Multilevel ensemble transform particle filtering

Gregory, A., Cotter, C. J. and Reich, S. (2016) Multilevel ensemble transform particle filtering. SIAM Journal on Scientific Computing, 38 (3). A1317-A1338. ISSN 1064-8275 doi: https://doi.org/10.1137/15M1038232 Available at https://centaur.reading.ac.uk/67471/

SIAM J. SCI. COMPUT.
Vol. 38, No. 3, pp. A1317–A1338

MULTILEVEL ENSEMBLE TRANSFORM PARTICLE FILTERING*

A. GREGORY†, C. J. COTTER†, AND S. REICH‡

Abstract. This paper extends the multilevel Monte Carlo variance reduction technique to nonlinear filtering. In particular, multilevel Monte Carlo is applied to a certain variant of the particle filter, the ensemble transform particle filter (EPTF). A key aspect is the use of optimal transport methods to re-establish correlation between coarse and fine ensembles after resampling; this controls the variance of the estimator. Numerical examples present a proof of concept of the effectiveness of the proposed method, demonstrating significant computational cost reductions (relative to the single-level ETPF counterpart) in the propagation of ensembles.

Key words. multilevel Monte Carlo, sequential data assimilation, optimal transport

AMS subject classifications. 65C05, 62M20, 93E11, 93B40, 90C05

DOI. 10.1137/15M1038232

1. Introduction.

Data assimilation is the process of incorporating observed data into model forecasts. In data assimilation, one is interested in computing statistics $\mathbb{E}_\eta[X]$ of solutions X to random dynamical systems with respect to a posterior measure (η) given partial observations of the system. In particle filtering [7, 4], this is done by using an empirical ensemble representing the posterior distribution η at any one time. The propagation in time of the members (particles) of this ensemble can be computationally expensive, especially in high dimensional systems.

Recently, the multilevel Monte Carlo (MLMC) method has been developed for achieving significant cost reductions in Monte Carlo simulations [9]. It has been applied to areas such as Markov chain Monte Carlo [13] and quasi–Monte Carlo [10] to return computational cost reductions from existing techniques. It has also been applied to uncertainty quantification within PDEs [6]. The idea is to consider a hierarchy of discretized models, balancing numerical error in cheap/coarse models against Monte Carlo variance in expensive/fine models. It is desirable to adapt MLMC to sequential Monte Carlo methods such as particle filters, and some first steps have been taken in this direction. First, the authors of [11] have developed a multilevel ensemble Kalman filter (EnKF), using MLMC estimators to calculate the mean and covariance of the posterior, in the case where the underlying distributions are Gaussian and the model is linear. However, for non-Gaussian distributions and nonlinear models, the EnKF is biased. The method does, however, converge to a "mean-field limit" [14]. Second, the authors of [3] proposed a multilevel sequential Monte Carlo method for Bayesian inference problems to give significant computational cost reductions from standard techniques. Our goal in this paper is to take a step further by applying MLMC to nonlinear filtering problems.

In general, MLMC works by computing statistics from pairs of coarser and finer

*Submitted to the journal's Methods and Algorithms for Scientific Computing section September 3, 2015; accepted for publication (in revised form) February 22, 2016; published electronically May 3, 2016.

http://www.siam.org/journals/sisc/38-3/M103823.html

†Department of Mathematics, Imperial College London, London SW7 2AZ, UK (a.gregory14@imperial.ac.uk, colin.cotter@imperial.ac.uk). The first author's research was supported by the Science and Solutions to a Changing Planet DTP and the Natural Environmental Research Council.

‡Institut für Mathematik, Universität Potsdam, D-14469 Potsdam, Germany, and Department of Mathematics and Statistics, University of Reading, Whiteknights, Reading RG6 6AX, UK (sreich@math.uni-potsdam.de).

correlated ensembles. For Monte Carlo simulation of SDEs, this correlation is achieved by using the same initial conditions and Brownian paths for each coarse/fine pair of ensemble members. The key challenge in applying MLMC to particle filtering is in maintaining this correlation after resampling. Giles [8] suggested that correlating coarse and fine ensembles could be achieved by minimizing the Wasserstein distance between the two ensembles. This can be formulated as a optimal transportation problem [20].

In this paper, we adapt the MLMC framework to the ensemble transform particle filter (ETPF) [19]. The ETPF is an efficient and effective nonlinear filter that uses optimal transportation transformations [22] instead of random resampling. In our multilevel ETPF (MLETPF) the coupling between coarse and fine ensembles is also maintained using optimal transportation. The sole aim of introducing MLMC to the ETPF is to reduce the computational cost of the propagation of particles. This is only a benefit if the computational cost dominates the optimal transportation transformation cost; while direct solvers for optimal transportation problems with one dimensional state space scale as $O\big(N\log(N)\big)$, solvers for problems with more than one dimension scale as $O\big(N^3\log(N)\big)$ with the ensemble size. To address this, a technique commonly used in the aforementioned EnKF known as localization can be used to reduce this optimal transportation cost significantly [5]. Our proposed MLETPF can return significant reductions in the overall computational cost of the ETPF where the particle propagation cost dominates. It will also return significant reductions in cases where optimal transportation computational cost dominates if the localized ETPF is used.

This paper proceeds as follows. Section 2 provides a background of the MLMC method, and section 3 describes the basic particle filtering framework together with the ETPF scheme. Then the proposed MLETPF method is presented in section 4, along with numerical examples demonstrating the effectiveness of the method. Finally, section 5 provides a summary and outlook.

2. The multilevel Monte Carlo method. The multilevel Monte Carlo estimator can be viewed as a variance reduction technique for a standard Monte Carlo estimator. Suppose one wishes to compute an approximation of $\mathbb{E}[X_L]$, where X_L is a numerical approximation of a random variable X (with discretization accuracy parameter[1] $h_L \propto M^{-L}$). The multilevel Monte Carlo (MLMC) method introduced in [9] considered the case where X is the solution to an SDE at time $T > 0$; the discretized solutions X_L are obtained from a given numerical method with stepsize h_L. This paper will instead consider X to be a solution, at time $T > 0$, to a general random dynamical system, with stochastic forcing and/or random initial conditions (drawn from a distribution π^0). In the simplest case let X_L^i, $i = 1, \ldots, N$, be $N \geq 1$ independent and identically distributed (i.i.d.) samples of X_L. The standard, unbiased, Monte Carlo estimator for $\mathbb{E}[X_L]$ is then

$$\bar{X}_L^{MC} = \frac{1}{N}\sum_{i=1}^{N} X_L^i. \tag{2.1}$$

Using a telescoping sum of expectations,

$$\mathbb{E}[X_L] = \mathbb{E}[X_0] + \sum_{l=1}^{L} \mathbb{E}[X_l] - \mathbb{E}[X_{l-1}], \tag{2.2}$$

[1] An example is the time stepsize.

one can define the MLMC approximation to $\mathbb{E}[X_L]$ as a sum of independent Monte Carlo estimators, $\bar{X}_L = \sum_{l=0}^{L} \hat{X}_l$, where

$$\hat{X}_l = \begin{cases} \sum_{i=1}^{N_0} \frac{1}{N_0} X_0^i, & l = 0, \\ \sum_{i=1}^{N_l} \frac{1}{N_l} \left(X_l^i - X_{l-1}^i\right), & l > 0, \end{cases} \tag{2.3}$$

leading to

$$\bar{X}_L = \frac{1}{N_0} \sum_{i=1}^{N_0} X_0^i + \sum_{l=1}^{L} \left(\frac{1}{N_l} \sum_{i=1}^{N_l} (X_l^i - X_{l-1}^i) \right). \tag{2.4}$$

Here N_l, $l = 0, \ldots, L$, are Monte Carlo sample sizes, of i.i.d. draws from X_l, X_{l-1}, respectively, for each of the $L+1$ estimators. We have

$$\mathbb{E}[\hat{X}_l] = \begin{cases} \mathbb{E}[X_0], & l = 0, \\ \mathbb{E}[X_l] - \mathbb{E}[X_{l-1}], & l > 0, \end{cases} \tag{2.5}$$

and hence the MLMC estimator is an unbiased approximation of $\mathbb{E}[X_L]$. We will call the estimators, $\hat{X}_l$, $l > 0$, "multilevel difference estimators." The important thing to note here is that the fine (level l) and coarse (level $l-1$) samples in each difference estimator must be positively correlated for each i in each of the $L+1$ multilevel difference estimators. This can be achieved by using the same random system input (e.g., initial conditions/stochastic forcing) for each i on both levels. On the other hand, the samples in different difference estimators must be uncorrelated.

For fixed T, the discretization bias (away from $\mathbb{E}[X]$) of the overall estimator is $O(h_L^\alpha)$ [9], where α is the global discretization bias (i.e., $|\mathbb{E}[X_L] - \mathbb{E}[X]|$) of the numerical method used to simulate X_l, $l \geq 0$. One notes from [9] that

$$|\mathbb{E}[X_l] - \mathbb{E}[X_{l-1}]| \leq (M-1) c h_l^\alpha, \tag{2.6}$$

where c is a positive constant, and that $(M-1)^{-1}|\mathbb{E}[X_L] - \mathbb{E}[X_{L-1}]|$ can be used as an estimate for the overall discretization bias, $\left|\bar{X}_L - \mathbb{E}[X]\right|$. As each estimator in (2.4) is independent of one another, the overall variance is given by the sum of the variances of each individual estimator. Given that there is a positive correlation between X_l and X_{l-1}, one can then expect that the sample variance of $X_l - X_{l-1}$, denoted by

$$V_l = \mathbb{V}[X_l - X_{l-1}] = \mathbb{V}[X_l] + \mathbb{V}[X_{l-1}] - 2\mathbb{C}\mathrm{ov}[X_l, X_{l-1}], \tag{2.7}$$

decays at a rate proportional to l, so that $\mathbb{V}_l = O(h_l^\beta)$, $\beta > 0$. The covariance in the last term in (2.6) is taken over the joint probability distribution of X_l and X_{l-1}. One can then trade off variance in fine/expensive estimators against discretization error in coarse/cheap estimators with lower variance by setting a decreasing sequence of Monte Carlo estimator sample sizes, $N_0 > N_1 > \cdots > N_L$. The overall computational cost of the MLMC estimator is

$$C_{ML} = \sum_{l=0}^{L} h_l^{-\gamma} N_l T, \tag{2.8}$$

where $h_l^{-\gamma}$ defines the computational cost of propagating one single sample (of X_l) through a discretized system. On the other hand, the cost of the standard Monte Carlo estimator in (2.1) is

$$C_{MC} = h_L^{-\gamma} N T. \tag{2.9}$$

One can choose the continuous variables N_l, $l = 0, \ldots, L$, at a rate that minimizes the variance of the MLMC estimator for a fixed computational cost, $N_l \propto \sqrt{V_l h_l^{\gamma}}$. In particular, by following a formula given in [9, 6], one can find optimal values of N_l, as well as the finest level L, and in doing so achieve a computational cost reduction relative to the standard Monte Carlo counterpart (2.1), with the same bound mean square error. Giles [9] proved the following result.

