Second-order accurate ensemble transform particle filters

Article

Accepted Version

SECOND-ORDER ACCURATE ENSEMBLE TRANSFORM PARTICLE FILTERS

WALTER ACEVEDO *, JANA DE WILJES†, AND SEBASTIAN REICH‡

Abstract. Particle filters (also called sequential Monte Carlo methods) are widely used for state and parameter estimation problems in the context of nonlinear evolution equations. The recently proposed ensemble transform particle filter (ETPF) (S. Reich, *A non-parametric ensemble transform method for Bayesian inference*, SIAM J. Sci. Comput., 35, (2013), pp. A2013–A2014) replaces the resampling step of a standard particle filter by a linear transformation which allows for a hybridization of particle filters with ensemble Kalman filters and renders the resulting hybrid filters applicable to spatially extended systems. However, the linear transformation step is computationally expensive and leads to an underestimation of the ensemble spread for small and moderate ensemble sizes. Here we address both of these shortcomings by developing second-order accurate extensions of the ETPF. These extensions allow one in particular to replace the exact solution of a linear transport problem by its Sinkhorn approximation. It is also demonstrated that the nonlinear ensemble transform filter (NETF) arises as a special case of our general framework. We illustrate the performance of the second-order accurate filters for the chaotic Lorenz-63 and Lorenz-96 models and a dynamic scene-viewing model. The numerical results for the Lorenz-63 and Lorenz-96 models demonstrate that significant accuracy improvements can be achieved in comparison to a standard ensemble Kalman filter and the ETPF for small to moderate ensemble sizes. The numerical results for the scene-viewing model reveal, on the other hand, that second-order corrections can lead to statistically inconsistent samples from the posterior parameter distribution.

Keywords. Bayesian inference, data assimilation, particle filter, ensemble Kalman filter, Sinkhorn approximation

AMS(MOS) subject classifications. 65C05, 62M20, 93E11, 62F15, 86A22

1. Introduction. Data assimilation (DA) denotes the broad topic of combining evolution models with partial observations of the underlying dynamical process [10, 15, 22]. DA algorithms come in the form of variational and/or ensemble-based methods [15]. In this paper, we focus on ensemble-based DA methods and their robust and efficient implementation. The ensemble Kalman filter (EnKF) [10] is by far the most popular ensemble-based DA method and has found widespread application in the geosciences. However, EnKFs lead to inconsistent approximations for partially observed nonlinear processes. On the contrary, particle filters (PF) (also called sequential Monte Carlo methods) [8] lead to consistent approximations but typically require ensemble sizes much larger than those required for EnKFs in order to track the underlying reference process [2].

In order to overcome these shortcomings, we are currently witnessing a strong trend towards hybrid filters which combine EnKFs with PFs and which are applicable to strongly nonlinear systems under small or moderate ensemble sizes. We mention here the Gaussian mixture filters (such as, for example, [24]), the rank histogram filter [1, 19], moment matching ensemble filters [28, 16, 25], the ensemble Kalman particle filter [11], and the hybrid ensemble transform particle filter [6].

In this paper, we focus on improved implementations of the ensemble transform particle filter (ETPF) [21, 22] and its hybridrization with the EnKF [6]. The ETPF requires the solution of a linear transport problem in each assimilation step, which renders the methods substantially more expensive than an EnKF. Computationally attractive alternatives, such as the Sinkhorn approximation [7], lead to unstable implementations since the ensemble becomes underdispersive. We address this problem by introducing a variant of the ETPF, which is second-order accurate independent of the actual solution procedure for the underlying optimal transport problem. An ensemble filter is called second-order accurate if the posterior mean and covariance matrix of the ensemble are in agreement with their importance sampling estimates from a Bayesian inference step. Second-order accurate particle filters have first been proposed in [28] and since then several variants of it have been developed [16, 25]. Here we instead consider second-order corrections to the ETPF. Such corrections require the solution of a continuous-time algebraic Riccati equation [27, 14]. The correction term vanishes as the ensemble size approaches infinity in agreement with the consistency of the ETPF [21].

The paper is organized as follows. The general framework of ensemble transform filters is summarized in Section 2. Section 3 summarizes the ETPF and introduces the second-order correction step. Numerical solution procedures for the associated continuous-time algebraic Riccati equation are discussed in Sectionsec:Riccati. The Sinkhorn approximation to the optimal transport problem of the ETPF is introduced in Section 5 and the overall second-order accurate implementation of the ETPF is summarized in Section 6. Numerical results are provided in Section 7, where the behavior of the new method is demonstrated for the highly nonlinear and

*Universität Potsdam, Institut für Mathematik, Karl-Liebknecht-Str. 24/25, D-14476 Potsdam, Germany

†Universität Potsdam, Institut für Mathematik, Karl-Liebknecht-Str. 24/25, D-14476 Potsdam, Germany

‡Universität Potsdam, Institut für Mathematik, Karl-Liebknecht-Str. 24/25, D-14476 Potsdam, Germany (sreich@math.uni-potsdam.de) and University of Reading, Department of Mathematics and Statistics, Whiteknights, PO Box 220, Reading RG6 6AX, UK.

chaotic Lorenz-63 [17] and Lorenz-96 [18] models. Here we repeat the experiments from [6] with the ETPF being replaced by a second-order accurate variant based on the Sinkhorn approximation to the underlying optimal transport problem. We finally also demonstrate the behavior of the new filters for parameter estimation of the scene-viewing model *SceneWalk* [9].

2. Ensemble-based forecasting-data assimilation systems. Let us assume that observations $\mathbf{y}^{\text{obs}}(t_k) \in \mathbb{R}^{N_y}$ become available at time instances t_k, $k = 1, \ldots, K$, and are related to the state variables $\mathbf{z} \in \mathbb{R}^{N_z}$ of an evolution model

$$\mathbf{z}(t_k) = \mathcal{M}(\mathbf{z}(t_{k-1})) \tag{2.1}$$

via the likelihood function

$$\pi(\mathbf{y}|\mathbf{z}) = \frac{1}{(2\pi)^{N_y/2}|\mathbf{R}|^{1/2}} \exp\left(-\frac{1}{2}(h(\mathbf{z}) - \mathbf{y})^{\mathrm{T}}\mathbf{R}^{-1}(h(\mathbf{z}) - \mathbf{y})\right), \tag{2.2}$$

where $\mathbf{R} \in \mathbb{R}^{N_y \times N_y}$ denotes the measurement error covariance matrix.

