mathbf{x}_k^i$, the optimal proposal achieves the minimum possible variance (namely, zero) of $\tilde{w}_k^i$ over a sample $\{\mathbf{x}_k^i\}$ drawn from the proposal with $\mathbf{x}_{k-1}^i$ and $\mathbf{y}_k$ fixed [again, see Doucet et al. (2000)]. Nevertheless, the weights may still vary substantially among the particles, and degeneracy can be a problem, because of the dependence in (6) of $\tilde{w}_k^i$ on $\mathbf{x}_{k-1}^i$, the particles at the previous time.

3. The optimal proposal in the context of linear, Gaussian systems

This section demonstrates that the asymptotic arguments of Bengtsson et al. (2008) and Snyder et al. (2008) are applicable to the optimal proposal when the system is linear and Gaussian. It also presents a linear, Gaussian example that illustrates both the asymptotic results and the potential benefits provided by the optimal proposal relative to the standard proposal. Although what follows is restricted to linear, Gaussian systems, the numerical simulations of Slivinski and Snyder (2015, manuscript submitted to *Mon. Wea. Rev.*) demonstrate that the asymptotic results are also informative in simple nonlinear systems.

a. Asymptotic relations following Bengtsson et al.

To extend the asymptotic arguments of Bengtsson et al. (2008) and Snyder et al. (2008) to the optimal proposal, we must show that $-\log \tilde{w}_k^i = -\log p(\mathbf{y}_k \,|\, \mathbf{x}_{k-1}^i)$ is approximately Gaussian for N_y large, when considered as a function of the random variable $\mathbf{x}_{k-1}^i$.

Restricting to the linear, Gaussian case, we write the system as

$$\mathbf{x}_k = \mathbf{M}\mathbf{x}_{k-1} + \eta_k, \quad \mathbf{y}_k = \mathbf{H}\mathbf{x}_k + \epsilon_k, \tag{9}$$

where $\eta_k \sim N(0, \mathbf{Q})$, $\epsilon_k \sim N(0, \mathbf{R})$, and the tilde symbol ($\sim$) means "is distributed as." For simplicity, we also assume that η_k and ϵ_k are independent of each other and across times. It follows from (9) that

$$\mathbf{y}_k \,|\, \mathbf{x}_k \sim N(\mathbf{H}\mathbf{x}_k, \mathbf{R}), \quad \mathbf{y}_k \,|\, \mathbf{x}_{k-1} \sim N(\mathbf{H}\mathbf{M}\mathbf{x}_{k-1}, \mathbf{R} + \mathbf{H}\mathbf{Q}\mathbf{H}^{\mathrm{T}}). \tag{10}$$

The second relation in (10) together with the definition of the probability density for a multivariate Gaussian implies

[1] In those papers, τ^2 is defined in terms of the explicit form (4) of $\tilde{w}_k^i$ for the standard proposal, which is a special case of (7). The rest of their derivation [e.g., section 4 of Snyder et al. (2008)] also follows with the definition (7), requiring only that $-\log \tilde{w}_k^i$ is approximately distributed as a Gaussian for large τ^2.

$$-\log\tilde{w}_k^i = \frac{1}{2}(\mathbf{y}_k - \mathbf{H}\mathbf{M}\mathbf{x}_{k-1}^i)^{\mathrm{T}}(\mathbf{R} + \mathbf{H}\mathbf{Q}\mathbf{H}^{\mathrm{T}})^{-1}(\mathbf{y}_k - \mathbf{H}\mathbf{M}\mathbf{x}_{k-1}^i). \tag{11}$$

Now let λ_j and the columns of $\mathbf{E}$ be, respectively, the eigenvalues and eigenvectors of

$$\mathrm{cov}[(\mathbf{R} + \mathbf{H}\mathbf{Q}\mathbf{H}^{\mathrm{T}})^{-1/2}\mathbf{H}\mathbf{M}\mathbf{x}_{k-1}]. \tag{12}$$

Defining $\tilde{\mathbf{y}}_k = \mathbf{E}(\mathbf{R} + \mathbf{H}\mathbf{Q}\mathbf{H}^{\mathrm{T}})^{-1/2}\mathbf{y}_k$ and $\tilde{\mathbf{H}} = \mathbf{E}(\mathbf{R} + \mathbf{H}\mathbf{Q}\mathbf{H}^{\mathrm{T}})^{-1/2}\mathbf{H}\mathbf{M}$, and noting that $\mathbf{E}^{\mathrm{T}}\mathbf{E} = \mathbf{I}$, (11) becomes

$$\begin{aligned} -\log\tilde{w}_k^i &= \frac{1}{2}(\tilde{\mathbf{y}}_k - \tilde{\mathbf{H}}\mathbf{x}_{k-1}^i)^{\mathrm{T}}(\tilde{\mathbf{y}}_k - \tilde{\mathbf{H}}\mathbf{x}_{k-1}^i) \\ &= \frac{1}{2}\sum_{j=1}^{N_y}[\tilde{y}_{k,j} - (\tilde{\mathbf{H}}\mathbf{x}_{k-1}^i)_j]^2, \end{aligned} \tag{13}$$

where $\tilde{y}_{k,j}$ and $(\tilde{\mathbf{H}}\mathbf{x}_{k-1}^i)_j$ are the jth components of $\tilde{\mathbf{y}}_k$ and $\tilde{\mathbf{H}}\mathbf{x}_{k-1}^i$, respectively.

The terms in the summation in (13) are mutually independent, given $\tilde{\mathbf{y}}_k$, because $\tilde{\mathbf{H}}\mathbf{x}_{k-1}^i$ is Gaussian and, as direct calculation shows, its covariance matrix is $\mathrm{diag}(\lambda_1^2, \ldots, \lambda_{N_y}^2)$. We, therefore, expect that the sum in many situations has a distribution that approaches a Gaussian for large N_y. As discussed in Bickel et al. (2008), this is true as long as

$$\frac{\lambda_j^2}{\sum_{i=1}^{N_y}\lambda_i^2} \ll 1, \tag{14}$$

which guarantees that no single term or set of terms dominates the sum.

The asymptotic relation (8) thus holds for the optimal proposal in the linear, Gaussian case when (14) is true, with the consequence that N_e must be at least as large as $\exp(\tau^2/2)$ to avoid degeneracy. An expression for τ^2 is given in Snyder et al. (2008):

$$\tau^2 = \sum_{j=1}^{N_y}\lambda_j^2\left(\frac{3}{2}\lambda_j^2 + 1\right), \tag{15}$$

which follows from the fact that the variance of the sum of mutually independent terms is the sum of variances of each term and from standard expressions for the moments of a Gaussian.

The results (11)–(15) can also be derived more quickly, but perhaps less transparently, by noting that the relations (10) imply that $p(\mathbf{y}_k \mid \mathbf{x}_{k-1})$ has the same form (as a function of $\mathbf{y}_k$ and $\mathbf{x}_{k-1}$) as $p(\mathbf{y}_k \mid \mathbf{x}_k)$ but with $\mathbf{x}_k$ replaced by $\mathbf{M}\mathbf{x}_{k-1}$ and $\mathbf{R}$ replaced by $\mathbf{R} + \mathbf{H}\mathbf{Q}\mathbf{H}^{\mathrm{T}}$. Thus, the results of section 5 of Snyder et al. (2008), concerning the Gaussianity of $-\log\tilde{w}_k^i$ for the standard proposal and the expression (15) for τ^2, may be immediately applied to the optimal proposal and $p(\mathbf{y}_k \mid \mathbf{x}_{k-1})$, as long as we replace $\mathbf{x}_k$ with $\mathbf{M}\mathbf{x}_{k-1}$ and $\mathbf{R}$ with $\mathbf{R} + \mathbf{H}\mathbf{Q}\mathbf{H}^{\mathrm{T}}$. In particular, the eigenvalues λ_j used in (15) for the standard proposal come from the matrix

$$\mathrm{cov}(\mathbf{R}^{-1/2}\mathbf{H}\mathbf{x}_k), \tag{16}$$

rather than (12).