THEOREM 2.1. *If the mean square error of $\bar{X}_L$ is bounded by $O(\epsilon^2)$, one can optimally choose L and N_l to allow the computational cost of the MLMC estimator to be bounded by*

$$C_{ML} \leq \begin{cases} c_1 \epsilon^{-2}, & \gamma < \beta, \\ c_2 \epsilon^{-2} \log(\epsilon)^2, & \gamma = \beta, \\ c_3 \epsilon^{-2-\frac{(\gamma-\beta)}{\alpha}}, & \gamma > \beta, \end{cases} \tag{2.10}$$

where c_1, c_2, c_3, γ, β are positive constants and $\alpha \geq \frac{1}{2}\min(\gamma, \beta)$.

For $\bar{X}_L^{MC}$ to have a mean square error of $O(\epsilon^2)$, a sample size N of $O(\epsilon^{-2})$ is required, as well as a discretization bias given by $h_L = O(\epsilon^{\frac{1}{\alpha}})$. Thus any of the computational costs in Theorem 2.1 are less than C_{MC} $(O(\epsilon^{-2-\frac{\gamma}{\alpha}}))$. The MLMC approach in principle is very simple to implement and can be very effective as long as one can satisfy the two constraints, $\alpha \geq \frac{1}{2}\min(\gamma, \beta)$ and $\beta > 0$. More detail on this method can be found in a generalized explanation, and related theorems, in [6].

3. Particle filtering. This section will outline the standard particle filtering methodology. In this context, one is interested in computing statistics of a random process X_t, conditioned on observations of a single realization of X, denoted by X', and referred to as the reference solution. The observations are random variables of the form

$$Y_{t_k} = H(X'_{t_k}) + \phi \tag{3.1}$$

at times $t \in (t_1, \ldots, t_{N_y})$, where $H : \mathbb{R}^d \to \mathbb{R}^y$ is an observation operator, and ϕ is a random variable representing measurement error. For simplicity, we choose $\phi \sim N(0, R)$, where R is a $y \times y$ covariance matrix, and we will take H to be the identity (so that $y = d$). We define X_{L,t_k} to be a numerical discretization of X_{t_k} with discretization accuracy parameter h_L. Our aim here is to sequentially approximate $\mathbb{E}_{\eta_{L,t_k}}[X_{L,t_k}]$, the expectation of X_{L,t_k} with respect to the measure η_{L,t_k}, where η_{L,t_k} is the posterior of $X_{L,t}$ given the observations $Y_{t_1,\ldots,t_k}$. Let $p(y|x)$ be the likelihood function of y given x and $q(x)$ be the prior of x. Then, for any $k \in [1, N_y]$, using importance sampling [7], one can draw N i.i.d. samples from the empirical approximation of the prior of X_{L,t_k},

$$\hat{q}(X_{L,t_k}) = \sum_{i=1}^{N} \tilde{w}_L(X^i_{L,t_{k-1}}) \delta(X_{L,t_k} - X^i_{L,t_k}), \tag{3.2}$$

and denote the normalized importance weights of sample $i \in [1, N]$ to be

$$\tilde{w}_L(X^i_{L,t_k}) = \frac{w_L(X^i_{L,t_k})}{\sum_{j=1}^{N} w_L(X^j_{L,t_k})}, \tag{3.3}$$

where

$$w_L(X^i_{L,t_k}) = p(Y_{t_k}|X^i_{L,t_k}) \tilde{w}_L(X^i_{L,t_{k-1}}). \tag{3.4}$$

Thus, the filter weights are defined iteratively, starting from $\tilde{w}_L(X^i_{L,t_0}) = \frac{1}{N}$. As the observations are given by a Gaussian distribution, the likelihood is

$$p(Y_{t_k}|X^i_{L,t_k}) = \frac{1}{\sqrt{2\pi}|R|^{y/2}} e^{-\frac{1}{2}\left(H(X^i_{L,t_k})-Y_{t_k}\right)^T R^{-1}\left(H(X^i_{L,t_k})-Y_{t_k}\right)}. \tag{3.5}$$

Finally, an estimator for the expectation of X_{L,t_k} with respect to the posterior η_{L,t_k} is

$$\bar{X}_{L,t_k} = \sum_{i=1}^{N} \tilde{w}_L(X^i_{L,t_k}) X^i_{L,t_k}. \tag{3.6}$$

This estimator, despite being biased by $O(N^{-1})$ due to the normalized importance weights in a single importance sampling update, is consistent with $\mathbb{E}_{\eta_{L,t_k}}[X_{L,t_k}]$. This is to say, as $N \to \infty$, the estimator converges in probability to $\mathbb{E}_{\eta_{L,t_k}}[X_{L,t_k}]$.

Typically, importance weights become degenerate as k increases [4]. In this case, it is necessary to duplicate higher weighted particles while removing lower weighted particles; this is known as resampling. Resampling resembles an unbiased transformation from the weighted ensemble, $\left\{X^i_{l,t_k}, \tilde{w}_l(X^i_{l,t_k})\right\}_{i=1,\dots,N}$ to an evenly weighted ensemble of resampled particles $\left\{\tilde{X}^i_{l,t_k}\right\}_{i=1,\dots,N}$. The scheme outlined above is known as the sequential importance resampling (SIR) method. For more information on this, see [7]. The ETPF, the subject of this paper, uses optimal transportation [22] to implement this transformation, which we describe next.

3.1. Ensemble transform particle filters. ETPFs are a variant of linear ensemble transform filters (LETFs) [19]. They present an alternative to the resampling step that takes place in the standard SIR methodology, replacing it with a linear transformation. The goal is to obtain a transformed set of evenly weighted particles, $\left\{\tilde{X}^i_{L,t_k}\right\}_{i=1,\dots,N}$, from the weighted set of particles $\left\{X^i_{L,t_k}\right\}_{i=1,\dots,N}$, with importance weights $\left\{\tilde{w}_L(X^i_{L,t_k})\right\}_{i=1,\dots,N}$, defining an empirical approximation to the posterior distribution η_{L,t_k}. This can be done with the following linear transformation:

$$\tilde{X}^j_{L,t_k} = \sum_{i=1}^{N} P_{i,j} X^i_{L,t_k} \tag{3.7}$$

for $i = 1, \dots, N$ and $j = 1, \dots, N$ with nonzero entries for $P_{i,j}$. Here $\sum_{i=1}^{N} P_{i,j} = 1$. Let Z_{L,t_k} denote the discrete random variable with samples X^i_{L,t_k} and associated probability vector $\tilde{w}_L(X^i_{L,t_k})$, $i = 1, \dots, N$. Then take $\tilde{Z}_{L,t_k}$ to be the discrete random variable with samples X^i_{L,t_k}, $i = 1, \dots, N$, all with equal probability. For the ETPF, one creates a coupling between Z_{L,t_k} and $\tilde{Z}_{L,t_k}$, denoted by the matrix $T_{i,j}$, of size $N \times N$, with nonnegative entries. The coupling defines the linear transformation matrix in (3.7) as $P_{i,j} = NT_{i,j}$. This coupling can be found by solving a linear transport problem by minimizing the expected Euclidean distance between Z_{L,t_k} and $\tilde{Z}_{L,t_k}$, subject to the constraints

$$\sum_{i=1}^{N} T_{i,j} = \frac{1}{N}, \quad \sum_{j=1}^{N} T_{i,j} = \tilde{w}_L(X^i_{L,t_k}). \tag{3.8}$$

This is in fact equivalent to maximizing the covariance between the two ensembles since

$$\begin{aligned}\mathbb{E}_{Z_{L,t_k},\tilde{Z}_{L,t_k}}[\|z_{L,t_k}-\tilde{z}_{L,t_k}\|^2] &= \mathbb{E}_{Z_{L,t_k}}[\|z_{L,t_k}\|^2]+\mathbb{E}_{\tilde{Z}_{L,t_k}}[\|\tilde{z}_{L,t_k}\|^2]\\ \cdots &- 2\mathbb{E}_{Z_{L,t_k}}[z_{L,t_k}]^T\mathbb{E}_{\tilde{Z}_{L,t_k}}[\tilde{z}_{L,t_k}]-2\mathrm{Tr}\big(\mathrm{Cov}_{Z_{L,t_k},\tilde{Z}_{L,t_k}}[z_{L,t_k},\tilde{z}_{L,t_k}]\big).\end{aligned} \tag{3.9}$$