An ensemble-based forecasting-data assimilation (FOR-DA) systems will produce two sets of ensembles of size M at any t_k. First we have the forecast ensemble $\{\mathbf{z}_i^{\text{f}}\}_{i=1}^M$ which approximates the conditional distribution $\pi(\mathbf{z}, t_k|\mathbf{y}_{1:k-1}^{\text{obs}})$ and, second, we have the analysis ensemble $\{\mathbf{z}_i^{\text{a}}\}_{i=1}^M$, which approximates the conditional distribution $\pi(\mathbf{z}, t_k|\mathbf{y}_{1:k}^{\text{obs}})$. Here

$$\mathbf{y}_{1:l}^{\text{obs}} = (\mathbf{y}^{\text{obs}}(t_1), \mathbf{y}^{\text{obs}}(t_2), \ldots, \mathbf{y}^{\text{obs}}(t_l)) \in \mathbb{R}^{N_y \times l} \tag{2.3}$$

denotes the complete set of observations from $t = t_1$ to $t = t_l$. Also note that

$$\mathbf{z}_i^{\text{f}}(t_k) = \mathcal{M}(\mathbf{z}_i^{\text{a}}(t_{k-1})) \tag{2.4}$$

and that FOR-DA systems primarily differ in the employed data assimilation algorithms.

The data assimilation algorithms considered in this paper are all of the form of a linear ensemble transform filter (LETF) [22]:

$$\mathbf{z}_j^{\text{a}}(t_k) = \sum_{i=1}^{M} \mathbf{z}_i^{\text{f}}(t_k)\, d_{ij}(t_k) \tag{2.5}$$

where the entries $d_{ij}(t_k)$ of the $M \times M$ transformation matrix $\mathbf{D}(t_k) = \{d_{ij}(t_k)\}$ are subject to the constraint

$$\sum_{i=1}^{M} d_{ij}(t_k) = 1 \tag{2.6}$$

for all $j = 1, \ldots, M$. In other words, provided that $M \leq N_z$, the members $\mathbf{z}_j^{\text{a}}(t_k)$ of the analysis ensemble lie in the $(M-1)$-dimensional hyperplane spanned by the forecast ensemble $\mathbf{z}_i^{\text{f}}(t_k)$ with $i \in \{1, \ldots, M\}$. Note that the entries of $\mathbf{D}$ can be negative. See [22]. One well known exemplary class of DA algorithms that has the LETF structure is the family of EnKFs [10, 22]. It has long been acknowledged that EnKFs are very robust yet the underlying Gaussianity and linearity assumptions limit their applicability to more general systems. To address the shortcomings of traditional techniques such as the EnKFs, other algorithms that are applicable for nonlinear model scenarios and are computational feasible when employed to high-dimensional systems have been proposed. For example, the nonlinear ensemble transform filter (NETF) of [28, 25] provides an example of a particle filter in the form of an LETF based upon the normalized importance weights

$$w_i(t_k) := \frac{\widehat{w}_i(t_k)}{\sum_{j=1}^{M} \widehat{w}_j(t_k)} \tag{2.7}$$

with $\widehat{w}_i(t_k) = \pi(\mathbf{y}(t_k)|\mathbf{z}_i^{\text{f}}(t_k))$, which reproduces the first and second-order moments of the posterior distribution. The main focus of this paper is, however, on the ETPF which can also be formulated in the form of (2.5) [10, 22] with $\mathbf{D}(t_k) = \{d_{ij}(t_k)\}$ being defined via minimization of the cost functional

$$J(\mathbf{D}(t_k)) = \sum_{i,j=1}^{M} d_{ij}(t_k)\|\mathbf{z}_i^{\text{f}}(t_k) - \mathbf{z}_j^{\text{f}}(t_k)\|^2 \tag{2.8}$$

subject to $d_{ij}(t_k) \geq 0$, (2.6) and

$$\frac{1}{M}\sum_{j=1}^{M} d_{ij}(t_k) = w_i(t_k) \tag{2.9}$$

[21]. The key idea of the ETPF is to approximate a transfer map between the random variable, $Z^{\mathrm{f}}(t_k)$, distributed according to $\pi_{Z^{\mathrm{f}}}(\mathbf{z}, t_k)$ and the random variable, $Z^{\mathrm{a}}(t_k)$, associated with $\pi_{Z^{\mathrm{a}}}(\mathbf{z}, t_k)$. This map induces a coupling of the respective densities that is optimal in the sense that it minimizes the expected distance between the two random variables, i.e.,

$$\mu_Z^* = \arg \inf_{\mu \in \Pi(\pi_{Z^{\mathrm{f}}}, \pi_{Z^{\mathrm{a}}})} \sqrt{\mathbb{E}||Z^{\mathrm{f}}(t_k) - Z^{\mathrm{a}}(t_k)||^2}. \tag{2.10}$$

Intuitively it is clear that the correlation between the forecast and the analysis random variable is increased via optimization of (2.10) and thus creates a strong relation between the prior and the posterior. Since we only rely on importance weights, our filter is also applicable to non-Gaussian likelihood functions. The ETPF can also be applied to spatially extended systems using the idea of localization [5] and has been combined with EnKFs in an hybridization approach [6]. While the ETPF convergence to the true posterior distribution in the limit of $M \to \infty$ [21], this is not the case for the EnKF or the NETF, in general. However, the ETPF is computationally expensive and underestimates the ensemble spread (covariance matrix) for finite ensemble sizes (see example 8.11 in [22]). Both of these shortcomings will be addressed by the LETFs proposed in Sections 3 and 5.

3. Second-order accurate LETFs. We now derive second-order accurate LETFs. Here second-order accuracy refers to reproducing the first and second-order moments exactly according to the importance sampling approach.

Definition 3.1. *An LETF (2.5) is called second-order accurate if the analysis mean satisfies*

$$\overline{\mathbf{z}}^{\mathrm{a}}(t_k) = \frac{1}{M}\sum_{i=1}^{M} \mathbf{z}_i^{\mathrm{a}}(t_k) = \sum_{i=1}^{M} w_i(t_k)\mathbf{z}_i^{\mathrm{f}}(t_k) \tag{3.1}$$

and the analysis covariance matrix

$$\widehat{\mathbf{P}}^{\mathrm{a}}(t_k) = \frac{1}{M}\sum_{i=1}^{M} (\mathbf{z}_i^{\mathrm{a}}(t_k) - \overline{\mathbf{z}}^{\mathrm{a}}(t_k))(\mathbf{z}_i^{\mathrm{a}}(t_k) - \overline{\mathbf{z}}^{\mathrm{a}}(t_k))^{\mathrm{T}} \tag{3.2}$$

is equal to the covariance matrix defined by the importance weights, i.e.

$$\mathbf{P}^{\mathrm{a}}(t_k) = \sum_{i=1}^{M} w_i(t_k)(\mathbf{z}_i^{\mathrm{f}}(t_k) - \overline{\mathbf{z}}^{\mathrm{a}}(t_k))(\mathbf{z}_i^{\mathrm{f}}(t_k) - \overline{\mathbf{z}}^{\mathrm{a}}(t_k))^{\mathrm{T}}. \tag{3.3}$$

Remark 3.1. *The covariance matrix (3.3) derived via importance sampling leads to the denominator M in case of equal weights $w_i = 1/M$. In line with this, the biased version of the empirical covariance (3.2) is used in this paper. Another option is to introduce the factor $\frac{M}{M-1}$ in (3.3) to obtain the unbiased variant (as is used, for example, in the NETF see [25]).*