Extending these results for the optimal proposal [especially (8)] to non-Gaussian, nonlinear systems hinges on showing that $\log\tilde{w}_k^i$ is nearly Gaussian. This is facilitated in the linear, Gaussian case by the fact that $\log\tilde{w}_k^i$ can always be written, as in (13), as a sum of independent terms. Although general statements are difficult, Bengtsson et al. (2008) and Bickel et al. (2008) discuss some conditions under which a similar central limit theorem holds for nonlinear, non-Gaussian systems, and Slivinski and Snyder (2015, manuscript submitted to *Mon. Wea. Rev.*) show that (8) is valid for a specific, significantly non-Gaussian system. Nevertheless, the linear, Gaussian case considered here is sufficient to establish that the optimal proposal does not avoid degeneracy.

The linear Gaussian case also provides insight into the potential advantages of the optimal proposal. Comparing (16) and (12) shows that τ^2 for the two proposals will be the same as $\mathbf{Q}$ becomes small, since then $\mathbf{R} + \mathbf{H}\mathbf{Q}\mathbf{H}^{\mathrm{T}} \to \mathbf{R}$ and $\mathbf{M}\mathbf{x}_{k-1} \to \mathbf{x}_k$. The benefits from the optimal proposal, therefore, depend on system noise and will be negligible when the system noise is sufficiently small. Using properties of the eigenvalues of symmetric matrices [see chapter 10 of Parlett (1998)], one can also with some effort demonstrate that the ith eigenvalue of (16) is always bounded below by the ith eigenvalue of (12), so that τ^2 given by (15) is always smaller for the optimal proposal, with equality only when $\mathbf{Q} = 0$.

b. *A simple system with i.i.d. degrees of freedom*

Consider the linear, Gaussian system (9) with $\mathbf{M} = a\mathbf{I}$, $a > 0$ a scalar, $\eta_{k-1} \sim N(0, q^2\mathbf{I})$, and $\epsilon_k \sim N(0, \mathbf{I})$. Each element of the state vector then evolves and is observed independently. We are interested in how the particle filter performs in this system over a single update step, say at time t_k, and, therefore, also assume a simple form for the state distribution at the previous step: $\mathbf{x}_{k-1} \sim N(0, \mathbf{I})$. There are two parameters: q, the standard deviation of the system noise; and a, the standard deviation of the prior for $\mathbf{x}_k$ when q becomes small, which measures the degree to which the deterministic system dynamics affect the forecast uncertainty.

For the standard proposal, the distributions needed for sampling $\mathbf{x}_k$ and which determine the weights [from (3) and (4), respectively] are

$$\mathbf{x}_k \,|\, \mathbf{x}_{k-1} \sim N(a\mathbf{x}_{k-1}, q^2\mathbf{I}), \quad \mathbf{y}_k \,|\, \mathbf{x}_k \sim N(\mathbf{x}_k, \mathbf{I}), \quad (17)$$

while those needed for the optimal proposal [from (5) and (6)] are

$$\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k \sim N\left(\frac{a\mathbf{x}_{k-1} + q^2\mathbf{y}_k}{1+q^2}, \frac{q^2}{1+q^2}\mathbf{I}\right),$$
$$\mathbf{y}_k \,|\, \mathbf{x}_{k-1} \sim N[a\mathbf{x}_{k-1}, (1+q^2)\mathbf{I}]. \quad (18)$$

The form of $\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k$ can be derived by beginning from the expression for $\mathbf{x}_k \,|\, \mathbf{x}_{k-1}$ and then conditioning on $\mathbf{y}_k$ using the standard Kalman filter update, noting that the prior $\mathbf{x}_k \,|\, \mathbf{x}_{k-1}$ has covariance $q^2\mathbf{I}$.

This simple system also allows explicit expressions for τ^2 for either proposal. To use (15), we need the eigenvalues λ_j^2 of the matrices (16) or (12). The N_y eigenvalues are equal to a^2+q^2 and $a^2(1+q^2)^{-1}$, respectively, since $\mathrm{cov}(\mathbf{x}_{k-1}) = \mathbf{I}$ and $\mathrm{cov}(\mathbf{x}_k) = (a^2+q^2)\mathbf{I}$. Thus,

$$\tau^2 = \begin{cases} N_y(a^2+q^2)\left(\frac{3}{2}a^2 + \frac{3}{2}q^2 + 1\right), & \text{standard proposal} \\ N_y(q^2+1)^{-2}a^2\left(\frac{3}{2}a^2 + q^2 + 1\right), & \text{optimal proposal} \end{cases}, \quad (19)$$

where τ^2 is proportional to the system dimension N_x; this follows from the fact that each degree of freedom is independent and independently observed and, as is clear from (15), is not a general property of all systems. We discuss the relation of τ^2 and N_x further in section 5.

We first check the validity of the asymptotic relation (8) for the optimal proposal. Figure 1 shows $E(1/w^{(N_e)})$ estimated from numerical simulations as a function of $(2\log N_e)^{1/2}/\tau$. The asymptotic result is clearly useful for the optimal proposal, implying that N_e must grow as $\exp(\tau^2/2)$ to avoid degeneracy, just as is the case for the standard proposal. As expected, (8) holds when $(2\log N_e)^{1/2}/\tau$ is not too large, with quantitative accuracy for $(2\log N_e)^{1/2}/\tau$ as large as 0.4.

Returning to (19), it is immediately clear that τ^2 is the same for the two proposals in the limit that $q^2 \to 0$. This is consistent with the fact that the two proposals approach each other as the system noise becomes small, since $p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1})$ and $p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k)$ approach delta functions centered at the same value of $\mathbf{x}_k$ (given by the deterministic map of $\mathbf{x}_{k-1}$ to t_k). As the system noise increases, τ^2 for the optimal proposal becomes smaller and smaller relative to τ^2 for the standard proposal. Both points are illustrated in Fig. 2, which displays the ratio of τ^2 for the two proposals as a function of a^2 and q^2.

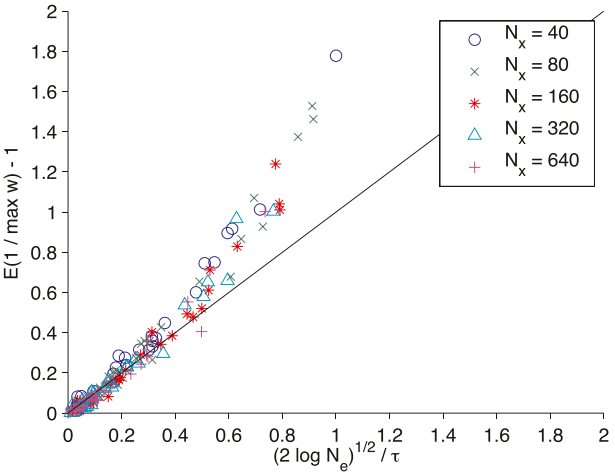

FIG. 1. Accuracy of the asymptotic relation (8) in numerical simulations using the optimal proposal. To obtain a range of values for $(2\log N_e)^{1/2}/\tau$, the simulations use $N_x = 5\times 2^m$ for $m = 3, 4, \ldots, 7$; $N_e = 2^n$ for $n = 4, 5, \ldots, 8$; and the parameters a and q are chosen randomly and independently from uniform distributions on $[0, 4]$. Results for different N_x are indicated by different symbols (as shown in the legend) and the expectation $E(1/w^{(N_e)})$ is approximated over 100 realizations of each simulation.