In a univariate case, we define an optimal coupling matrix, $T_{i,j}$, as one which minimizes the cost function,

$$\sum_{i=1}^{N}\sum_{j=1}^{N}T_{i,j}(X^i_{L,t_k}-X^j_{L,t_k})^2. \tag{3.10}$$

Theoretical analysis of the above transformation is given in [19]. Once the transformed particles in (3.7) are found, the posterior mean is now estimated by,

$$\bar{X}_{L,t_k}=\frac{1}{N}\sum_{j=1}^{N}\tilde{X}^j_{L,t_k}. \tag{3.11}$$

It is important to note that this linear transformation, which is deterministic, will give the same estimator as in (3.6) and thus does not add considerable extra variance to the estimator from random resampling. This is a consistent estimator for the previous posterior mean estimator $\big(\sum_{i=1}^{N}\tilde{w}_L(X^i_{L,t_k})X^i_{L,t_k}\big)$ since

$$\begin{aligned}\bar{X}_{L,t_k}&=\frac{1}{N}\sum_{j=1}^{N}\tilde{X}^j_{L,t_k}=\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{N}T_{i,j}NX^i_{L,t_k}=\sum_{j=1}^{N}\sum_{i=1}^{N}T_{i,j}X^i_{L,t_k}\\ &=\sum_{i=1}^{N}\tilde{w}_L(X^i_{L,t_K})X^i_{L,t_k}.\end{aligned} \tag{3.12}$$

In the univariate case, the matrix $T_{i,j}$ can easily be found by an $O\big(N\log(N)\big)$ algorithm in [20]. This will become an important observation when discussing localization in the next section. The above constraints lead to a maximum of $2N-1$ nonzero elements in $T_{i,j}$, leading to a very sparse matrix calculation, and thus the ensemble transformation process can be achieved in an $O(N)$ computational cost. The $O\big(N\log(N)\big)$ computational cost comes from the fact that one has to sort the univariate particles prior to the algorithm. In our numerical experiments at the end of this paper, this sorting was a negligible part of the ensemble transform computational cost. This allows one to be able to carry the ensemble transform out on every assimilation step without the computational expense of this being of a higher order of magnitude than the propagation of the particles in between assimilation steps, but more analysis will cover this observation in the next section.

In the multivariate case, the same linear transport problem prevails; however, one is required to minimize the cost function,

$$\sum_{i=1}^{N}\sum_{j=1}^{N}T_{i,j}\left\|X^i_{L,t_k}-X^j_{L,t_k}\right\|^2, \tag{3.13}$$

whose minimum defines the Wasserstein distance between Z_{L,t_k} and $\tilde{Z}_{L,t_k}$. This can be solved in an $O\big(N^3\log(N)\big)$ computational cost, using algorithms such as the

FastEMD algorithm [16]. However, this means that with many systems this optimal transportation computational cost will dominate over model costs of the system, and thus the scheme is not efficient. The model costs of the particle filter are defined to be the computational cost needed to propagate the N particles through the system in between assimilation steps and is given in (2.9) (determined by a constant γ). Thankfully, a technique called *localization* can aid this problem, and it can also provide a pivotal change to the scheme when applying it to high dimensional systems.

3.1.1. Localization. Localization, a scheme frequently used in the EnKF for high dimensional systems, can also be applied to the ETPF [5]. In the simplest form, localization applied to the ETPF means that one can reduce the computational cost of designing a multivariate coupling to d times the cost of designing a univariate coupling. Localization allows one to construct an individual transformation in (3.7) for each of the d components of a multivariate X_{L,t_k}. A simple definition for the localization matrix C [5] that describes the spatial correlation structure of the ensemble $\{X_{L,t_k}\}_{i=1,\dots,N}$ could be

$$C_{m,n} = \begin{cases} 1 - \frac{1}{2}\left(\frac{s_{m,n}}{r_{loc,c}}\right), & \left(\frac{s_{m,n}}{r_{loc,c}}\right) \le 2, \\ 0 & \text{otherwise.} \end{cases} \tag{3.14}$$

Here $m, n = 1, \dots, d$ are the indices of the spatial components of X_{L,t_k},

$$s_{m,n} = \min\left\{|m-n-N|, |m-n|, |m-n+N|\right\}, \tag{3.15}$$

and $r_{loc,c}$ is a constant. The above form for $C_{m,n}$ explicitly takes spatial-periodicity into account. One can now decompose the linear transport problem in (3.13) into d separate linear transport problems, to find a coupling matrix $T_{i,j}(m)$, $i = 1, \dots, N$, $j = 1, \dots, N$, for each component $m = 1, \dots, d$. The objective of these linear transport problems is minimizing the cost function

$$\sum_{i=1}^{N}\sum_{j=1}^{N} T_{i,j}(m)\left\|X^i_{L,t_k} - X^j_{L,t_k}\right\|^2_m, \tag{3.16}$$

where

$$\left\|X^i_{L,t_k} - X^j_{L,t_k}\right\|^2_m = \sum_{n=1}^{d} C_{m,n}\left(X^i_{L,t_k}(n) - X^j_{L,t_k}(n)\right)^2, \tag{3.17}$$

subject to the constraints

$$\sum_{i=1}^{N} T_{i,j}(m) = \frac{1}{N}, \quad \sum_{j=1}^{N} T_{i,j}(m) = \tilde{w}_L(X^i_{L,t_k}). \tag{3.18}$$

Here $X^i_{L,t_k}(m)$ is the mth component of X^i_{L,t_k}. Then one can define the approximation of the marginal posterior mean for each $m = 1, \dots, N$ as

$$\bar{X}_{L,t_k}(m) = \frac{1}{N}\sum_{j=1}^{N} \tilde{X}^j_{L,t_k}(m), \tag{3.19}$$

where the transformed components are given by

$$\tilde{X}^j_{L,t_k}(m) = \sum_{i=1}^{N} P_{i,j}(m) X^i_{L,t_k}(m), \tag{3.20}$$

and $P_{i,j}(m) = T_{i,j}(m)N$. Note that the cost functions in (3.17) do not achieve the minimum of (3.13). When $r_{loc,c} = 0$, exhibiting the most computationally efficient scenario, one has the interesting case where d univariate linear transport problems need to be solved, thus transforming all components individually. One can simply use the univariate algorithm in [20], mentioned in the previous section, with a computational cost of $O\big(N\log(N)\big)$, for each linear transport problem, to get an overall $O\big(dN\log(N)\big)$ computational cost. In practice, when $r_{loc,c} = 0$, one can also reorder each of the transformed sets of components into the rank structure of the original ensemble. This preserves the copula structure [21] of the original ensemble.

If the model costs of a system are less than that of the multivariate optimal transportation, using $r_{loc,c} = 0$ is the only case in which the model costs can return to being the dominative cost in the ETPF estimator. Despite this, it is very reasonable to imagine that this model cost will dominate that of the optimal transportation in some systems, especially for high dimensional PDEs (i.e., where γ is high). Localization does have an effect on the performance of the ETPF by adding bias into the posterior mean approximation stemming from the fact that one is generating deterministic couplings $T_{i,j}$ that are minimizing different (simplified) cost functions to the full, multivariate one in (3.13). This bias is thus caused by the decay in correlation between the components. Despite this, numerical experiments conducted in [5] find the localized ETPF to be effective even in the chaotic, highly nonlinear Lorenz equations.

Localization is also needed in the likelihood evaluation of multivariate particles in the ETPF. Although this is not critical to the aim of this paper, only briefly covered here, it is essential for the ETPF to be able to successfully filter high dimensional systems due to the curse of dimensionality. Standard sequential Monte Carlo (SMC) methods fail to track high dimensional systems due to exponentially degenerate importance weights. However, while there have only been some suggestions for a solution to this problem in SMCs, such as in [18], one can alter the above localization scheme in the ETPF to solve this problem swiftly [5]. It is also needed to reduce the computational cost of likelihood evaluations when the dimension of the state space is greater than the sample size ($d > N$). For each component ($m = 1, \dots, d$) in the particles $\big\{X^i_{L,t_k}\big\}_{i=1,\dots,N}$, generate an separate importance weight given by

$$w_L(X^i_{L,t_k}(m)) \propto \frac{1}{\sqrt{2\pi}|R|^{y/2}} e^{-\frac{1}{2}\left(H(X^i_{L,t_k})-Y_{t_k}\right)^T(\tilde{C}_m)R^{-1}\left(H(X^i_{L,t_k})-Y_{t_k}\right)}, \tag{3.21}$$

where

$$(\tilde{C}_m)_{n,n} = \begin{cases} 1 - \frac{1}{2}\big(\frac{s_{m,n}}{r_{loc,R}}\big), & \big(\frac{s_{m,n}}{r_{loc,R}}\big) \le 2, \\ 0 & \text{otherwise} \end{cases} \tag{3.22}$$

for $n = 1, \dots, N$ ($\tilde{C}_m$ is diagonal) and the value of $r_{loc,R}$ can be independent of $r_{loc,c}$. Of course, H should be a local operator; see [1] for details of the use of localization within the EnKF. These weights are then used in the constraints in the linear transport problems for each individual component transformation in (3.18). The two "radii" of localization, $r_{loc,c}$ and $r_{loc,R}$, will henceforth be referred to as the particular settings of localization used.

4. Multilevel ensemble transform particle filter (MLETPF). The proposed MLETPF framework is demonstrated in this section. It creates an estimator consistent with the standard ETPF estimator in (3.6), for the same discretization accuracy level, L. The term "single level" estimator will henceforth be a reference

to the corresponding standard ETPF estimator, conditioned on the same observations, with the same discretization level L and variance as the proposed MLETPF estimator. The general premise of the MLETPF is to run $L+1$ independent ETPF estimators, with N_l samples, forward in time (and space), in the coupled multilevel framework. When updating the weights of each particle in each estimator, the same method as the ETPF holds for each of the $L+1$ estimators. Thus, we define the MLETPF estimator of $\mathbb{E}_{\eta_{L,t_k}}[X_{L,t_k}]$ as the following, where the importance weights $\tilde{w}_l(X_{l,t_k})$ target η_{l,t_k}, the posterior of each d-dimensional discretization X_{l,t_k}, given the observations $Y_{t_1,\ldots,t_k}$, $k \in [1, N_y]$:

(4.1)
$$\bar{X}_{L,t_k} = \left(\sum_{i=1}^{N_0} \tilde{w}_0(X^i_{0,t_k}) X^i_{0,t_k} \right) + \sum_{l=1}^{L} \left(\sum_{i=1}^{N_l} \left(\tilde{w}_l(X^i_{l,t_k}) X^i_{l,t_k} - \tilde{w}_{l-1}(X^i_{l-1,t_k}) X^i_{l-1,t_k} \right) \right).$$

We assume here that $h_0 \leq \Delta t$, where $\Delta t = t_{k+1} - t_k$ for all $k \in [1, N_y - 1]$, so that all of the $L+1$ estimators are conditioned on the same observations. This does mean that one has to set a bound on the frequency of the data assimilation, given the time-step of the minimum level, h_0. We note that it is possible to adjust the framework here slightly to incorporate frequent observations only available on finer levels at certain times. This could be done by only updating the weights for finer ensembles at those observations and then proceeding with the ensemble transform stages when both the coarse and the fine levels in each difference estimator have had an importance weight update.