Since only the DA step of a FOR-DA system is considered in this and the following sections, we drop the explicit time-dependence for notational convenience from now on. We introduce the $N_z \times M$ matrix of the forecast ensemble

$$\mathbf{Z}^{\mathrm{f}} = (\mathbf{z}_1^{\mathrm{f}}, \mathbf{z}_2^{\mathrm{f}}, \ldots, \mathbf{z}_M^{\mathrm{f}}) \in \mathbb{R}^{N_z \times M} \tag{3.4}$$

and an analog expression

$$\mathbf{Z}^{\mathrm{a}} = (\mathbf{z}_1^{\mathrm{a}}, \mathbf{z}_2^{\mathrm{a}}, \ldots, \mathbf{z}_M^{\mathrm{a}}) \in \mathbb{R}^{N_z \times M} \tag{3.5}$$

for the analysis ensemble. Then an LETF (2.5) can be represented in the form

$$\mathbf{Z}^{\mathrm{a}} = \mathbf{Z}^{\mathrm{f}}\mathbf{D}. \tag{3.6}$$

We also introduce the vector $\mathbf{1} = (1, 1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^{M\times 1}$, the vector

$$\mathbf{w} = (w_1, \ldots, w_M)^{\mathrm{T}} \in \mathbb{R}^{M\times 1} \tag{3.7}$$

of normalized importance weights (2.7), and the diagonal $M \times M$ matrix $\mathbf{W} = \mathrm{diag}\,(\mathbf{w})$. Since the analysis mean is provided by (3.1), an LETF is first-order accurate if

$$\frac{1}{M}\mathbf{Z}^{\mathrm{a}}\mathbf{1} = \mathbf{Z}^{\mathrm{f}}\mathbf{w}. \tag{3.8}$$

Equation (3.8) holds if $\mathbf{D}$ satisfies (2.9), i.e.

$$\frac{1}{M}\mathbf{D}\mathbf{1} = \mathbf{w}. \tag{3.9}$$

Recall that the transformation matrix is also subject to (2.6), which is equivalent to $\mathbf{D}^{\mathrm{T}}\mathbf{1} = \mathbf{1}$ [22]. In the following, the focus is on first-order accurate LETF characterized by transformation matrices, $\mathbf{D}$, in the class

$$\mathcal{D}_1 = \{\mathbf{D} \in \mathbb{R}^{M\times M} |\, \mathbf{D}^{\mathrm{T}}\mathbf{1} = \mathbf{1},\ \mathbf{D}\mathbf{1} = M\mathbf{w}\,\}. \tag{3.10}$$

These conditions are, for example, satisfied by the transformation matrix

$$\mathbf{D}_0 = \mathbf{w}\mathbf{1}^{\mathrm{T}}, \tag{3.11}$$

which leads to the analysis ensemble

$$\mathbf{Z}^{\mathrm{a}} = \overline{\mathbf{z}}^{\mathrm{a}}\mathbf{1}^{\mathrm{T}}. \tag{3.12}$$

REMARK 3.2. *An EnKF also leads to transformations of the form (3.6) with the associated* $\mathbf{D}_{\mathrm{EnKF}}$ *satisfying* $\mathbf{D}_{\mathrm{EnKF}}^{\mathrm{T}}\mathbf{1} = \mathbf{1}$ *but in general not (3.9) [22]. Hence* $\mathbf{D}_{\mathrm{EnKF}} \notin \mathcal{D}_1$*, in general. The simple modification*

$$\widehat{\mathbf{D}}_{\mathrm{EnKF}} = \mathbf{D}_{\mathrm{EnKF}}\left(\mathbf{I} - \frac{1}{M}\mathbf{1}\mathbf{1}^{\mathrm{T}}\right) + \mathbf{D}_0 \tag{3.13}$$

leads to $\widehat{\mathbf{D}}_{\mathrm{EnKF}} \in \mathcal{D}_1$*.*

Note that the analysis covariance matrix (3.2) can be equivalently written in the form

$$\widehat{\mathbf{P}}^{\mathrm{a}} = \frac{1}{M}\mathbf{Z}^{\mathrm{f}}(\mathbf{D} - \mathbf{w}\mathbf{1}^{\mathrm{T}})(\mathbf{D} - \mathbf{w}\mathbf{1}^{\mathrm{T}})^{\mathrm{T}}(\mathbf{Z}^{\mathrm{f}})^{\mathrm{T}} \tag{3.14}$$

for any $\mathbf{D} \in \mathcal{D}_1$. In order to achieve second-order accuracy, (3.14) has to be equal to the importance sampling estimate of the posterior covariance matrix (3.3) which can now be expressed in the following form

$$\mathbf{P}^{\mathrm{a}} = \mathbf{Z}^{\mathrm{f}}(\mathbf{W} - \mathbf{w}\mathbf{w}^{\mathrm{T}})(\mathbf{Z}^{\mathrm{f}})^{\mathrm{T}}. \tag{3.15}$$

The class of second-order accurate LETFs, considered in this paper, is now characterized by the set

$$\mathcal{D}_2 = \{\mathbf{D} \in \mathcal{D}_1 |\, (\mathbf{D} - \mathbf{w}\mathbf{1}^{\mathrm{T}})(\mathbf{D} - \mathbf{w}\mathbf{1}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{W} - \mathbf{w}\mathbf{w}^{\mathrm{T}}\,\}. \tag{3.16}$$

REMARK 3.3. *There is an important subclass* $\mathcal{D}_1^+ \subset \mathcal{D}_1$ *that satisfies the additional constraint* $d_{ij} \geq 0$*, i.e.*

$$\mathcal{D}_1^+ = \{\mathbf{D} \in \mathcal{D}_1 |\, d_{ij} \geq 0 \text{ for all } i, j = 1, \ldots, M\}. \tag{3.17}$$

Then $\mathbf{D} \in \mathcal{D}_1^+$ *are left stochastic matrices and thus can be interpreted as resampling schemes that produce realizations* $\mathbf{z}_j^a$ *with respect to the transition probabilities in column* j *in* $\mathbf{D}$ *for* $j \in 1, \ldots, M$*. However, if such a stochastic matrix is used deterministically to produce an analysis ensemble, such as in the ETPF, then the particles* $\mathbf{z}_j^a$ *are associated with the expected value of the random variable induced by each column* j *of* $\mathbf{D}$*. Consider, for example, the simple transformation matrix* $\mathbf{D}_0 \in \mathcal{D}_1^+$ *given in (3.11). In this case,* $\mathbf{z}_j^{\mathrm{a}} = \overline{\mathbf{z}}^{\mathrm{a}}$ *for all* $j \in \{1, \ldots, M\}$ *and the implied analysis covariance matrix (3.14) becomes identical to zero, which is clearly undesirable, and* $\mathbf{D}_0 \notin \mathcal{D}_2$*. The ETPF is designed such that this effect is minimized and vanishes asymptotically as* $M \to \infty$ *[21, 22]. More broadly speaking, one has* $\mathcal{D}_1^+ \cap \mathcal{D}_2 = \emptyset$ *generically.*