Even at moderate values of the system noise variance, the decrease of τ^2 with the optimal proposal implies very large gains in performance. For example, taking $a^2 = q^2 = 0.5$, so that the prior variance a^2+q^2 at t_k is equal to the observation-error variance, τ^2 is reduced by a factor of 5 when using the optimal proposal. Fixing a^2 and q^2, if we suppose that the optimal proposal needs, say, $N_e \approx 100$ to achieve a certain value for $E(1/w^{(N_e)})$, then the asymptotic theory predicts that the standard proposal would require $N_e \approx 10^{10}$ to achieve that same value, since N_e depends exponentially on τ^2.

The predictions of the asymptotic theory are confirmed in Fig. 3, which shows the minimum N_e such that $E(1/w^{(N_e)}) < (0.9)^{-1}$ for various values of N_x. The ensemble size grows exponentially with N_x for both proposals. At the same time, the exponent is smaller and the growth with N_x is much slower for the optimal proposal. Quantitatively, the slopes of the best-fit lines on the log-linear plot have a ratio of 4.6, which agrees closely with the asymptotic prediction of 5 when $a^2 = q^2 = 0.5$.

Although more minor, the optimal proposal has the additional advantage that it yields better estimates for a given value of $E(w^{(N_e)})$ (i.e., for a given degree of degeneracy). When $E(w^{(N_e)}) \geq 0.8$, the mean squared error

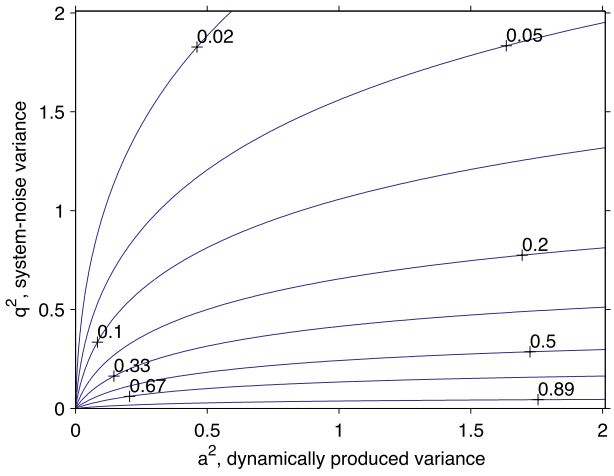

FIG. 2. Contours for the ratio of τ^2 for the optimal proposal to that for the standard proposal, as a function of a^2 and q^2.

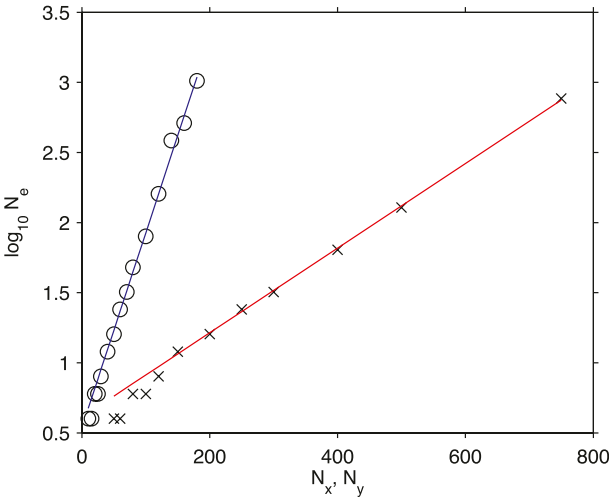

FIG. 3. The minimum N_e for which $E(1/w^{(N_e)}) < (0.9)^{-1}$, as a function of N_x and using $a^2 = q^2 = 0.5$ in the simple system. Results from numerical simulations are shown for the standard proposal (circles) and the optimal proposal (crisscrosses), together with best-fit lines for each proposal that omit the data for the four smallest values of N_x. The expectation of $1/w^{(N_e)}$ is computed over 10^3 realizations.

(MSE) from both proposals is significantly larger than that of the correct posterior mean, but the MSE using the optimal proposal can be a factor of 2–5 smaller than that using the standard proposal (Fig. 4a). The advantage of the optimal proposal clearly comes from moving the particles at t_k toward the new observations. The estimates of the posterior variance, however, are of similar quality for either proposal at fixed $E(w^{(N_e)})$ (Fig. 4b).

4. Performance bounds from the optimal proposal

We next demonstrate that particle filters using the optimal proposal have minimal degeneracy, first explaining on an informal, intuitive level why this is so and then presenting a rigorous proof.

Our arguments apply to "single step" algorithms for particle filters, that is, algorithms in which the sample at t_k is generated using only the sample generated previously at t_{k-1} and the observations $\mathbf{y}_k$. More complex algorithms are possible, such as the block resampling of Doucet et al. (2006) in which sampling at t_k utilizes observations $\mathbf{y}_k$, $\mathbf{y}_{k-1}$, …, $\mathbf{y}_{k-L}$ for some $L > 0$ and involves regenerating the samples at t_{k-1}, …, t_{k-L} given $\mathbf{y}_k$, …, $\mathbf{y}_{k-L}$. We will comment further on block-resampling algorithms at the end of this section.

a. An intuitive view

It will be helpful to first review sequential importance sampling as applied in single-step particle filters. Crucially, the target distribution is the joint conditional distribution $p(\mathbf{x}_{k-1}, \mathbf{x}_k \,|\, \mathbf{y}_k)$. This means that the weight update (1) and (2) follows from the ratio of the target distribution to the proposal, evaluated at the joint sample $(\mathbf{x}^i_{k-1}, \mathbf{x}^i_k)$,

$$w^i_k \propto w^{*,i}_k = \frac{p(\mathbf{x}^i_{k-1}, \mathbf{x}^i_k \,|\, \mathbf{y}_k)}{\pi(\mathbf{x}^i_{k-1}, \mathbf{x}^i_k \,|\, \mathbf{y}_k)}, \tag{20}$$

where we have introduced the unnormalized importance weights $w^{*,i}_k$, following Kong et al. (1994). To have a sequential scheme, we must also assume a joint proposal for $\mathbf{x}_{k-1}$ and $\mathbf{x}_k$ of the following form:

$$\pi(\mathbf{x}_{k-1}, \mathbf{x}_k \,|\, \mathbf{y}_k) = \pi(\mathbf{x}_{k-1})\pi(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k). \tag{21}$$