One notes that as the standard ETPF estimator for each level of discretization, $l \geq 0$, is consistent with $\mathbb{E}_{\eta_{l,t_k}}[X_{l,t_k}]$, the above estimator is consistent with the $\mathbb{E}_{\eta_{L,t_k}}[X_{L,t_k}]$, given the linearity of expectation shown in (2.2). Here each of the particles from the fine and coarse ensembles in each of the multilevel difference estimators are positively correlated in between assimilation steps, as in the standard MLMC method. This correlation is required for the variance of each difference estimator to decay with $l \to \infty$, as discussed in the opening section. However, now in the ETPF context, when one comes to transform the fine and coarse ensembles in each multilevel difference estimator, the two ensembles cannot be transformed independently of one another, and they need to have a positive correlation imparted between them ready for the next phase of particle propagation, especially if the transformations are happening frequently. If the random input to the system is simply a random initial condition, in a system with no stochastic forcing, these particles from the fine and coarse ensembles will certainly diverge instantly if they are not positively correlated after the ensemble transformations. In this paper, this positive correlation is achieved using a multilevel coupling step after the standard ensemble transform stage. This requires one to first carry out the ensemble transform (3.7) on the coarse and fine ensembles, $\{X^i_{l-1,t_k}\}_{i=1,\ldots,N_l}$, $\{X^i_{l,t_k}\}_{i=1,\ldots,N_l}$, with weights $\{w_{l-1}(X^i_{l-1,t_k})\}_{i=1,\ldots,N_l}$, $\{w_l(X^i_{l,t_k})\}_{i=1,\ldots,N_l}$, respectively, to get evenly weighted particles, $\{\tilde{X}^i_{l-1,t_k}\}_{i=1,\ldots,N_l}$ and $\{\tilde{X}^i_{l,t_k}\}_{i=1,\ldots,N_l}$. If localization is needed, one can implement this with the required parameters on both ensemble transforms, as in the last section. It is very important that the same localization settings are used on all estimators, so that the overall MLETPF estimator is consistent with the single level ETPF estimator with the same localization settings. The key point is that the fine and coarse ensembles from each discretization level will have the same systematic localization bias as one another. This means, such as with the discretization bias, that the localization biases can cancel each other out in the telescoping sum of estimators (4.1), leaving only the

systematic localization bias of the finest level equal to that of the localized single level estimator. At this point, one notes that (4.1) becomes

$$\bar{X}_{L,t_k} = \left(\frac{1}{N_0}\sum_{i=1}^{N_0} \tilde{X}^i_{0,t_k}\right) + \sum_{l=1}^{L}\left(\frac{1}{N_l}\sum_{i=1}^{N_l}\left(\tilde{X}^i_{l,t_k} - \tilde{X}^i_{l-1,t_k}\right)\right). \tag{4.2}$$

Now one needs to positively couple the fine and coarse ensembles of transformed particles from each estimator above. We propose building another coupling between $\tilde{X}_{l,t_k}$ and $\tilde{X}_{l-1,t_k}$, denoted by $T^{F/C}_{i,j}$, that minimizes the cost function

$$\sum_{i=1}^{N_l}\sum_{j=1}^{N_l} T^{F/C}_{i,j}\left\|\tilde{X}^i_{l,t_k} - \tilde{X}^j_{l-1,t_k}\right\|^2, \tag{4.3}$$

with constraints

$$\sum_{i=1}^{N} T^{F/C}_{i,j} = \frac{1}{N_l}, \qquad \sum_{j=1}^{N} T^{F/C}_{i,j} = \frac{1}{N_l}. \tag{4.4}$$

This is an assignment problem, and in the multivariate case it can be solved by the Hungarian algorithm [15] with a computational cost equal to the multivariate linear transport problem algorithms discussed previously, and hence it is the same order of magnitude as the corresponding ensemble transform stage in the standard ETPF method. In the univariate case, one can simply use the cheap algorithm in [20], exactly like the ensemble transform stage. One notes that the above assignment problem returns a coupling with one element in each row and column ($\frac{1}{N_l}$), resulting in particles simply being reordered and not transformed. This therefore returns exactly the same transformed particles in each ensemble. The reordering can be seen as finding the transformation matrix $P^{F/C}_{i,j} = T^{F/C}_{i,j}N_l$ and then applying the standard ensemble transform in (3.7) to both the fine and the coarse transformed ensembles to get new ensembles of $\{\tilde{\tilde{X}}^i_{l,t_k}\}_{i=1,\dots,N_l}$ and $\{\tilde{\tilde{X}}^i_{l-1,t_k}\}_{i=1,\dots,N_l}$ which are now positively correlated. Each multilevel difference estimator can now be estimated by

$$\hat{X}_{l,t_k} = \frac{1}{N_l}\sum_{j=1}^{N_l}\left(\tilde{\tilde{X}}^j_{l,t_k} - \tilde{\tilde{X}}^j_{l-1,t_k}\right). \tag{4.5}$$

Using a calculation similar to (3.12), we now show that the estimator in (4.5) is consistent with the term

$$\left(\sum_{i=1}^{N_l}\left(\tilde{w}_l(X^i_{l,t_k})X^i_{l,t_k} - \tilde{w}_{l-1}(X^i_{l-1,t_k})X^i_{l-1,t_k}\right)\right) \tag{4.6}$$

from (4.1). Let $T^F_{i,j}$ and $T^C_{i,j}$ $(i,j = 1,\dots,N)$ be the coupling matrices used for the ensemble transform on the finer and coarse ensembles, respectively; then

(4.7)

$$\hat{X}_{l,t_k} = \frac{1}{N_l}\sum_{j=1}^{N_l}\left(\tilde{\tilde{X}}^j_{l,t_k} - \tilde{\tilde{X}}^j_{l-1,t_k}\right) = \sum_{j=1}^{N_l}\sum_{i=1}^{N_l} T^{F/C}_{i,j}\left(\tilde{X}^i_{l,t_k} - \tilde{X}^i_{l-1,t_k}\right)$$

$$= \frac{1}{N_l}\sum_{i=1}^{N_l}\sum_{j=1}^{N_l} N_l\left(T^F_{i,j}X^j_{l,t_k} - T^C_{i,j}X^j_{l-1,t_k}\right) = \sum_{j=1}^{N_l}\left(\tilde{w}_l(X^j_{l,t_k})X^j_{l,t_k} - \tilde{w}_{l-1}(X^j_{l-1,t_k})X^j_{l-1,t_k}\right).$$

The estimator in (4.2) can therefore be written as

$$\bar{X}_{L,t_k} = \left(\frac{1}{N_0}\sum_{i=1}^{N_0} \tilde{\tilde{X}}^i_{0,t_k}\right) + \sum_{l=1}^{L}\left(\frac{1}{N_l}\sum_{i=1}^{N_l}\left(\tilde{\tilde{X}}^i_{l,t_k} - \tilde{\tilde{X}}^i_{l-1,t_k}\right)\right), \tag{4.8}$$

with covariance

$$\mathbb{C}\mathrm{ov}[\bar{X}_{L,t_k}] = \sum_{l=0}^{L}\frac{\mathbb{V}_l}{N_l}, \tag{4.9}$$

where

$$\mathbb{V}_l = \begin{cases} \mathbb{C}\mathrm{ov}[\tilde{\tilde{X}}_{0,t_k}], & l = 0, \\ \mathbb{C}\mathrm{ov}[\tilde{\tilde{X}}_{l,t_k} - \tilde{\tilde{X}}_{l-1,t_k}], & l > 0, \end{cases} \tag{4.10}$$

due to the fact that each estimator in (4.8) is independent. The above estimator is consistent with the single level (localized, with the same settings as used in the MLETPF) ETPF estimator due to (4.7) and (4.1). Therefore, in the absence of localization, it is also consistent with $\mathbb{E}_{\eta_{L,t_k}}[X_{L,t_k}]$.

The multilevel coupling $T^{F/C}_{i,j}$ minimizes the expected distance between the two transformed ensembles, which maximizes the covariance between them via (3.9) and then finally minimizes V_l. The multilevel coupling procedure above minimizes $\mathbb{V}_l$ in each multilevel difference estimator; we also ensure that pairs of particles in the coarse and fine ensembles are positively coupled in between assimilation steps (by using the same random input). The aim of this is to make the covariance $\mathbb{V}_l$ decrease at an asymptotic rate $O(h_l^\beta)$ ($\beta > 0$) required for the variance reduction of the multilevel framework to work. This is because the fine and coarse ensembles are coupled both in between assimilation steps and during them via the coupling. This will be demonstrated in numerical experiments later in the paper. Designing this coupling between both transformed ensembles $\tilde{X}_{l-1,t_k}$ and $\tilde{X}_{l,t_k}$ is the key to the proposed MLETPF method, and it enforces correlation amongst both transformed ensembles while remaining consistent with the single level ETPF estimator. Most importantly, it also suits the ETPF method since the coupling can be generated simply and cheaply when using the $r_{loc,c} = 0$ localization that can be used freely in the ETPF. This will be shown later.