We now propose a general methodology of how to turn a transformation matrix $\mathbf{D} \in \mathcal{D}_1$ into a transformation matrix $\widehat{\mathbf{D}} \in \mathcal{D}_2$. We start from the *ansatz*

$$\widehat{\mathbf{D}} = \mathbf{D} + \mathbf{\Delta} \tag{3.18}$$

with $\mathbf{D} \in \mathcal{D}_1$, $\mathbf{\Delta} \in \mathbb{R}^{M\times M}$ such that $\mathbf{\Delta 1} = \mathbf{0}$, $\mathbf{\Delta}^{\mathrm{T}}\mathbf{1} = \mathbf{0}$, and $\mathbf{P}^{\mathrm{a}} = \widehat{\mathbf{P}}^{\mathrm{a}}$ with

$$\widehat{\mathbf{P}}^{\mathrm{a}} = \frac{1}{M}\mathbf{Z}^{\mathrm{f}}(\widehat{\mathbf{D}} - \mathbf{w}\mathbf{1}^{\mathrm{T}})(\widehat{\mathbf{D}} - \mathbf{w}\mathbf{1}^{\mathrm{T}})^{\mathrm{T}}(\mathbf{Z}^{\mathrm{f}})^{\mathrm{T}}\,. \tag{3.19}$$

The condition

$$\mathbf{0} = \mathbf{P}^{\mathrm{a}} - \widehat{\mathbf{P}}^{\mathrm{a}} = \mathbf{Z}^{\mathrm{f}}\left\{(\mathbf{W} - \mathbf{w}\mathbf{w}^{\mathrm{T}}) - \frac{1}{M}(\widehat{\mathbf{D}} - \mathbf{w}\mathbf{1}^{\mathrm{T}})(\widehat{\mathbf{D}} - \mathbf{w}\mathbf{1}^{\mathrm{T}})^{\mathrm{T}}\right\}(\mathbf{Z}^{\mathrm{f}})^{\mathrm{T}}, \tag{3.20}$$

together with (3.18) lead to the following quadratic equation in the correction $\mathbf{\Delta}$:

$$M(\mathbf{W} - \mathbf{w}\mathbf{w}^{\mathrm{T}}) - (\mathbf{D} - \mathbf{w}\mathbf{1}^{\mathrm{T}})(\mathbf{D} - \mathbf{w}\mathbf{1}^{\mathrm{T}})^{\mathrm{T}} = (\mathbf{D} - \mathbf{w}\mathbf{1}^{\mathrm{T}})\mathbf{\Delta}^{\mathrm{T}} + \mathbf{\Delta}(\mathbf{D} - \mathbf{w}\mathbf{1}^{\mathrm{T}})^{\mathrm{T}} + \mathbf{\Delta}\mathbf{\Delta}^{\mathrm{T}}. \tag{3.21}$$

If we also choose $\mathbf{\Delta}$ to be symmetric, then the special case (3.11) leads to

$$M(\mathbf{W} - \mathbf{w}\mathbf{w}^{\mathrm{T}}) = \mathbf{\Delta}\mathbf{\Delta} \tag{3.22}$$

and a solution of (3.21) is simply given by the symmetric square root

$$\mathbf{\Delta} = \sqrt{M}(\mathbf{W} - \mathbf{w}\mathbf{w}^{\mathrm{T}})^{1/2}, \tag{3.23}$$

which recovers the NETF [25, 28]. Note that $\mathbf{\Delta Q}$ with $\mathbf{Q}$ an $M \times M$ orthogonal matrix such that $\mathbf{Q1} = \mathbf{1}$ also provide a solution to (3.21) if $\mathbf{D} = \mathbf{w}\mathbf{1}^{\mathrm{T}}$.[1] The following lemma states how to choose the orthogonal matrix $\mathbf{Q}$ in an optimal way.

LEMMA 3.2. *Let $\mathbf{\Delta}$ be any $M \times M$ matrix such that (i) $\mathbf{\Delta 1} = \mathbf{0}$ and (ii)*

$$\frac{1}{M}\mathbf{\Delta}\mathbf{\Delta}^{\mathrm{T}} = \mathbf{W} - \mathbf{w}\mathbf{w}^{\mathrm{T}} \tag{3.24}$$

and let us assume that $M \leq N_z + 1$. Define the $M \times M$ orthogonal matrix

$$\mathbf{Q}_{\mathrm{opt}} := \mathbf{U}_{\mathrm{opt}}\mathbf{V}_{\mathrm{opt}}^{\mathrm{T}} \tag{3.25}$$

with the two $M \times M$ orthogonal matrices $\mathbf{U}_{\mathrm{opt}}$ and $\mathbf{V}_{\mathrm{opt}}$ given by the singular value decomposition of the $M \times M$ matrix

$$\mathbf{S} = \mathbf{\Delta}(\widehat{\mathbf{Z}}^{\mathrm{f}})^{\mathrm{T}}\widehat{\mathbf{Z}}^{\mathrm{f}}, \qquad \widehat{\mathbf{Z}}^{\mathrm{f}} := \mathbf{Z}^{\mathrm{f}} - \frac{1}{M}\mathbf{Z}^{\mathrm{f}}\mathbf{1}\mathbf{1}^{\mathrm{T}}, \tag{3.26}$$

i.e. $\mathbf{S} = \mathbf{U}_{\mathrm{opt}}\mathbf{\Lambda}_{\mathrm{opt}}\mathbf{V}_{\mathrm{opt}}^{\mathrm{T}}$. Then the transformation matrix

$$\mathbf{D}_{\mathrm{opt}} = \mathbf{w}\mathbf{1}^{\mathrm{T}} + \mathbf{\Delta}\mathbf{Q}_{\mathrm{opt}} \tag{3.27}$$

results in a second-order accurate LETF, which minimizes

$$\widehat{J}(\mathbf{D}) = \frac{1}{M}\sum_{i=1}^{M}\|\mathbf{z}_i^{\mathrm{a}} - \mathbf{z}_i^{\mathrm{f}}\|^2 \tag{3.28}$$

over all second-order accurate transformation matrices $\mathbf{D} \in \mathcal{D}_2$.