This allows a sample from the joint proposal to be generated by beginning from $\mathbf{x}^i_{k-1}$ drawn from $\pi(\mathbf{x}_{k-1})$ at t_{k-1} and drawing $\mathbf{x}^i_k$ from $\pi(\mathbf{x}_k \,|\, \mathbf{x}^i_{k-1}, \mathbf{y}_k)$ as described in section 2. Equations (1) and (2) then follow from (20) by inserting (21) in the denominator and writing the numerator as

$$p(\mathbf{x}^i_{k-1}, \mathbf{x}^i_k \,|\, \mathbf{y}_k) = p(\mathbf{x}^i_{k-1})p(\mathbf{x}^i_k \,|\, \mathbf{x}^i_{k-1})p(\mathbf{y}_k \,|\, \mathbf{x}^i_k)/p(\mathbf{y}_k). \tag{22}$$

Now, to understand the optimal proposal at a heuristic level, consider the joint conditional distribution factored according to the identity:

$$p(\mathbf{x}_{k-1}, \mathbf{x}_k \,|\, \mathbf{y}_k) = p(\mathbf{x}_{k-1} \,|\, \mathbf{y}_k)p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k). \tag{23}$$

In the context of a Monte Carlo algorithm, (23) states that one can draw from the joint distribution conditioned on $\mathbf{y}_k$ in a sequential fashion, first drawing $\mathbf{x}^i_{k-1}$ from $p(\mathbf{x}_{k-1} \,|\, \mathbf{y}_k)$ and then generating $\mathbf{x}^i_k$ from $p(\mathbf{x}_k \,|\, \mathbf{x}^i_{k-1}, \mathbf{y}_k)$. The optimal proposal would therefore be "perfect," in the sense that it would directly provide a

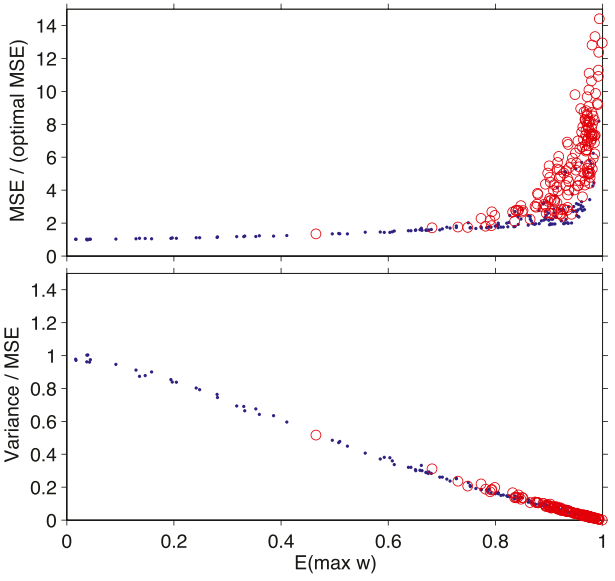

FIG. 4. The ratio of mean squared error (MSE), averaged over 100 realizations, to (top) the MSE of the optimal, conditional-mean estimate and (bottom) the ratio of the estimate posterior variance to the MSE as a function of the expected maximum weight. Results from numerical simulations for the standard proposal (circles) and optimal proposal (dots) are shown, with different points corresponding to different values of a^2, q^2, N_x, and N_e in the simple system, generated as in Fig. 1.

sample from $p(\mathbf{x}_k \,|\, \mathbf{y}_k)$ and the incremental weights $\tilde{w}_k^i$ would be equal for all $\mathbf{x}_k^i$, if the algorithm began with a sample from $p(\mathbf{x}_{k-1} \,|\, \mathbf{y}_k)$.

In a single-step particle filter, however, the best that we can hope for is that $\mathbf{x}_{k-1}^i$ is drawn (perhaps after resampling at t_{k-1}) from $p(\mathbf{x}_{k-1})$, the distribution of the state conditioned on all information *before* t_k. For the optimal proposal, this results in an incremental weight $\tilde{w}_k^i = p(\mathbf{x}_{k-1}^i \,|\, \mathbf{y}_k)/p(\mathbf{x}_{k-1}^i) \propto p(\mathbf{y}_k \,|\, \mathbf{x}_{k-1}^i)$, as in (6). Even if we choose the optimal proposal, the weights w_k^i will vary to account for the fact that $\mathbf{x}_{k-1}^i$ was not drawn from $p(\mathbf{x}_{k-1} \,|\, \mathbf{y}_k)$.

Comparison of (23) and (21) yields similar intuition. Choosing the optimal proposal makes the second density on the rhs agree, but the proposal lacks the conditioning on $\mathbf{y}_k$ in the first density on the rhs.

These heuristic arguments indicate that the fundamental limitation of a single-step particle filter is not how cleverly the proposal is chosen but rather that the algorithm does not correct particles at earlier times to reflect new observations.

b. *Another look at optimality of the optimal proposal*

The foregoing, intuitive argument suggests that the optimal proposal really is the best possible proposal for a single-step particle filter, since it is exactly the second distribution on the rhs of (23). We next give a rigorous result, showing that the optimal proposal minimizes the variance of $w_k^{*,i}$, over draws from the joint proposal distribution. This result has not been noted previously in the literature, which instead emphasizes that the optimal proposal yields weights with minimal variance (namely zero) over draws of $\mathbf{x}_k^i$ alone [see proposition 2 of Doucet et al. (2000)]. The variance of $w_k^{*,i}$ is of interest because $1 + \mathrm{var}(w_k^{*,i})$ approximates the increase in sampling variance for the mean of an arbitrary function of $\mathbf{x}_k$ relative to a draw from the posterior distribution [see (13) in Kong et al. (1994)].

A general, inductive proof covering times from t_0 to t_k appears in appendix B. Here, we focus on how degeneracy occurs in the step from t_{k-1} to t_k, by assuming that we are given $\mathbf{x}_{k-1}^i$ from $p(\mathbf{x}_{k-1})$ (i.e., a random draw from the posterior distribution at t_{k-1}).

With this assumption, the joint proposal for $(\mathbf{x}_{k-1}, \mathbf{x}_k)$ has the form (21) with $\pi(\mathbf{x}_{k-1})$ replaced by $p(\mathbf{x}_{k-1})$:

$$\pi(\mathbf{x}_{k-1}, \mathbf{x}_k \,|\, \mathbf{y}_k) = p(\mathbf{x}_{k-1})\pi(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k).$$

Define the random variable

$$w_k^* = \frac{p(\mathbf{x}_{k-1}, \mathbf{x}_k \,|\, \mathbf{y}_k)}{\pi(\mathbf{x}_{k-1}, \mathbf{x}_k \,|\, \mathbf{y}_k)}, \tag{24}$$

so that $w_k^{*,i}$ in (2) is w_k^* evaluated at $(\mathbf{x}_{k-1}^i, \mathbf{x}_k^i)$. We now ask which choice of $\pi(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k)$ minimizes $\mathrm{var}(w_k^*)$, with expectation taken over $\pi(\mathbf{x}_{k-1}, \mathbf{x}_k \,|\, \mathbf{y}_k)$.