4.1. Algorithm. In this section, we present an algorithm to implement the MLETPF in practice. In this paper, a predefined recurrence relation for the decay of N_l as $l \to \infty$ will be set ($N_{l+1} = f(N_l)$), and these sample sizes will be kept fixed throughout the filtering process. The finest discretization level $L > 0$ will be kept arbitrary for now. The algorithm is now presented:

1. Start at t_0 (thus $k = 0$) and with $l = 0$. Choose N_0.
2. Calculate $N_l = \lceil f(N_{l-1})\rceil$ if $l > 0$. Sample $\left\{X^i_{0,t_0}\right\}_{i=1,\dots,N_0} \sim \pi^0$ if $l = 0$ or $\left\{X^i_{l,t_0}, X^i_{l-1,t_0}\right\}_{i=1,\dots,N_l} \sim \pi^0$, where $X^i_{l,t_0} = X^i_{l-1,t_0}$ if $l > 0$, at time t_0.
3. Propagate all samples forward according to system dynamics until time t_{k+1}. If $l > 0$, the fine and coarse pairs of samples in each estimator must be coupled by using the same random input.
4. Derive the normalized importance weights for $X_{l,t_{k+1}}$:

$$\tilde{w}_l(X^i_{l,t_{k+1}}) = \frac{w_l(X^i_{l,t_{k+1}})}{\sum_{j=1}^{N_l} w_l(X^j_{l,t_{k+1}})}. \tag{4.11}$$

5. Transform $\{X^i_{l,t_k}, \tilde{w}_l(X^i_{l,t_k})\}_{i=1,\dots,N_l}$ and $\{X^i_{l-1,t_k}, \tilde{w}_{l-1}(X^i_{l-1,t_k})\}_{i=1,\dots,N_l}$ into the evenly weighted ensembles, $\{\tilde{X}^i_{l,t_k}\}_{i=1,\dots,N_l}$ and $\{\tilde{X}^i_{l-1,t_k}\}_{i=1,\dots,N_l}$, using the linear transformation in (3.7). Couple them using the multilevel coupling matrix $T^{F/C}_{i,j}$ to produce reordered ensembles $\{\tilde{\tilde{X}}^i_{l,t_k}\}_{i=1,\dots,N_l}$ and $\{\tilde{\tilde{X}}^i_{l-1,t_k}\}_{i=1,\dots,N_l}$.

6. Move on to step 7 if $l = L$, or if not, iterate $l+ = 1$ and then repeat steps 2–6 for $k = 0$ or steps 3–6 for $k > 0$.

7. Iterate k+=1. Start again from step 3 with $l = 0$. The MLETPF approximation of $\mathbb{E}_{\eta_{L,t_k}}[X_{L,t_k}]$ is given by (4.8).

4.2. Computational cost of the MLETPF and ETPF. As previously noted, in a multidimensional case, it is computationally expensive to generate the coupling matrices needed to couple the fine and coarse transformed particles. Localization is used to reduce the computational cost of the ensemble transforms down to $O\big(dN\log(N)\big)$ when $r_{loc,c} = 0$ in the standard and multilevel ETPF methods, and this too can also reduce the computational cost of generating the multilevel coupling $T^{F/C}_{i,j}$. When localization is used along with $r_{loc,c} = 0$, one can break the multivariate coupling $T^{F/C}_{i,j}$ down into d separate univariate couplings. As the components are transformed individually, one can simply find a coupling $T^{F/C}_{i,j}(m)$ for each individual component m in the transformed coarse and fine ensembles with the cost function,

$$\sum_{i=1}^{N_l}\sum_{j=1}^{N_l} T^{F/C}_{i,j}(m)\big(\tilde{X}^i_{l,t_k}(m) - \tilde{X}^j_{l-1,t_k}(m)\big)\big)^2, \tag{4.12}$$

and the same constraints as in (4.4). Each of these couplings can again be found using the cheap, $O\big(N_l\log(N_l)\big)$ (for each l) univariate algorithm in [20]. One can then reorder/transform each component of the fine and coarse ensembles separately using the same methodology as in the previous section. This performs a role similar to resampling N_l particles from $F^{-1}_{f,m}(u)$ and $F^{-1}_{c,m}(u)$, where $F^{-1}_{f,m}/F^{-1}_{c,m}$ are the marginal (empirical) inverse cumulative distribution functions of $\tilde{X}_{l,t_k}(m)$ and $\tilde{X}_{l-1,t_k}(m)$, respectively. Here the same uniform variate $u \in [0,1]$ is used for each pair of N_l samples. If localization is carried out along with $r_{loc,c} > 0$, the computational cost of the optimal transport in the standard ETPF and thus the multilevel coupling, minimizing the full multivariate cost function in (4.3), in the MLETPF will rise to $O\big(dN_l^3\log(N_l)\big)$. This is because d different localized, but still multivariate, optimal transportation problems will have to be solved. Therefore, in this scenario, the model costs are likely to be dwarfed by these optimal transportation costs, which are fixed by definition. In this case there is no justification for implementing the multilevel framework, as it only aims to reduce model cost and not the optimal transportation computational expense. However, to consider the case where model cost does dominate that of the multivariate optimal transportation, the full multivariate coupling will be demonstrated in the numerical experiments at the conclusion of this paper.

The paper now looks at the overall computational cost of the ETPF and MLETPF estimators when localized, with $r_{loc,c} = 0$. In this case, as explained above, one can reduce the computational expense of not only the ensemble transform stage, but the multilevel coupling stage as well, in the MLETPF scheme from a potential $O\big(N_l^3\log(N_l)\big)$ to $O\big(dN_l\log(N_l)\big)$ for each multilevel difference estimator. This is enough, with suitable assumptions, to expect that the model computational cost bounds for the standard MLMC method in Theorem 2.1 are of the same order of magnitude as that of the entire MLETPF, including the ensemble transform and coupling

stages, as one can simply "hide" the optimal transportation costs behind the particle propagation costs. This follows from the next proposition.

PROPOSITION 4.1. *If $\Delta t \geq h_0$ is constant, and one can bound the computational cost of all N_y ensemble transform/multilevel coupling stages (with the last term being the cost associated to the sort prior to the algorithm) of the MLETPF by*

$$C_{ET} \leq \sum_{l=0}^{L} \Big(dc_1 N_l N_y + de_1 N_l \log(N_l) N_y \Big), \tag{4.13}$$

where c_1 is a positive constant, then the total computational cost of the MLETPF is bounded by

$$C_{MLETPF} \leq \sum_{l=0}^{L} \Big(c_2 N_l C_{l,d} t_{N_y} + de_1 N_l \log(N_l) N_y \Big), \tag{4.14}$$

where $C_{l,d}$ is the cost of propagation of one particle on level l (dependent on d), e_1 is a positive constant, and c_2 is a positive constant.

Proof. Let the total computational cost of the MLETPF be given by the sum of the model cost and the ensemble transform cost (including the localized likelihood evaluation in (3.21), which scales at $O(N_l)$ due to the sparse diagonal $d \times d$ matrix $\tilde{C}$, with $r_{loc,R}$ assumed constant ($r_{loc,R} = O(1)$) and $\ll d$):

$$C_{MLETPF} = C_{ET} + C_{MODEL}. \tag{4.15}$$

Then bounding C_{ET} as in the claim and using the model cost in (2.8),

$$\begin{aligned} C_{MLETPF} &\leq \sum_{l=0}^{L} \Big(N_l (dc_1 N_y + C_{l,d} c_3 t_{N_y}) + de_1 N_l \log(N_l) N_y \Big) \\ &\leq \sum_{l=0}^{L} \Big(N_l (dc_4 t_{N_y} + C_{l,d} c_3 t_{N_y}) + de_1 N_l \log(N_l) N_y \Big) \\ &\leq \sum_{l=0}^{L} \Big(c_2 N_l C_{l,d} t_{N_y} + de_1 N_l \log(N_l) N_y \Big). \end{aligned} \tag{4.16}$$

Here we assume that $C_{l,d}$ will grow at least linearly with d, $c_4 = \frac{c_1}{\Delta t}$ and that c_3 is a positive constant. ∎

The last term of each of the expressions above comes from the sorting of the univariate particles in the cheap algorithm. Although this computational cost scales at $O\big(N_l \log(N_l)\big)$, the constant e_1 is typically very small with respect to the other constants in the cost expression; furthermore, this scaling is a worst case estimate, but in many cases it would not be reached. Therefore, assuming that this part is less than the remainder of the localized ensemble transform cost, one can bound the overall cost of the MLETPF by the model cost and keep the corresponding reductions outlined in Theorem 2.1. However, even if this sorting cost does dominate the ensemble transform cost, one still expects to recover computational cost reductions relative to the single level estimator for a fixed error. Despite the model cost in this case not necessarily being the highest computational expense in the localized MLETPF, with the cost of the sorting/ensemble transform dominating, it is important to note that for a fixed

error bound this sorting cost/ensemble transform will still typically be less than the model computational cost of the single level ETPF. The reductions in Theorem 2.1 would then be a slight underestimation; however, there would still be evident reductions of computational cost relative to the single level ETPF. This is indeed the case for the numerical examples in the next section, as we show there.

4.3. Numerical examples. Numerical examples of the MLETPF method, applied to classical data assimilation problems, are given in this section. Three problems will be studied: the multivariate, chaotic stochastic Lorenz-63 equations; the univariate but nonlinear double-well OU process; and the high dimensional stochastic Lorenz-96 equations. The algorithm above will be used to generate experimental MLETPF estimators (that are compared against the single level ETPF estimators) for the last two of the three problems above, along with varying levels of predefined accuracy. This predefined level of accuracy, $O(\epsilon)$, will determine L (the finest level/overall numerical discretization bias), the fixed sample sizes for each estimator in the MLETPF method (N_l), and the fixed overall sample size in the corresponding single level ETPF estimator required to achieve the order of magnitude of this error in both estimators. This follows from the standard Monte Carlo approximation error decomposition given by the central limit theorem, as in [9]. One can then compare the computational cost, given as the number of operations, for both the single level and the multilevel estimators, which should be in line with Theorem 2.1 given the same predefined error. The error is not bounded exactly due to variations in the variance at each assimilation step, but a proof of concept from a practitioner's viewpoint can be established.