Proof. Since $\widehat{\mathbf{Z}}^{\mathrm{f}}\mathbf{1} = \mathbf{0}$, the matrix $\mathbf{S}$ also satisfies $\mathbf{S1} = \mathbf{0}$ in addition to $\mathbf{S}^{\mathrm{T}}\mathbf{1} = \mathbf{0}$, which implies that $\mathbf{Q}_{\mathrm{opt}}\mathbf{1} = \mathbf{1}$ and (3.27) is second-order accurate. Also note that

$$\left(\mathbf{\Delta}(\widehat{\mathbf{Z}}^{\mathrm{f}})^{\mathrm{T}}\widehat{\mathbf{Z}}^{\mathrm{f}}(\widehat{\mathbf{Z}}^{\mathrm{f}})^{\mathrm{T}}\widehat{\mathbf{Z}}^{\mathrm{f}}\mathbf{\Delta}\right)^{-1/2}\mathbf{\Delta}(\widehat{\mathbf{Z}}^{\mathrm{f}})^{\mathrm{T}}\widehat{\mathbf{Z}}^{\mathrm{f}} = \left(\mathbf{S}\mathbf{S}^{\mathrm{T}}\right)^{-1/2}\mathbf{S} \tag{3.29}$$

$$= \left(\mathbf{U}_{\mathrm{opt}}\mathbf{\Lambda}_{\mathrm{opt}}^{-1}\mathbf{U}_{\mathrm{opt}}^{\mathrm{T}}\right)\mathbf{U}_{\mathrm{opt}}\mathbf{\Lambda}_{\mathrm{opt}}\mathbf{V}_{\mathrm{opt}}^{\mathrm{T}} \tag{3.30}$$

[1] The NETF, as proposed in [25], uses randomly chosen orthogonal matrices which satisfy $\mathbf{Q1} = \mathbf{1}$ while the NETF of [28] is based on a non-symmetric square root of $\mathbf{W} - \mathbf{w}\mathbf{w}^{\mathrm{T}}$.

which has been shown in [20] to minimize (3.28) for given forecast and analysis means and covariance matrices and the optimality of $\mathbf{Q}_{\text{opt}} = \mathbf{U}_{\text{opt}}\mathbf{V}_{\text{opt}}^{\text{T}}$ follows. See also [22]. □

A couple of comments should be made on the requirement of $M \leq N_z + 1$ in Lemma 3.2. First, if the number of samples, M, exceeds the dimensions of state space, N_z, then it is computationally preferable to implement the optimal transformation in the form

$$\mathbf{z}_i^{\text{a}} = \overline{\mathbf{z}}^{\text{a}} + \mathbf{T}(\mathbf{z}_i^{\text{f}} - \overline{\mathbf{z}}^{\text{f}}), \tag{3.31}$$

where $\mathbf{T} \in \mathbb{R}^{N_z \times N_z}$ is an appropriately defined symmetric matrix [20, 22]. Second, one could still proceed with (3.25) but should multiply $\mathbf{Q}_{\text{opt}}$ by the projection matrix $\mathbf{I} - \mathbf{1}\mathbf{1}^{\text{T}}/M$ from the right in order to keep the resulting transformation matrix (3.27) mean preserving, i.e., $\mathbf{D}_{\text{opt}}\mathbf{1} = \mathbf{w}$. This additional operation arises from the fact that the matrix $\mathbf{S}$ will have multiple zero singular values.

4. Continuous-time algebraic Riccati equation. We now return to the general case of a first-order accurate transformation matrix $\mathbf{D}$. Then (3.21) leads to a continuous-time algebraic Riccati equation in the symmetric correction $\boldsymbol{\Delta}$. More specifically, upon introducing

$$\mathbf{B} = \mathbf{D} - \mathbf{w}\mathbf{1}^{\text{T}}, \qquad \mathbf{A} = M(\mathbf{W} - \mathbf{w}\mathbf{w}^{\text{T}}) - \mathbf{B}\mathbf{B}^{\text{T}} \tag{4.1}$$

and assuming that $\boldsymbol{\Delta}$ is symmetric, equation (3.21) can be expressed as the continuous-time algebraic Riccati equation

$$\mathbf{A} = \mathbf{B}\boldsymbol{\Delta} + \boldsymbol{\Delta}\mathbf{B}^{\text{T}} + \boldsymbol{\Delta}\boldsymbol{\Delta}. \tag{4.2}$$

Note that (4.2) arises as the stationary solution of the dynamic Riccati equation

$$\frac{\text{d}}{\text{d}\tau}\boldsymbol{\Delta} = -\mathbf{B}\boldsymbol{\Delta} - \boldsymbol{\Delta}\mathbf{B}^{\text{T}} + \mathbf{A} - \boldsymbol{\Delta}\boldsymbol{\Delta}. \tag{4.3}$$

Since (4.3) is controllable [27], solutions, $\boldsymbol{\Delta}(\tau)$, of (4.3) with initial condition $\boldsymbol{\Delta}(0) = \mathbf{0}$ will converge to a solution of (4.2) as $\tau \to \infty$ [4]. Hence numerical time-stepping of (4.3) with the explicit Euler method for sufficiently many iterations will result in an approximate solution to (4.2). This approach has been used for the numerical results displayed later in this paper.

REMARK 4.1. *Alternatively, (4.2) can be solved by applying the Schur vector approach of [14]. The Schur vector approach is based on the extended Hamiltonian matrix*

$$\mathbf{H} = \begin{pmatrix} \mathbf{B}^{\text{T}} & \mathbf{I} \\ \mathbf{A} & -\mathbf{B} \end{pmatrix} \tag{4.4}$$

and its upper triangular Schur decomposition

$$\mathbf{U}^{\text{T}}\mathbf{H}\mathbf{U} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{0} & \mathbf{S}_{22} \end{pmatrix} \tag{4.5}$$

with the real part of the spectrum of $\mathbf{S}_{11}$ being negative and the real parts of the spectrum of $\mathbf{S}_{22}$ being positive. With the orthogonal matrix $\mathbf{U}$ partitioned accordingly, the solution of (3.21) is given by

$$\boldsymbol{\Delta} = \mathbf{U}_{21}\mathbf{U}_{11}^{-1}. \tag{4.6}$$

This computational approach requires that the matrix pair $(\mathbf{A}^{1/2}, \mathbf{B})$ is detectable [27, 14]. Since this condition may not always be satisfied for (4.2), we recommend to use (4.3) in order to find approximative solutions to (4.2). Alternatively, one could exploit more general Lagrangian invariant subspace techniques as discussed in [12].

We can now either use the $\mathbf{D}$ from the ETPF and derive a second-order accurate version of the ETPF or we compute an approximate solution $\mathbf{D} \in \mathcal{D}_1^+$ to the optimal transport problem (2.8) and are still able to turn it into a second-order accurate PF. This aspect will be discussed in more detail in Section 5.

5. Sinkhorn approximation to the optimal transport problem. The Sinkhorn approximation to the optimal transport problem defined by the cost functional (2.8) and $\mathbf{D} \in \mathcal{D}_1^+$ is provided by the regularised cost functional

$$\mathbf{D}_{\min}(\lambda) = \arg\min J_{\mathrm{SH}}(\mathbf{D}) = \sum_{i,j=1}^{M} \left\{ d_{ij} \|\mathbf{z}_i^{\mathrm{f}} - \mathbf{z}_j^{\mathrm{f}}\|^2 + \frac{1}{\lambda} d_{ij} \ln \frac{d_{ij}}{d_{ij}^0} \right\} \tag{5.1}$$

where $\lambda > 0$ is a regularization parameter and d_{ij}^0 are the entries of $\mathbf{D}_0$ defined in (3.11). Each parameter λ is associated with a specific $\mathbf{D}_{\min}(\lambda) \in \mathcal{D}_1^+$ and $\lambda \to \infty$ leads back to the original cost function (2.8). While, one the other hand, the choice $\lambda \to 0$ leads to (3.11) as the unique minimizer. This follows from the fact that the regularization term in (5.1) is minimal for $d_{ij} = d_{ij}^0$, i.e., $\lim_{\lambda \to 0} \mathbf{D}_{\min}(\lambda) = \mathbf{D}_0$.