The variance can be computed directly from $\mathrm{var}(w_k^*) = E(w_k^{*2}) - E(w_k^*)^2$. Using the definition of w_k^* above, we have

$$E(w_k^*) = \int w_k^* \pi(\mathbf{x}_{k-1}, \mathbf{x}_k \,|\, \mathbf{y}_k)\, d\mathbf{x}_{k-1}\, d\mathbf{x}_k = 1.$$

Proceeding similarly but factoring the numerator according to (23) gives

$$E(w_k^{*2}) = \int \frac{p(\mathbf{x}_{k-1} \,|\, \mathbf{y}_k)^2}{p(\mathbf{x}_{k-1})} \left[\int \frac{p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k)^2}{\pi(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k)} d\mathbf{x}_k \right] d\mathbf{x}_{k-1},$$

which yields

$$\mathrm{var}(w_k^*) = -1 + \int f(\mathbf{x}_{k-1}, \mathbf{y}_k; \pi) \frac{p(\mathbf{x}_{k-1} \,|\, \mathbf{y}_k)^2}{p(\mathbf{x}_{k-1})} d\mathbf{x}_{k-1}, \tag{25}$$

where

$$f(\mathbf{x}_{k-1}, \mathbf{y}_k; \pi) = \int \frac{p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k)^2}{\pi(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k)} d\mathbf{x}_k. \tag{26}$$

The function $f(\mathbf{x}_{k-1}, \mathbf{y}_k; \pi)$ consists of an integral whose integrand is the ratio of the square of a probability density to a second probability density. Appendix A shows that such integrals are greater than or equal to 1, with equality if and only if the two densities are the same. Hence, $f(\mathbf{x}_k, \mathbf{y}_k; \pi) \geq 1$ and

$$\mathrm{var}(w_k^*) \geq -1 + \int \frac{p(\mathbf{x}_{k-1} \mid \mathbf{y}_k)^2}{p(\mathbf{x}_{k-1})} d\mathbf{x}_{k-1},$$

with the lower bound achieved only for the optimal proposal, $\pi(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{y}_k) = p(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{y}_k)$.

This means that other single-step particle filters will always exhibit degeneracy that is more pronounced than that for the optimal proposal at a given N_e and will always require larger N_e to limit degeneracy in the weights to a desired level. In particular, the implicit particle filter and the equivalent-weights particle filter are both single-step sequential algorithms and both will exhibit worse degeneracy of weights than a particle filter based on the optimal proposal.

In contrast to our arguments, Bocquet et al. (2010) compare particle filters using the standard and optimal proposal in an idealized system and find advantages for the standard proposal in certain parameter regimes. The system they consider, however, is deterministic, a setting in which the standard and optimal proposals are identical. Moreover, the filters they implement include a fictitious system noise when drawing from the respective proposals. We conclude that any advantages of the standard proposal in their experiments arise from specific details of those implementations.

c. *Block-resampling algorithms*

Sections 4a and 4b explain how the performance of single-step particle filters is limited by the need to correct particles at t_{k-1} given more recent observations $\mathbf{y}_k$ at t_k. Doucet et al. (2006) derive an approach to accomplish this correction, which they term block resampling. Introducing the notation $\mathbf{x}_{i:j}$ to indicate the concatenation of states from t_i through t_j, the idea is that, in addition to generating particles $\mathbf{x}_k^i$, the particles $\mathbf{x}_{k-L+1:k}$ within a window of length L will also be resampled using a proposal distribution that depends on $\mathbf{y}_k$. Additional algorithms for updating past particles with new observations are given in Lin et al. (2013). In the geophysical literature, Weir et al. (2013) also emphasize the potential for improving particle filters by updating particles at earlier times given new observations.

For block resampling, the counterpart of the optimal proposal is

$$\pi(\mathbf{x}_{k-L+1:k} \mid \mathbf{y}_{k-L+1:k}, \mathbf{x}_{k-L}) = p(\mathbf{x}_{k-L+1:k} \mid \mathbf{y}_{k-L+1:k}, \mathbf{x}_{k-L}), \tag{27}$$

with the incremental weights:

$$\tilde{w}_k^i \propto p(\mathbf{y}_k \mid \mathbf{y}_{k-L+1:k-1}, \mathbf{x}_{k-L}^i). \tag{28}$$

As in the single-step algorithm, the proposal (27) clearly minimizes the variance of incremental weights over draws of $\mathbf{x}_{k-L+1:k}^i$, since the weights are independent of $\mathbf{x}_{k-L+1:k}^i$. The arguments of section 4b can also be extended to give similar bounds for block-resampling algorithms, which will again be achieved if and only if the optimal proposal (27) is used.

Block resampling using (27) is one potential way to reduce N_e without leading to degeneracy. The variance of the incremental weights (28) will decrease as the length L of the resampling window increases as long as the system dynamics are not deterministic [i.e., as long as $p(\mathbf{x}_k \mid \mathbf{x}_{k-1})$ is not a delta function], since the probability of $\mathbf{y}_k$ given observations over the window will become less dependent on $\mathbf{x}_{k-L}^i$, the state at the beginning of the window. Doucet et al. (2006) show in certain important cases that the dependence on $\mathbf{x}_{k-L}^i$ decreases exponentially with L.

Block resampling is not without drawbacks. As was the case for the optimal proposal, block resampling using (27) depends crucially on the specification of the system noise. More important, sampling from the proposal distribution is no longer easy or inexpensive. Indeed, its difficulty approaches that of sampling directly from the posterior distribution as L increases. Although the techniques of Morzfeld et al. (2012) offer promise for the future, implementation of block resampling remains prohibitive at present for many geophysical applications such as numerical weather prediction.

5. τ^2, number of observations, and system dimension

The interest of our results lies in their implications for particle filters in high-dimensional systems. Equation (8) relates the maximum weight to τ^2 rather than directly to the system dimension N_x. In the simple system of section 3b, τ^2 is proportional to N_y and N_x, and relating the degeneracy of the weights to N_x is straightforward.

In general, however, τ^2 has a more subtle relation to N_x and N_y. It is clear from (15) that τ^2 can depend on the observing network [$\mathbf{R}$ and $\mathbf{H}$ in (16) and (12), for the linear Gaussian case], properties of the system [$\mathbf{M}$ and $\mathbf{Q}$ in (12)] and the statistics of the prior, which will depend on both system properties and the observing network. Moreover, for the standard and optimal proposals, $\tilde{w}_k^i$ depends only on observed quantities. Thus, τ^2 cannot be directly related to N_x, since some portion of the state may be invisible to the observations and that part of the state will not affect τ^2. (As an example, τ^2 will not depend on the prior statistics for winds if only temperatures are observed.)

This leaves the question of how large τ^2 is in specific applications. We make a rough estimate for global numerical weather prediction (NWP) as follows, beginning

with the standard proposal. Methods for more direct estimates can be found in Slivinski and Snyder (2015, manuscript submitted to *Mon. Wea. Rev.*). Chorin and Morzfeld (2013) also explore the feasibility of high-dimensional assimilation, but from a different perspective.

For the standard proposal, τ^2 depends, via (15) and (16), on the prior covariance as expressed in the observed variables and normalized by the observation-error covariance. The balanced component of the prior errors has correlation length of 500 km or less in the troposphere (Rabier et al. 1998), while prior errors in moisture and the unbalanced components, and the observation errors, have smaller correlation lengths. Thus, if we tile the globe with 100 regions, each having spatial dimensions of 1000 km, it is plausible that the assimilation problem on each tile is approximately independent of the others and that the global spectrum λ_i of (16) can be approximated as the union of the spectra on the 500 tiles individually, each of which is approximately the same. Each tile likely has upward of 10 λ_i that are significant (i.e., the prior covariance on the tile has significant variance for at least 10 eigenvectors), since the tiles have dimensions comparable to the error correlation length in the horizontal and the errors also have multiple vertical correlation scales within the domain. Modern global data-assimilation systems utilize 10^7 observations, so that each tile contains roughly 10^5 observations and those 10 λ_i are likely all significantly greater than unity (i.e., the forecast error in each of those 10 directions is relatively well observed). The resulting estimate is that $\tau^2 \geq 10^4$ if the standard proposal is applied to global NWP.