The error in the estimators will be estimated by the time-averaged root mean square error (RMSE), given by

$$RMSE = \sqrt{\frac{1}{N_y}\sum_{k=1}^{N_y}\left\|\bar{X}_{L,t_k} - \mathbb{E}_{\eta_{t_k}}[X_{t_k}]\right\|^2}, \tag{4.17}$$

where η_{t_k} is the posterior distribution of X_{t_k} given the observations $Y_{t_1,\ldots,t_k}$. An approximation of $\mathbb{E}_{\eta_{t_k}}[X_{t_k}]$ will be used in the RMSE calculations above, by computing a standard ETPF estimator for it, of which the numerical discretization bias and sample size produce an estimator with an error orders less than any ϵ used in the following experiments. In cases where localization is used in the single level and multilevel estimators for the problems above, the estimators are inconsistent with $\mathbb{E}_{\eta_{L,t_k}}[X_{L,t_k}]$, but they are crucially consistent with each other when the same localization settings are used. The ETPF approximation of $\mathbb{E}_{\eta_{t_k}}[X_{t_k}]$ will use the same localization settings to correctly compare the ETPF and MLETPF estimators like for like.

The recurrence relation for N_l, $l \geq 1$, given N_0 (dependent on ϵ), used in both numerical experiments will be set to $N_{l+1} = \lceil N_l M^{-3/2} \rceil$; however, the optimality of this depends on the relative value of β with respect to γ [9]. This is only optimal for $\beta > \gamma$, $(\beta + \gamma) = 3$; this is the case in the numerical examples below. One notes that for an RMSE of $O(\epsilon)$ one requires that $N_l = O(\epsilon^{-2})$, and that $N = O(\epsilon^{-2})$ for the single level ETPF. Also, the discretization bias of the multilevel and single level ETPF estimators from $\mathbb{E}_{\eta_{t_k}}[X_{t_k}]$ should be $O(\epsilon)$. Thus, for a numerical scheme that has a global discretization bias of $O(t_{N_y} h_L^\alpha)$, one requires that

$$L = \left\lceil \frac{\log\left((t_{N_y}^\alpha d)/\epsilon\right)}{\alpha \log(M)} \right\rceil \tag{4.18}$$

for the sum of the multilevel and single level ETPF estimator components' bias to be of $O(\epsilon)$. One could also use the maximum among all components' biases to be a suitable measure here. The sample covariances of the independent ETPF estimators will also be measured by a sum among all components, $\mathrm{Tr}(\mathbb{V}_l)$, the trace of the covariance matrix.

For the single level ETPF, using the above analysis, one can see how requiring $h_L = O(\epsilon)$ and $N = O(\epsilon^{-2})$, the model cost in (2.9) will be $O(\epsilon^{-3})$, dominating the localized (with $r_{loc,c} = 0$) ensemble transform cost $O\big(dN\log(N)\big)$, which is equivalent to $O\big(d\epsilon^{-2}\log(\epsilon^{-2})\big)$; this supports the point made in the previous section. This is also greater than the localized ensemble transform cost of the MLETPF, $O\big(dN_l\log(N_l)\big)$, again equivalent to $O\big(d\epsilon^{-2}\log(\epsilon^{-2})\big)$. Thus the computational cost reductions of the MLETPF, even if the sorting cost of the localized ensemble transform dominates, is apparent in these cases. The computational cost for these numerical examples is defined as the theoretical number of operations needed to compute each approximation, including the ensemble transform stage and the multilevel coupling for the MLETPF. This is computed simply by inserting a step-counter into the numerical implementation (functions, etc.) in the Python code used for these experiments. Finally, $M = 2$ is used for each numerical example.

4.3.1. Stochastic Lorenz-63 equations. The simple 3-component chaotic nonlinear system in $X = (x, y, z)$,

$$\frac{dX}{dt} = \begin{cases} \sigma(y - x) + \phi\frac{dW}{dt}, \\ x(\rho - z) - y + \phi\frac{dW}{dt}, \\ xy - \beta z + \phi\frac{dW}{dt}, \end{cases} \tag{4.19}$$

with $\rho = 28$, $\sigma = 10$, $\beta = 8/3$, $\phi = 0.4$, will be used to demonstrate the effect of the multivariate multilevel coupling, $T^{F/C}_{i,j}$, without localization and thus with cost function in (4.3). The localized coupling will also be used as a comparison. Here the Brownian motion W will be the same for each component to keep the strong nonlinearity in the equations. Computational cost against accuracy comparisons with the standard ETPF method will not be investigated here given the low model cost of this test problem, and thus the dominating effects of the multilevel coupling and/or the ensemble transform stage in both the ETPF and the MLETPF will make the model cost reductions of the multilevel framework unnoticeable. The MLETPF estimator with $N_0 = 500$, and $L = 5$, using the full multivariate cost function in (4.3) to find the multilevel couplings $T^{F/C}_{i,j}$ in the aforementioned algorithm, is used to compute an approximation to $\mathbb{E}_{\eta_{L,t_k}}[X_{L,t_k}]$, with X as above, for $k \in [1, 5120]$, with $h_0 = M^{-7} = \Delta t$, and thus $t_{N_y} = 40$. Here X_L is the solution to the above Lorenz-63 equations using the forward Euler numerical scheme. The reason that we choose the minimum level to be equivalent to $l = 7$ is for stability when using the Euler method. Using different numerical methods, with greater stability at greater time-steps, and thus lower levels, would be able to decrease this minimum level. The observations are given by a measurement error with $R = 2I$, where I is the 3×3 identity matrix and weights are thus based on observations of all components x, y, and z. Figure 1 shows the mean estimates of $Tr(\mathbb{V}_l)$ ($l \in [1, 5]$) over all assimilation steps $k \in [1, N_y]$. The asymptotic decay of the above estimates show importantly that the multilevel coupling, with the multivariate cost function, successfully produces the variance decay, $\mathrm{Tr}(\mathbb{V}_l) = O(h_l^{\beta})$; in this case, $\beta \approx 1$. The figure also shows the value of β (variance decay) for the case where $r_{loc,c} = 0$ localization for the coupling. Localization being

used in a problem such as the strongly nonlinear Lorenz 63 system is dangerous due to the decay in correlations between components, but with the parameters above, it is used simply to compare the rates of variance decay with the nonlocalized case. One recovers $\beta \approx 2$ in the localized case; this shows that refining the optimal transport down to the one dimensional localization case is beneficial for variance reduction but comes with the sacrifice of inconsistency of the reference "single level" estimator.

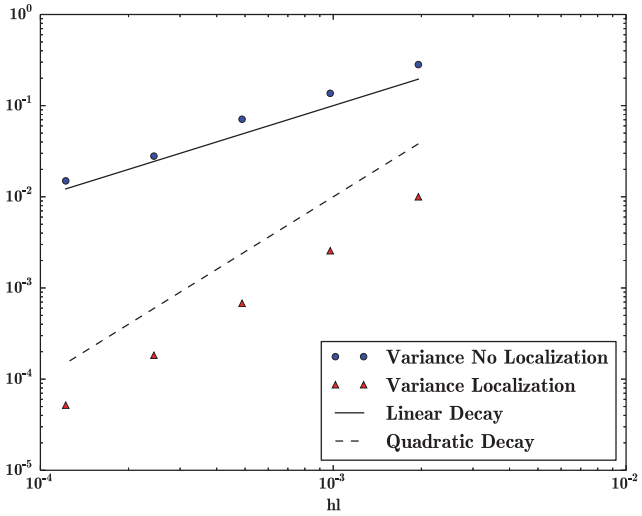

FIG. 1. *Mean, over all assimilation steps, $k \in [1, N_y]$, of the estimates of* $\mathrm{Tr}(\mathbb{V}_l)$*, with $l \in [1, 5]$ for the stochastic Lorenz-63 equations. Both the nonlocalized and the $r_{loc,c} = 0$ localized cases are shown.*

4.3.2. Double-well OU process. This is a univariate test problem that demonstrates the cost effective, consistent, MLETPF estimator of $\mathbb{E}_{\eta_{L,t_k}}[X_{L,t_k}]$, where X_{L,t_k} is a numerical discretized solution to the double-well, nonlinear OU process,

$$dX_{t_k} = -V'(X_{t_k})dt + \xi dW_{t_k}, \tag{4.20}$$

$k \in [1, N_y]$, and W_{t_k} is a standard Brownian motion. Here $V(X_{t_k}) = \frac{1}{4}X_{t_k}^4 - \frac{1}{2}X_{t_k}^2$. This example uses $h_l = 2^{-4-l}$, but it is arbitrarily chosen. The stochastic forcing is set to $\xi = 0.5$. The observations and assimilation times were given by $R = 0.6$, $t_{N_y} = 50$, where $\Delta t = h_0$, and so $N_y = 800$. The numerical discretizations of X_{t_k} are computed by the Euler–Maruyama numerical scheme. The parameters above produce a stable numerical solution for a single realization of the above system when using this scheme for the time frame above. A very accurate simulation ($N_0 = 10000$, $L = 7$) of the MLETPF estimator is run to demonstrate the mean asymptotic decay of V_l and $|\hat{X}_{l,t_k}|$ ($l \in [1, 7]$) over all assimilation steps. These are shown in Figure 2. The values of $\alpha \approx 1$, $\beta \approx 2$ are as expected given the Euler–Maruyama global discretization bias of $O(h_l)$ and the additive noise in the OU process contributing to the variance.