REMARK 5.1. *After determining $\mathbf{D}_{\min}(\lambda)$ it is possible to add an appropriate corresponding second-order correction term $\mathbf{\Delta}(\lambda)$ which, depending on λ, leads to different second-order accurate particle filters, e.g., $\lambda \to 0$ leads to the NETF and $\lambda \to \infty$ to the second-order corrected ETPF. In other words, varying λ allows one to naturally bridge between the NETF and the second-order corrected ETPF.*

There exists a straightforward iterative method for finding the minimizer of (5.1). First one notes that the minimizer is of the form

$$\mathbf{D}_{\min}(\lambda) = \mathrm{diag}(\mathbf{u})\, \mathbf{K}\, \mathrm{diag}(\mathbf{v}), \tag{5.2}$$

where $\mathbf{u} \in \mathbb{R}^{M \times 1}$ and $\mathbf{v} \in \mathbb{R}^{M \times 1}$ are two non-negative vectors and $\mathbf{K}$ has entries

$$k_{ij} = e^{-\lambda \|\mathbf{z}_i^{\mathrm{f}} - \mathbf{z}_j^{\mathrm{f}}\|^2}. \tag{5.3}$$

The unknown vectors $\mathbf{u}$ and $\mathbf{v}$ can be computed by Sinkhorn's fixed point iteration

$$\{M w_i / (\mathbf{K}\mathbf{v})_i\} \to \mathbf{u}, \qquad \{1/(\mathbf{K}\mathbf{u})_i\} \to \mathbf{v}. \tag{5.4}$$

The Sinkhorn approximation requires $\mathcal{O}(M^2)$ operations. See [7] for an efficient implementation and additional details.

Let us denote the iterates of $\mathbf{u}$ and $\mathbf{v}$ by $\mathbf{u}^l$ and $\mathbf{v}^l$, respectively, where we always update $\mathbf{u}$ first according to the formula to the left in (5.4). Then the associated

$$\mathbf{D}^l = \mathrm{diag}(\mathbf{u}^l)\, \mathbf{K}\, \mathrm{diag}(\mathbf{v}^l) \tag{5.5}$$

satisfies $(\mathbf{D}^l)^{\mathrm{T}} \mathbf{1} = \mathbf{1}$ and the weights

$$\mathbf{w}^l = \frac{1}{M} \mathbf{D}^l \mathbf{1} \tag{5.6}$$

converge to $\mathbf{w}$ as $l \to \infty$. If we stop the iteration at an index l_*, then we define the associated transformation matrix by

$$\mathbf{D} = \mathbf{D}^{l_*} - (\mathbf{w}^{l_*} + \mathbf{w}) \mathbf{1}^{\mathrm{T}}. \tag{5.7}$$

The index l_* can be determined by the condition

$$\|\mathbf{w}^{l_*} - \mathbf{w}\| \leq \varepsilon \tag{5.8}$$

for sufficiently small $\varepsilon > 0$, e.g. $\varepsilon = 10^{-8}$.

6. Algorithmic summary. We summarise the key steps of the second-order accurate ETPF implementation based upon the Sinkhorn approximation to the optimal transport problem. The Sinkhorn approximation can, of course, be replaced by any available direct solver for the optimal transport problem.

We assume that a set of forecast ensemble members, $\mathbf{Z}^{\mathrm{f}}$, and a vector of importance weights, $\mathbf{w}$, are given. Then the following steps are performed:

(i) Select a regularization parameter $\lambda > 0$ for the Sinkhorn approximation to the optimal transport algorithm. Compute the matrix $\mathbf{K}$ according to (5.3). Normalize the entries of $\mathbf{K}$ such that all entries satisfy $-\lambda^{-1} \ln k_{ij} \leq 1$. Recursively compute vectors $\mathbf{u}^l$ and $\mathbf{v}^l$ according to the update formula (5.4). Start with $\mathbf{v}^0 = \mathbf{1}$. Iterate till the transformation matrix (5.5) and its associated weight vector (5.6) satisfy (5.8). Note that (5.5) should satisfy $\mathbf{1}^{\mathrm{T}} \mathbf{D}^l = \mathbf{1}^{\mathrm{T}}$ in each iteration. We used $\varepsilon = 10^{-8}$ in our experiments. One finally obtains the transform matrix $\mathbf{D}$ using (5.7).

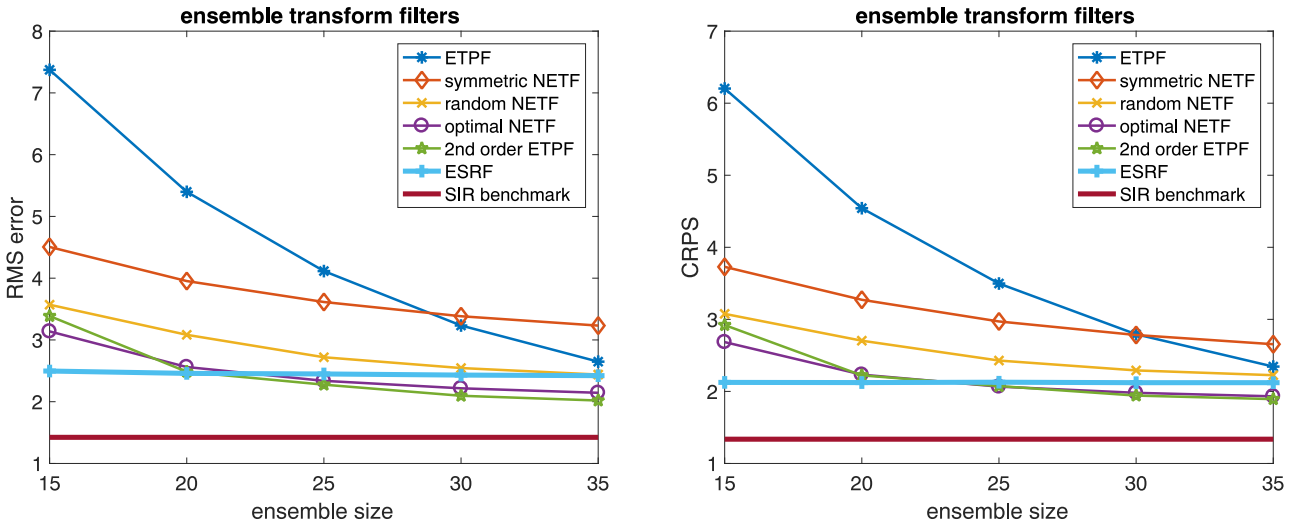

FIG. 7.1. *RMS errors (left panel) and CRPS (right panel) for various second-order accurate LETFs compared to the ETPF and the ESRF as a function of the ensemble size, M, for the Lorenz-63 model. We also provide the RMS error and the CRPS obtained from a standard particle filter with resampling and $M = 1000$ ensemble members.*