Extending this estimate to the optimal proposal requires assumptions about the magnitude of the system noise appropriate for global NWP. Little is known about the system noise and, indeed, many operational assimilation systems ignore system noise. It, therefore, seems reasonable to assume that the deficiencies of the forecast model are not the dominant contributions to short-range forecast errors. In the context of the simple system, the correct parameter regime is $a^2 \geq q^2$ and Fig. 2 indicates that τ^2 for the optimal proposal is factor of perhaps 5 or 10 smaller than for the standard proposal as long as the prior variance is comparable to or smaller than the observation-error variance (i.e., $a^2 + q^2 \leq 1$). This gives an estimate of $\tau^2 \geq 10^3$ for the optimal proposal applied to NWP.

6. Summary and discussion

This paper has shown that particle filters using the optimal proposal are subject to the asymptotic results of BBS08 that relate the expectation of one over the maximum weight to $(\tau^2/2\log N_e)^{1/2}$, where τ^2 is the logarithm of the variance of the incremental weights. The asymptotics extend directly to the optimal proposal in the case of linear, Gaussian systems and imply that the optimal proposal requires $N_e \sim \exp(\tau^2/2)$ to avoid degeneracy of the particle weights.

Various properties of the optimal proposal can be illustrated in the simple system given in section 3b. The simple system consists of N_x independent degrees of freedom, each of which is independently observed. In this system, τ^2 is proportional to N_x and N_y and particle filtering using either the standard or optimal proposals needs an ensemble size N_e that increases exponentially with the dimension N_x. Nevertheless, the optimal proposal reduces τ^2 relative to that for the standard proposal and thus its advantage over the standard proposal can be dramatic, since N_e depends exponentially on τ^2. The main parameter determining the optimal proposal's advantage is the variance of the system noise; larger system noise increases the optimal proposal's benefits relative to the standard proposal.

A second important result (section 4) is that the optimal proposal is optimal in the sense that it minimizes a specific measure of weight degeneracy. More precisely, among all "single step" proposal distributions for $\mathbf{x}_k^i$, that is, those proposals depending only on the particle $\mathbf{x}_{k-1}^i$ at the previous time and the new observations $\mathbf{y}_k$, the optimal proposal minimizes the variance of the unnormalized weights (24) [or, more generally, (B2)] overdraws from the proposal at $t_0, t_1, \ldots, t_k$. The optimality shown here extends the usual result presented in the particle-filtering literature, which is that the optimal proposal yields weights with zero (and thus minimal) variance overdraws of the new particles $\mathbf{x}_k^i$. This optimality of the optimal proposal, together with the asymptotic results applied to the optimal proposal, implies that other single-step particle filters, such as the implicit and equivalent-weights filters, cannot avoid the need for N_e that grows exponentially with τ^2 (although they may reduce τ^2 relative to the standard proposal).

We next consider the relation of our results to the study of Ades and Van Leeuwen (2015). They apply the equivalent-weights particle filter in experiments with simulated two-dimensional turbulence and $N_x \approx 6 \times 10^4$. They find that the analysis mean provides a good estimate of the system state, even with only 32 particles, although higher moments of the posterior distribution are not well represented. They conclude, in contradiction to the results presented here, that the equivalent-weights filter has "overcome" the need for very large ensembles in high-dimensional systems.

How can their results be reconciled with ours that show that the optimal proposal bounds the performance of the equivalent-weights filter? First, the equivalent-weights proposal as implemented in Ades and Van Leeuwen (2015) becomes sharper as N_e increases (via their parameter ϵ, which they set to $10^{-3}/N_e$). The dependence of

the proposal on N_e facilitates keeping weights approximately equal, but also has the effect that substantial portions of the state space will contain no particles even as N_e increases. This places the specific implementation of the algorithm used by Ades and Van Leeuwen (2015) outside the scope of both standard convergence theorems for sequential importance sampling and our asymptotic results. We note, however, that versions of the algorithm with a more standard, N_e-independent proposal do suffer from degeneracy, as mentioned in Ades and Van Leeuwen (2015) following their Eq. (A18).

Second, as discussed in section 5, the degeneracy of the weights does not depend directly on N_x and so it may be that τ^2 for the optimal proposal is small enough in Ades and Van Leeuwen (2015) that degeneracy can be avoided with $N_e = 32$. Decaying two-dimensional turbulence exhibits most rapid error growth at the larger, energy-containing scales (Rotunno and Snyder 2008) and this is realized in the physical variables as error in the position of the coherent vortices (Boffetta et al. 1997). The simulations in Ades and Van Leeuwen (2015) have $O(10)$ vortices and thus forecast errors in their experiments may have as few as100–200 significant degrees of freedom, a much smaller number than N_x. Figure 3 shows that, in the simple system, the optimal proposal avoids degeneracy for $N_e \approx 30$ when there are as many as 300 independent degrees of freedom.

Overall, the optimal proposal offers substantial improvements over the standard proposal when the system noise is not too small, requiring orders of magnitude fewer ensemble members in many moderately large problems. At the same time, N_e must still grow exponentially with τ^2 to avoid degeneracy of the weights, even with the optimal proposal. The back-of-the-envelope estimate of section 5 indicates that a particle filter using the optimal proposal will not be feasible for global NWP for many years.

It is important to emphasize that our results hold only for particle filters using sequential importance sampling. Filters that seek to apply importance sampling directly to the marginal, conditional distribution at t_k have been proposed (Klaas et al. 2005; Nakano 2014), but the weights will be well behaved only when the proposal is close to the desired conditional distribution. An open question is whether the most convenient proposals, such as those based on the update step of the ensemble Kalman filter, will be sufficient in high dimensions.

In our view, further progress in particle filtering for large, spatially distributed problems, such as global NWP, will rest on the incorporation of some form of spatial localization into the algorithm. Localization (Houtekamer and Mitchell 1998, 2001; Hamill et al. 2001) capitalizes on the common property in geophysical systems that state variables separated by a sufficient distance are nearly independent, and is the key idea that allows the ensemble Kalman filter to perform well with $N_e = O(100)$ for a variety of geophysical applications. Bengtsson et al. (2003) were the first to investigate various possibilities for localization in a particle-filter update and note the complications introduced by the nonlinearity of the update and by the need to preserve spatial continuity when constructing random samples. Transform methods (Reich 2013; Metref et al. 2014), in which the emphasis is to find a mapping from a given sample to a sample consistent with the posterior distribution, are especially suitable for localization, as the transform may be computed locally, based on a local set of observations, much as in the local ensemble transform Kalman filter (Hunt et al. 2007). Spatial localization also has the appeal that it will be effective even when the system noise is small and its implementation does not require any information about the system noise.

APPENDIX A

Demonstration that $f(\mathbf{x}_{k-1}, \mathbf{y}_k; \pi) \geq 1$

Clearly $f(\mathbf{x}_{k-1}, \mathbf{y}_k; \pi) = 1$ for the optimal proposal density, $\pi(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k) = p(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k)$. In the general case, the function $f(\mathbf{x}_{k-1}, \mathbf{y}_k; \pi)$ has the form $\int [\rho(\mathbf{x})^2/\mu(\mathbf{x})]\, d\mathbf{x}$ with $\rho(\mathbf{x})$ and $\mu(\mathbf{x})$ probability density functions that may depend on $\mathbf{x}_{k-1}$ and $\mathbf{y}_k$.