Figure 3 shows the computational cost against the accuracy (RMSE) for the MLETPF and the single level ETPF estimators over varying values of ϵ. Here one sets $N_0 = \epsilon^{-2}$ for the MLETPF and $N = \epsilon^{-2}$ for the single level ETPF estimator. One can clearly see the expected orders of growth for the computational cost of the

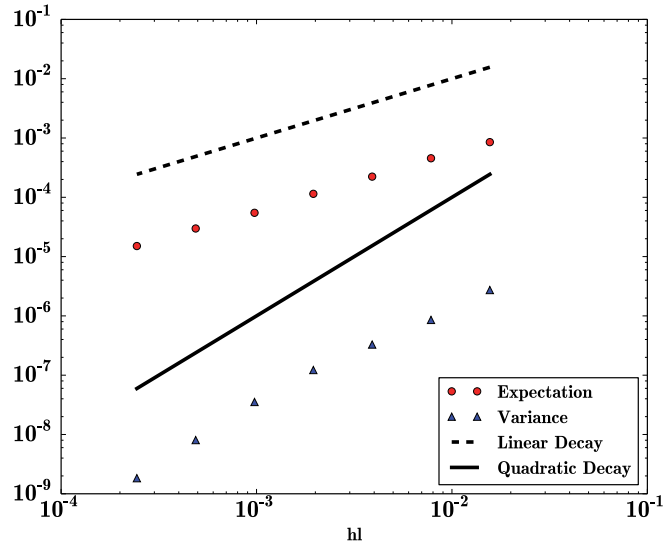

FIG. 2. *Mean, over all assimilation steps, $k \in [1, N_y]$, of the estimates of $\mathbb{V}_l$ (variance) and $|\hat{X}_{l,t_k}|$ (expectation) with $l \in [1, 7]$ for the double-well OU process.*

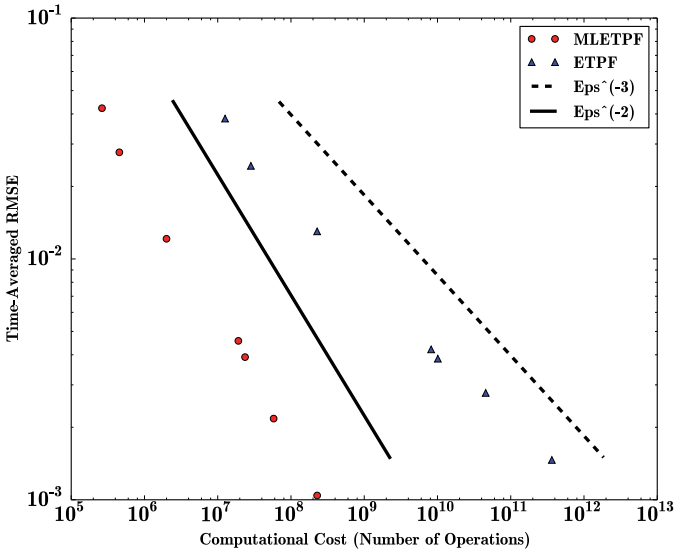

FIG. 3. *Computational cost (number of operations) against the time-averaged RMSE of the ETPF and MLETPF estimators for the double-well OU process. Reference lines show the orders of decay of $RMSE^{-2}$ and $RMSE^{-3}$.*

standard ETPF ($O(\epsilon^{-3})$, as $\gamma = 1$ and $\alpha = 1$) and the MLETPF ($O(\epsilon^{-2})$, given that $\gamma < \beta$) that were shown in Theorem 2.1 for the predefined RMSE of $O(\epsilon)$.

4.3.3. The stochastic Lorenz-96 equations. The final numerical test for the MLETPF method in this paper is the high dimensional ($d = 40$ in this case) stochastic

Lorenz-96 equations, given by

$$\frac{dX_j}{dt} = -\frac{\big(X(j-1)X(j+1) - X(j-2)X(j-1)\big)}{3\Delta X} - X(j) + F + \sigma^2 \frac{dW_j}{dt}, \tag{4.21}$$

with $j \in [1, N_x]$, $N_x = 40$, $N_x \Delta x = 10$, and thus $\Delta X = 0.25$; dW_j/dt are 40 i.i.d. Brownian motions. Here F is a constant forcing ($F = 8$) and $\sigma^2 = 0.4$. Periodic boundary conditions are used, so that $X(-1) = X(N_x)$. The observations were given by a measurement error of $R = 6I$, where I is the 40×40 identity matrix and assimilation times were set to $t_{N_y} = 100$, where $\Delta t = h_0$, meaning that $N_y = 1600$ as $h_l = 2^{-4-l}$ (again simply arbitrary). The Euler–Maruyama method is used to find X_{l,t_k} once again here. In this numerical example, the ETPF and MLETPF estimators use $r_{loc,R} = 1$. The localization setting of $r_{loc,c} = 0$ is used for both the multilevel and the single level ETPF estimators here due to the model cost, $C_{l,d}$, being simply equal to $h_l^{-1} d$ and thus much lower than that of high optimal transportation costs in multiple dimensions.

Once again, a very accurate simulation ($N_0 = 1000$, $L = 10$) of the MLETPF was generated to demonstrate the mean asymptotic decays of $\sum_{i=1}^{40} |\hat{X}_{l,t_k}(i)|$ and $\mathrm{Tr}(\mathbb{V}_l)$ ($l \in [1, 10]$) over all assimilation steps, and these are shown in Figure 4. They follow the same, expected values of $\alpha \approx 1$, $\beta \approx 2$, as with the last example.

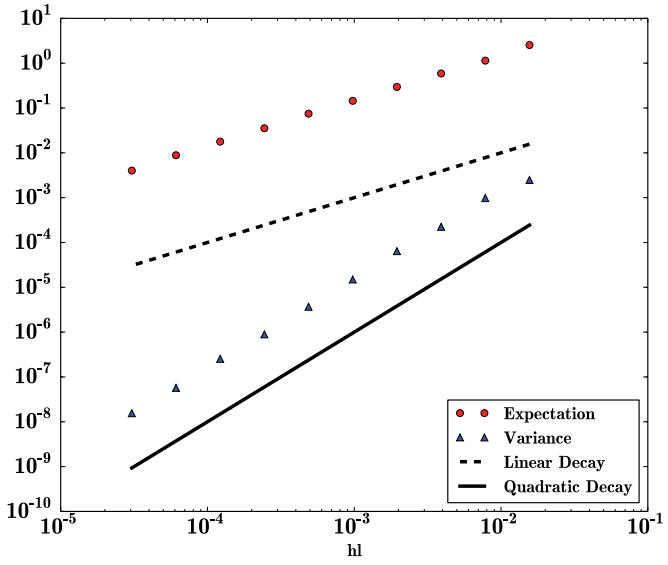

FIG. 4. *Mean, over all assimilation steps, $k \in [1, N_y]$, of the estimates of* $\mathrm{Tr}(\mathbb{V}_l)$ *(variance) and* $\sum_{i=1}^{40} |\hat{X}_{l,t_k}(i)|$ *(expectation) with* $l \in [1, 10]$ *for the stochastic Lorenz-96 equations.*

Next, the stability of the MLETPF is considered. Since N_l is fixed and does not change to bound error over time, we can look at the errors from the reference solution, X'_{t_k}, compared to the observational errors, to study the stability of the MLETPF estimator over time and check that the errors do not increase. One expects that, for a successful particle filter, the errors from the estimator should be less than the observational errors and remain stable. Figure 5 shows this expected behavior, where the cumulative time-averaged RMSE values from X'_{t_k}, of both the observations

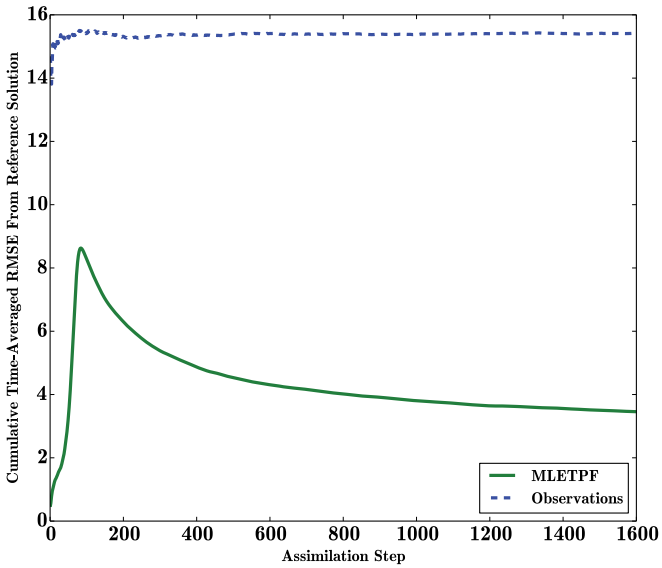

FIG. 5. *The cumulative time-averaged RMSE of the observations and MLETPF estimator away from the reference solution, X'_{t_k}, for the stochastic Lorenz-96 equations.*

and the MLETPF estimator (using arbitrary values of $N_0 = 1000$ and $L = 10$) are shown. For $k \in [1, N_y]$, these are defined to be

$$\sqrt{\frac{1}{k}\sum_{i=1}^{k}\left\|Y_{t_i} - X'_{t_i}\right\|^2} \tag{4.22}$$

for the observations and

$$\sqrt{\frac{1}{k}\sum_{i=1}^{k}\left\|\bar{X}_{L,t_i} - X'_{t_i}\right\|^2} \tag{4.23}$$

for the MLETPF estimator. To demonstrate the stability of the variance of the particle filter, cumulative time-averaged RMSE values for the second moments of $\bar{X}_{L,t_k}$ and Y_{t_k} are shown in Figure 6. Finally, to compare the MLETPF with its standard counterpart for this set of equations, Figure 7 shows the computational cost against the accuracy (RMSE) for the standard ETPF and the MLETPF estimators over varying values of ϵ. Once again, one sets $N_0 = \epsilon^{-2}$ for the MLETPF and $N = \epsilon^{-2}$ for the single level ETPF estimator. This follows the successful cost reductions achieved in the last example, defined in Theorem 2.1.