(ii) Solve the Riccati equation (4.2) for the correction $\mathbf{\Delta}$ by solving the dynamic Riccati equation (4.3) with the explicit Euler method, step-size $\Delta\tau = 0.1$, and initial condition $\mathbf{\Delta}(0) = \mathbf{0}$. The iteration is stopped whenever

$$\|\mathbf{\Delta}((k+1)\Delta\tau) - \mathbf{\Delta}(k\Delta\tau)\|_\infty \leq 10^{-3} \tag{6.1}$$

and we set $\mathbf{\Delta} = \mathbf{\Delta}((k+1)\Delta\tau)$.

(iii) The analysis ensemble is given by

$$\mathbf{Z}^{\mathrm{a}} = \mathbf{Z}^{\mathrm{f}}\widehat{\mathbf{D}} = \mathbf{Z}^{\mathrm{f}}(\mathbf{D} + \mathbf{\Delta}). \tag{6.2}$$

We mention that the proposed second-order accurate ETPF can be used instead of the standard ETPF in a hybrid filter, as described in [6], and, when applied to spatially extended system, can also be used with localization. More specifically, a hybrid filter is based on factorizing the likelihood (2.2) into

$$\pi(\mathbf{y}|\mathbf{z}) = \pi(\mathbf{y}|\mathbf{z})^{\alpha} \times \pi(\mathbf{y}|\mathbf{z})^{1-\alpha} \tag{6.3}$$

and applying different filters to each of the two factors. R-localization, on the other hand, leads to different transformation matrices $\mathbf{D}(x_k)$ at each grid point x_k of the computational domain. See [5, 22] for further details.

7. Numerical examples. We now demonstrate the numerical behavior of the proposed second-order accurate ETPF as summarized in Section 6. The first two experiments are based on the Lorenz-63 and Lorenz-96 models, respectively, and its data assimilation setting of [6]. We finally apply the second-order accurate filters to parameter estimation of the scene-viewing model *SceneWalk* [9].

7.1. Lorenz-63. We use the chaotic Lorenz-63 system [17] with the standard parameter setting $\sigma = 10$, $\rho = 28$, and $\beta = 8/3$, and observe the first component of the three dimensional system in observation intervals of $\Delta t_{\mathrm{obs}} = 0.12$ with observation error variance $R = 8$. A total of $K = 500,000$ assimilation steps are performed. Since the model dynamics is deterministic, particle rejuvenation

$$\mathbf{z}_j^{\mathrm{a}} \to \mathbf{z}_j^{\mathrm{a}} + \sum_{i=1}^{M} (\mathbf{z}_i^{\mathrm{f}} - \bar{\mathbf{z}}^{\mathrm{f}}) \frac{\beta\xi_{ij}}{\sqrt{M-1}} \tag{7.1}$$

is applied with $\beta = 0.2$ and ξ_{ij} independent and identically distributed Gaussian random variables with mean zero and variance one. Simulations with $\beta = 0.15$ and $\beta = 0.25$ gave similar results to those reported here. This data assimilation setting has already been used in [6] and [5] since it leads to non-Gaussian forecast and analysis distributions and a particle filter is able to outperform EnKFs in the limit of large ensemble sizes.

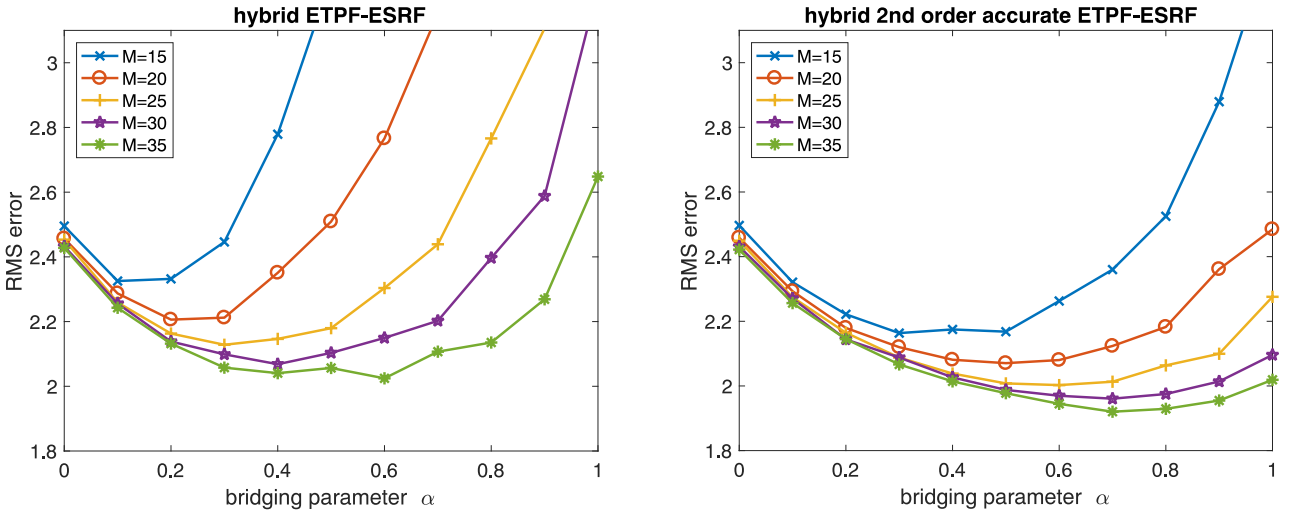

FIG. 7.2. *Hybrid filter with standard EPTF (left panel) and second-order accurate ETPF (right panel) applied to the Lorenz-63 model. Time-averaged RMS errors are displayed as a function of the bridging parameter α. Please note that $\alpha = 0$ corresponds to the standard ESRF, while $\alpha = 1$ corresponds to the ETPF and the second-order corrected ETPF, respectively.*