Defining $\Delta(\mathbf{x}) = \rho(\mathbf{x}) - \mu(\mathbf{x})$, the integrand may be rewritten as

$$\frac{\rho(\mathbf{x})^2}{\mu(\mathbf{x})} = \mu(\mathbf{x}) + 2\Delta(\mathbf{x}) + \frac{\Delta(\mathbf{x})^2}{\mu(\mathbf{x})}.$$

Now it is easy to see that $\int \mu(\mathbf{x})\, d\mathbf{x} = 1$ and $\int \Delta(\mathbf{x})\, d\mathbf{x} = 0$. Therefore, if $\rho(\mathbf{x}) \neq \mu(\mathbf{x})$,

$$\int \frac{\rho(\mathbf{x})^2}{\mu(\mathbf{x})}\, d\mathbf{x} > 1,$$

since the final term $\Delta(\mathbf{x})^2/\mu(\mathbf{x})$ in the integrand is nonnegative and nonzero.

APPENDIX B

General Proof of Optimality

For single-step algorithms using sequential importance sampling, the proposal distribution for $\mathbf{x}_{0:k}$ has the following form:

$$\pi(\mathbf{x}_{0:k} \,|\, \mathbf{y}_{1:k}) = \pi(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \mathbf{y}_k)\pi(\mathbf{x}_{0:k-1} \,|\, \mathbf{y}_{1:k-1}), \quad \text{(B1)}$$

where we use the shorthand notation $\mathbf{x}_{0:k}$ introduced in section 4c to refer to the state at all times $t_0, t_1, \ldots, t_k$.

The corresponding expression for the unnormalized importance weights is

$$w_k^* = \frac{p(\mathbf{x}_{0:k} \mid \mathbf{y}_{1:k})}{\pi(\mathbf{x}_{0:k} \mid \mathbf{y}_{1:k})}, \tag{B2}$$

which extends (24).

We show by induction that for proposals of the form (B1), choosing $\pi(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{y}_k) = p(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{y}_k)$ at each step minimizes the variance of w_k^*, with expectation taken over $\mathbf{x}_{0:k}$ distributed as $\pi(\mathbf{x}_{0:k} \mid \mathbf{y}_{1:k})$.

The arguments of section 4b hold with $k = 1$ and the initial particles $\mathbf{x}_0$ drawn from an arbitrary proposal $\pi(\mathbf{x}_0)$ rather than $p(\mathbf{x}_0)$. This proves that the optimal proposal minimizes $\mathrm{var}(w_1^*)$.

For the general step from t_{k-1} to t_k, we let $\pi(\mathbf{x}_{0:k-1} \mid \mathbf{y}_{1:k-1})$ be fixed and examine how $\mathrm{var}(w_k^*)$ depends on the choice of $\pi(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{y}_k)$. As in section 4b, the variance can be computed from $\mathrm{var}(w_k^*) = E(w_k^{*2}) - E(w_k^*)^2$. Again,

$$E(w_k^*) = \int w_k^* \, \pi(\mathbf{x}_{0:k} \mid \mathbf{y}_{1:k}) \, d\mathbf{x}_{0:k} = 1.$$

Calculating $E(w_k^{*2})$ is facilitated by the following factorization:

$$\begin{aligned} p(\mathbf{x}_{0:k} \mid \mathbf{y}_{1:k}) &= p(\mathbf{x}_k \mid \mathbf{x}_{0:k-1}, \mathbf{y}_{1:k}) p(\mathbf{x}_{0:k-1} \mid \mathbf{y}_{1:k}) \\ &= p(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{y}_k) p(\mathbf{y}_k \mid \mathbf{x}_{k-1}) \\ &\quad \times p(\mathbf{x}_{0:k-1} \mid \mathbf{y}_{1:k-1}) / p(\mathbf{y}_k \mid \mathbf{y}_{1:k-1}), \end{aligned}$$

where the second equality follows using Bayes's rule and the usual assumptions that $\mathbf{x}_k \mid \mathbf{x}_{k-1}$ is independent of $\mathbf{x}_j$, $j < k-1$, and $\mathbf{y}_k \mid \mathbf{x}_k$ is independent of $\mathbf{x}_j$, $j < k$. Then, using (B1) and the result of appendix A, we have

$$E(w_k^{*2}) = \int \frac{p(\mathbf{y}_k \mid \mathbf{x}_{k-1})^2 p(\mathbf{x}_{0:k-1} \mid \mathbf{y}_{1:k-1})^2}{p(\mathbf{y}_k \mid \mathbf{y}_{1:k-1})^2 \pi(\mathbf{x}_{0:k-1} \mid \mathbf{y}_{1:k-1})} \left[\int \frac{p(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{y}_k)^2}{\pi(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{y}_k)} d\mathbf{x}_k \right] d\mathbf{x}_{0:k-1} \geq \int \frac{p(\mathbf{y}_k \mid \mathbf{x}_{k-1})^2 p(\mathbf{x}_{0:k-1} \mid \mathbf{y}_{1:k-1})^2}{p(\mathbf{y}_k \mid \mathbf{y}_{1:k-1})^2 \pi(\mathbf{x}_{0:k-1} \mid \mathbf{y}_{1:k-1})} d\mathbf{x}_{0:k-1}.$$

This lower bound is achieved only if $\pi(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{y}_k) = p(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{y}_k)$. Thus, regardless of how we chose $\pi(\mathbf{x}_{0:k-1} \mid \mathbf{y}_{1:k-1})$, the optimal proposal minimizes $\mathrm{var}(w_k^*)$ over all possible $\pi(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{y}_k)$. This concludes the induction.

We have shown that the proposal

$$\pi(\mathbf{x}_{0:k} \mid \mathbf{y}_{1:k}) = \pi(\mathbf{x}_0) \prod_{j=1}^{k} p(\mathbf{x}_j \mid \mathbf{x}_{j-1}, \mathbf{y}_j)$$

is the optimal importance function for the density $p(\mathbf{x}_{0:k} \mid \mathbf{y}_{1:k})$. This optimality is with respect to $\mathrm{var}(w_k^*)$, where expectations are based on $\pi(\mathbf{x}_{0:k} \mid \mathbf{y}_{1:k})$, and over all importance functions of that satisfy the recursion (B1) (i.e., over all importance function that allow for sequential sampling).

REFERENCES

Ades, M., and P. J. Van Leeuwen, 2015: The equivalent-weights particle filter in a high-dimensional system. *Quart. J. Roy. Meteor. Soc.*, **141**, 484–503, doi:10.1002/qj.2370.

Arulampalam, M., S. Maskell, N. Gordon, and T. Clapp, 2002: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Process.*, **50**, 174–188, doi:10.1109/78.978374.

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

——, P. Bickel, and B. Li, 2008: Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. *Probability and Statistics: Essays in Honor of David A. Freedman*, D. Nolan and T. Speeds, Eds., Vol. 2, Institute of Mathematical Statistics, 316–334, doi:10.1214/193940307000000518.

Bickel, P., B. Li, and T. Bengtsson, 2008: Sharp failure rates for the bootstrap particle filter in high dimensions. *Pushing the Limits of Contemporarly Statistics: Contributions in Honor of Jayanta K. Ghosh*, B. Clarke and S. Ghosal, Eds., Vol. 3, Institute of Mathematical Statistics, 318–329, doi:10.1214/074921708000000228.