5. Summary and outlook. This paper has demonstrated a proof of concept for the application of MLMC to nonlinear filtering. The ETPF, coupled with localization, allows one to simply and cheaply carry out a multilevel coupling between each fine and coarse ensemble in each independent Monte Carlo estimator in the MLMC framework. A recent study has also proposed a framework to apply MLMC to nonlinear filtering with a modified random resampling step in the standard particle filtering methodology to couple particles from coarse and fine levels [12]. In contrast, the coupling

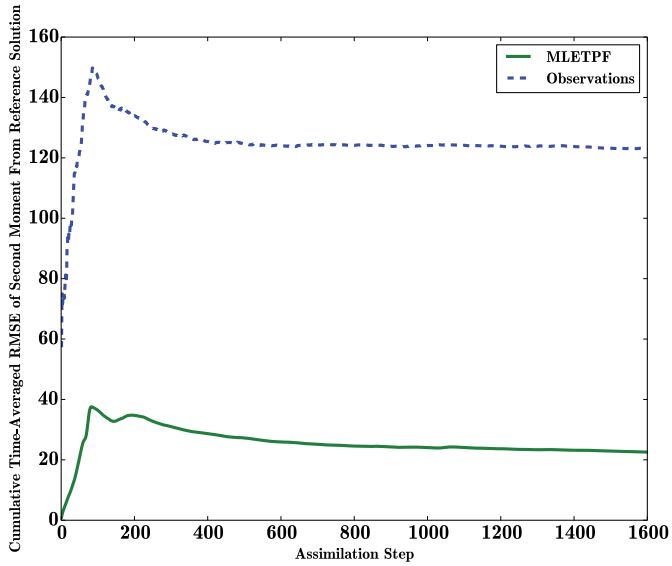

FIG. 6. *The cumulative time-averaged RMSE of the second moment of the observations and MLETPF estimator away from the second moment of the reference solution,* $(X'_{t_k})^2$*, for the stochastic Lorenz-96 equations.*

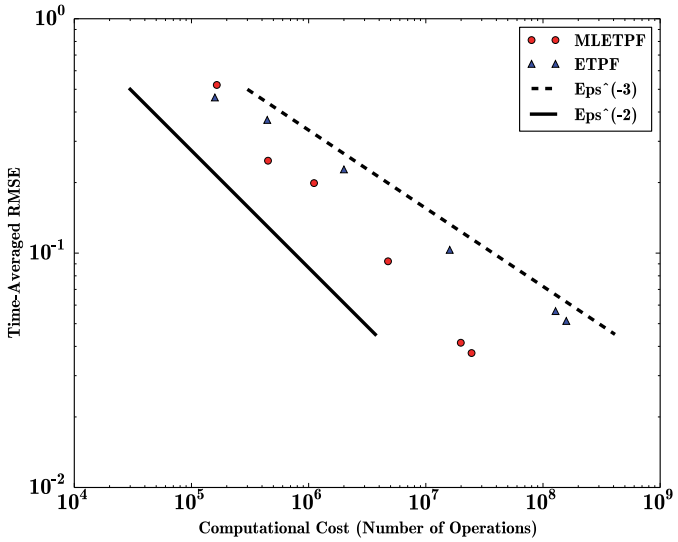

FIG. 7. *Computational cost (number of operations) against the time-averaged RMSE of the ETPF and MLETPF estimators for the stochastic Lorenz-96 equations. Reference lines show the orders of decay of* $RMSE^{-2}$ *and* $RMSE^{-3}$*.*

in the present paper is designed to minimize the Wasserstein distance between the distributions of these transformed ensembles (in the standard ETPF methodology), originally suggested in [8]. It has been shown through numerical experiments that one can restore positive correlation between fine and coarse ensembles which might have

been lost if they had been transformed independently of one another. This in turn satisfies the necessary constraints on the sample variance of each independent multilevel estimator, allowing the proposed MLETPF method to reduce the computational cost of the propagation of particles in the localized ETPF method.

In general, localization with $r_{loc,c} = 0$ makes the computational cost of this coupling and the ensemble transform in the MLETPF cheap enough that the multilevel framework can return overall computational cost reductions from the standard ETPF methods; this is the aim of the paper. It must be noted that although this paper has only touched on the case where the very crude $r_{loc,c} = 0$ localization is considered, due to small model cost test problems, this method could also be applied to high dimensional systems where the model cost dominates that of the multivariate optimal transportation. One can do this without any crude constraints on $r_{loc,c}$ using the full multivariate coupling methodology presented in this paper and demonstrated numerically. However, whether the variance decay of $\mathbb{V}_l$ from such a multivariate coupling would hold, producing equally strong results, in the limit of $d \gg N$ is unknown, and thus the issue of how one would adjust this coupling to be used alongside other values of $r_{loc,c}$ to reduce the dimensionality of these multivariate couplings remains to be explored.

Iterative and approximate schemes for solving discrete optimal transportation problems have been an area of rapid research in the last few years [17], and this offers the chance to improve the multilevel coupling in the proposed method by reducing computational cost. This could be done by trading off between the optimality and computational cost of the coupling for each l, e.g., more expensive/optimal couplings for greater l with lower sample sizes N_l.

The form of the coupling used in this paper is simple to implement and has the potential to be used in plenty of applications, in and outside of data assimilation, whenever one wishes to establish consistent correlation between two distributions for variance reduction. Considering an extension for the multidimensional example presented in this paper, one could also apply a spatial multilevel framework, setting the spatial resolution (ΔX_l) to be dependent on the level of discretization, as done in [6, 2] to gain even more significant cost reductions.

REFERENCES

[1] J. L. Anderson, *Localization and sampling error correction in ensemble Kalman filter data assimilation*, Mon. Weather Rev., 140 (2012), pp. 2359–2371, http://dx.doi.org/10.1175/MWR-D-11-00013.1.

[2] A. Barth and A. Lang, *Multilevel Monte Carlo method with applications to stochastic partial differential equations*, Int. J. Comput. Math., 89 (2012), pp. 2479–2498, http://dx.doi.org/10.1080/00207160.2012.701735.

[3] A. Beskos, A. Jasra, K. Law, R. Tempone, and Y. Zhou, *Multilevel Sequential Monte Carlo Samplers*, preprint, arXiv:1503.07259, 2015.

[4] O. Cappé, S. J. Godsill, and E. Moulines, *An overview of existing methods and recent advances in sequential Monte Carlo*, Proc. IEEE, 95 (2007), pp. 899–924, http://dx.doi.org/10.1109/JPROC.2007.893250.

[5] Y. Cheng and S. Reich, *A McKean Optimal Transportation Perspective on Feynman-Kac Formulae with Application to Data Assimilation*, preprint, arXiv:1311.6300, 2013.

[6] K. A. Cliffe, M. B. Giles, R. Scheichl, and A. L. Teckentrup, *Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients*, Comput. Vis. Sci., 14 (2011), pp. 3–15, http://dx.doi.org/10.1007/s00791-011-0160-x.

[7] A. Doucet and A. M. Johansen, *A tutorial on particle filtering and smoothing: Fifteen years later*, in The Oxford Handbook of Nonlinear Filtering, Oxford University Press, Oxford, UK, 2011, pp. 656–704, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.

1.157.772.

[8] M. Giles, *Multilevel Monte Carlo methods*, in Monte Carlo and Quasi-Monte Carlo Methods 2012, Springer Proc. Math. Stat. 65, Springer, Heidelberg, 2013, pp. 83–103, http://dx.doi.org/10.1007/978-3-642-41095-6_4.

[9] M. B. Giles, *Multilevel Monte Carlo path simulation*, Oper. Res., 56 (2008), pp. 607–617, http://dx.doi.org/10.1287/opre.1070.0496.

[10] M. B. Giles and B. J. Waterhouse, *Multilevel quasi-Monte Carlo path simulation*, in Advanced Financial Modelling, Radon Ser. Comput. Appl. Math. 8, Walter de Gruyter, Berlin, 2009, pp. 165–181.

[11] H. Hoel, K. J. H. Law, and R. Tempone, *Multilevel Ensemble Kalman Filtering*, preprint, arXiv:1502.06069, 2015.

[12] A. Jasra, K. Kamatani, K. J. H. Law, and Y. Zhou, *Multilevel Particle Filter*, preprint, arXiv:1510.04977, 2015.

[13] C. Ketelsen, R. Scheichl, and A. L. Teckentrup, *A Hierarchical Multilevel Markov Chain Monte Carlo Algorithm with Applications to Uncertainty Quantification in Subsurface Flow*, preprint, Center for Applied Scientific Computing, Livermore, CA, 2013.

[14] F. Le Gland, V. Monbet, and V. Tran, *Large sample asymptotics for the ensemble Kalman filter*, in The Oxford Handbook of Nonlinear Filtering, Oxford University Press, Oxford, UK, 2011, pp. 598–631.

[15] J. Munkres, *Algorithms for the assignment and transportation problems*, J. Soc. Indust. Appl. Math., 5 (1957), pp. 32–38, http://dx.doi.org/10.1137/0105003.

[16] O. Pele and M. Werman, *Fast and robust earth mover's distances*, in Proceedings of the 12th IEEE International Conference on Computer Vision, 2009, pp. 460–467, http://dx.doi.org/10.1109/ICCV.2009.5459199.

[17] R. Ravi, *Iterative methods in combinatorial optimization*, in Proceedings of the 29th Symposium on Theoretical Aspects of Computer Science, Vol. 14, 2012, p. 24, http://dx.doi.org/10.4230/LIPIcs.STACS.2012.24.

[18] P. Rebeschini and R. Van Handel, *Can local particle filters beat the curse of dimensionality?*, Ann. Appl. Probab., 25 (2015), pp. 2809–2866, http://dx.doi.org/10.1214/14-AAP1061.

[19] S. Reich, *A nonparametric ensemble transform method for Bayesian inference*, SIAM J. Sci. Comput., 35 (2013), pp. A2013–A2024, http://dx.doi.org/10.1137/130907367.

[20] S. Reich and C. J. Cotter, *Probabilistic Forecasting and Bayesian Data Assimilation*, Cambridge University Press, New York, 2015.

[21] R. Schefzik, T. L. Thorarinsdottir, and T. Gneiting, *Uncertainty quantification in complex simulation models using ensemble copula coupling*, Statist. Sci., 28 (2013), pp. 616–640, http://dx.doi.org/10.1214/13-STS443.

[22] C. Villani, *Optimal Transport: Old and New*, Grundlehren Math. Wiss. 338, Springer, Berlin, 2009.