Bocquet, M., C. A. Pires, and L. Wu, 2010: Beyond Gaussian statistical modeling in geophysical data assimilation. *Mon. Wea. Rev.*, **138**, 2997–3023, doi:10.1175/2010MWR3164.1.

Boffetta, G., A. Celani, A. Crisanti, and A. Vulpiani, 1997: Predictability in two-dimensional decaying turbulence. *Phys. Fluids*, **9**, 724–734, doi:10.1063/1.869227.

Chorin, A. J., and X. Tu, 2009: Implicit sampling for particle filters. *Proc. Natl. Acad. Sci. USA*, **106**, 17 249–17 254, doi:10.1073/pnas.0909196106.

——, and M. Morzfeld, 2013: Conditions for successful data assimilation. *J. Geophys. Res. Atmos.*, **118**, 11 522–11 533, doi:10.1002/2013JD019838.

——, ——, and X. Tu, 2013: A survey of implicit particle filters for data assimilation. *State Space Models: Applications in Economics and Finance*, Y. Zeng and S. Wu, Eds., Springer, 63–88.

Doucet, A., S. Godsill, and C. Andrieu, 2000: Sequential Monte-Carlo methods for Bayesian filtering. *Stat. Comput.*, **10**, 197–208, doi:10.1023/A:1008935410038.

——, M. Briers, and S. Sénécal, 2006: Efficient block sampling strategies for sequential Monte Carlo methods. *J. Comput. Graph. Stat.*, **15**, 693–711, doi:10.1198/106186006X142744.

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

Hamill, T. M., J. S. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.*, **129**, 2776–2790, doi:10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, **126**, 796–811, doi:10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2.

——, and ——, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129**, 123–137, doi:10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2.

Hunt, B. R., E. J. Kostelich, and I. Szunyogh, 2007: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D*, **230**, 112–126, doi:10.1016/j.physd.2006.11.008.

Klaas, M., A. Doucet, and N. D. Frietas, 2005: Towards practical n^2 Monte Carlo: The marginal particle filter. *Proc. 21st Annual Conf. on Uncertainty in Artificial Intelligence*, Arlington, VA, Association for Uncertainty in Artificial Intelligence, 308–315.

Kong, A., J. Liu, and W. Wong, 1994: Sequential imputations and Bayesian missing data problems. *J. Amer. Stat. Assoc.*, **89**, 278–288, doi:10.1080/01621459.1994.10476469.

Lin, M., R. Chen, and J. S. Liu, 2013: Lookahead strategies for sequential Monte Carlo. *Stat. Sci.*, **28**, 69–94, doi:10.1214/12-STS401.

Liu, J. S., and R. Chen, 1998: Sequential Monte Carlo methods for dynamic systems. *J. Amer. Stat. Assoc.*, **93**, 1032–1044, doi:10.1080/01621459.1998.10473765.

Metref, S., E. Cosme, C. Snyder, and P. Brasseur, 2014: A non-Gaussian analysis scheme using rank histograms for ensemble data assimilation. *Nonlinear Processes Geophys.*, **21**, 869–885, doi:10.5194/npg-21-869-2014.

Morzfeld, M., X. Tu, E. Atkins, and A. J. Chorin, 2012: A random map implementation of implicit filters. *J. Comput. Phys.*, **231**, 2049–2066, doi:10.1016/j.jcp.2011.11.022.

Nakano, S., 2014: Hybrid algorithm of ensemble transform and importance sampling for assimilation of non-Gaussian observations. *Tellus*, **66A**, 21429, doi:10.3402/tellusa.v66.21429.

Papadakis, N., E. Mémin, A. Cuzol, and N. Gengembre, 2010: Data assimilation with the weighted ensemble Kalman filter. *Tellus*, **62A**, 673–697, doi: 10.1111/j.1600-0870.2010.00461.x.

Parlett, B. N., 1998: *The Symmetric Eigenvalue Problem*. SIAM, xxiv + 391 pp.

Rabier, F., A. McNally, E. Andersson, P. Courtier, P. Undén, J. Eyre, A. Hollingsworth, and F. Bouttier, 1998: The ECMWF implementation of three-dimensional variational assimilation (3D-Var). Part II: Structure functions. *Quart. J. Roy. Meteor. Soc.*, **124**, 1809–1829, doi:10.1002/qj.49712455003.

Reich, S., 2013: A nonparametric ensemble transform method for Bayesian inference. *SIAM J. Sci. Comput.*, **35**, A2013–A2024, doi:10.1137/130907367.

Rotunno, R., and C. Snyder, 2008: A generalization of Lorenz's model for the predictability of flows with many scales of motion. *J. Atmos. Sci.*, **65**, 1063–1076, doi:10.1175/2007JAS2449.1.

Snyder, C., 2012: Particle filters, the ''optimal'' proposal and high-dimensional systems. *ECMWF Seminar on Data Assimilation for Atmosphere and Ocean*, Shinfield, United Kingdom, ECMWF, 161–170.

——, T. Bengtsson, P. Bickel, and J. Anderson, 2008: Obstacles to high-dimensional particle filtering. *Mon. Wea. Rev.*, **136**, 4629–4640, doi:10.1175/2008MWR2529.1.

van Leeuwen, P. J., 2009: Particle filtering in geophysical systems. *Mon. Wea. Rev.*, **137**, 4089–4114, doi:10.1175/2009MWR2835.1.

——, 2010: Nonlinear data assimilation in geosciences: An extremely efficient particle filter. *Quart. J. Roy. Meteor. Soc.*, **136**, 1991–1999, doi:10.1002/qj.699.

Weir, B., R. N. Miller, and Y. H. Spitz, 2013: A potential implicit particle method for high-dimensional systems. *Nonlinear Processes Geophys.*, **20**, 1047–1060, doi:10.5194/npg-20-1047-2013.

Corrigendum

CHRIS SNYDER

National Center for Atmospheric Research,[a] *Boulder, Colorado*

THOMAS BENGTSSON

Genentech, San Francisco, California

MATHIAS MORZFELD

Department of Mathematics, University of California, Berkeley, Berkeley, California

(Manuscript received and in final form 29 April 2016)

There is a typographical error in the text preceding (12) and preceding (16) in Snyder et al. (2015). The text preceding (12) should read, ''Now let λ_j^2 and the columns of $\mathbf{E}$ be, respectively, the eigenvalues and eigenvectors of . . .'' The text preceding (16) should read, ''In particular, the eigenvalues λ_j^2 used in (15) for the standard proposal come from the matrix . . .'' Thus, the correct text should define λ_j^2, rather than λ_j, as the eigenvalues of either $\mathrm{cov}[(\mathbf{R}+\mathbf{HQH}^{\mathrm{T}})^{-1/2}\mathbf{HM}\mathbf{x}_{k-1}]$ or $\mathrm{cov}(\mathbf{R}^{-1/2}\mathbf{H}\mathbf{x}_k)$, depending on the choice of the standard proposal distribution or the optimal proposal, respectively.

This error has no effect on the conclusions of the paper.

REFERENCE

Snyder, C., T. Bengtsson, and M. Morzfeld, 2015: Performance bounds on particle filters using the optimal proposal. *Mon. Wea. Rev.*, **143**, 4750–4761, doi:10.1175/MWR-D-15-0144.1.

[a] The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Corresponding author address: Chris Snyder, NCAR, P.O. Box 3000, Boulder, CO 80307.
E-mail: chriss@ucar.edu

DOI: 10.1175/MWR-D-16-0161.1