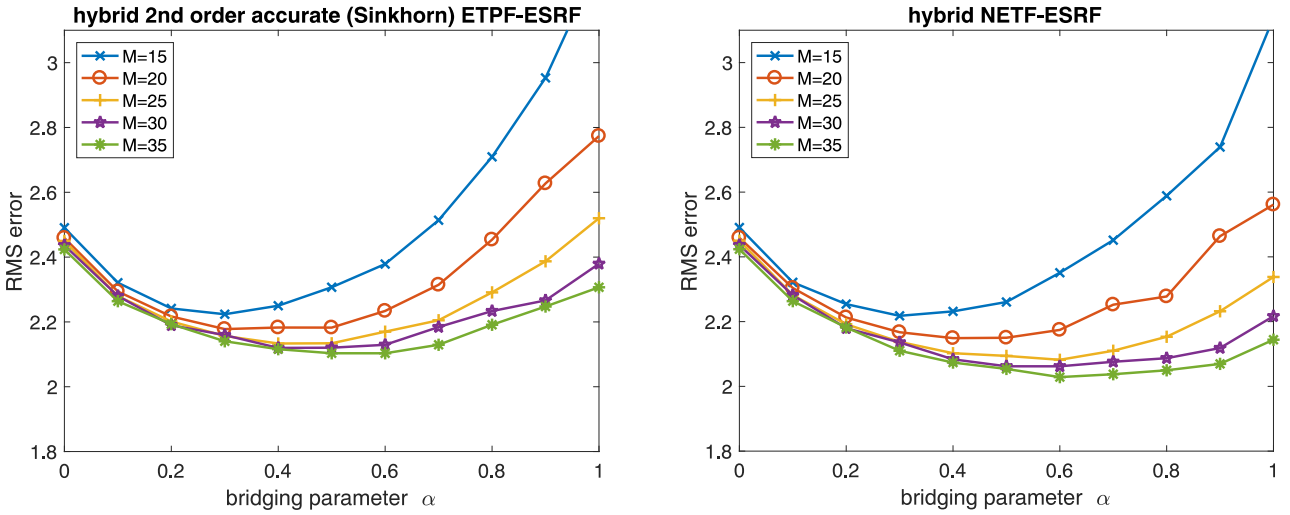

FIG. 7.3. *Second-order hybrid ETPF-ESRF with the optimal transport problem solved by the Sinkhorn approximation with $\lambda = 10$ (left panel) and hybrid NETF-ESRF with the the orthogonal matrix $\mathbf{Q}$ as defined in (3.25) (right panel) applied to the Lorenz-63 model. Time-averaged RMS errors are displayed as a function of the bridging parameter α.*

A comparison between a standard particle filter with resampling, the EnKF, and the ETPF can be found in [5]. Here we are, however, interested in the performance of second-order accurate filters for small ensemble sizes in the range $M \in \{15, 20, \ldots, 35\}$. See Figure 7.1 for the resulting time-averaged RMS errors. It can be clearly seen that the second-order corrected ETPF and the NETF with optimally chosen rotation matrix leads to the smallest RMS errors for $M \geq 25$, while the standard ensemble square root filter (ESRF) [10] is optimal for smaller ensemble sizes. It can be seen that the standard ETPF is not competitive except for $M = 35$. The same findings apply for the continuous ranked probability score (CRPS) [3], which we computed for the observed component of the Lorenz-63 system. The results can be found in Figure 7.1.

We also display the RMS errors for implementations of the NETF with randomly chosen orthogonal matrices, $\mathbf{Q}$, as suggested by [25], and with $\mathbf{Q} = \mathbf{I}$ in Figure 7.1. It can be seen that both choices lead to substantially increased RMS errors.

We now test the second-order accurate transform filters within the hybrid filter framework proposed in [6]. More specifically, the hybrid filter of [6] with a second-order accurate transform filter applied first is implemented for ensemble sizes varying between $M = 15$ and $M = 35$. The bridging parameter, α, of the hybrid filter approach is chosen such that $\alpha = 0$ corresponds to the standard ESRF while $\alpha = 1$ leads to a purely second-

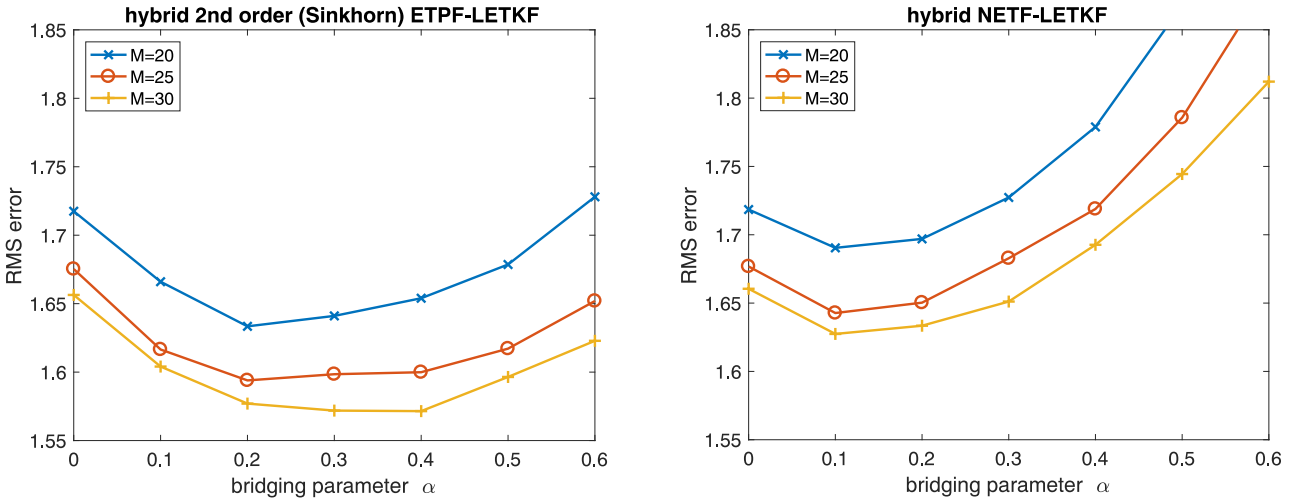

FIG. 7.4. *Second-order hybrid ETPF-LETKF with the optimal transport problem solved by the Sinkhorn approximation with* $\lambda = 10$ *(left panel) and hybrid NETF-LETKF with the the orthogonal matrix* $\mathbf{Q}$ *at each grid point defined as in (3.25) (right panel) applied to the Lorenz-96 model. Time-averaged RMS errors are displayed as a function of the bridging parameter* α*. The choice* $\alpha = 0$ *corresponds to the LETKF*

order accurate ETPF. We perform experiments for fixed bridging parameters $\alpha \in \{0, 0.1, 0.2, \ldots, 0.9, 1.0\}$ and compare the resulting RMS errors to those from a hybrid method based on the standard ETPF in Figure 7.2. The improvement achieved by the second-order correction is clearly visible. In both cases, the ETPF has been implemented using a direct solver for the underlying optimal transport problem.

We next replace the direct solver for the optimal transport problem by the Sinkhorn approximation with regularization parameters $\lambda = 10$ and $\lambda = 40$. The RMS errors for the resulting hybrid filter with $\lambda = 10$ can be found in Figure 7.3, while $\lambda = 40$ leads to RMS errors which are very close to those displayed in the right panel of Figure 7.2, which are based on a direct solver for the optimal transport problem.

We also implement the hybrid filter of [6] with the ETPF being replaced by the second-order accurate NETF with the rotation matrix, $\mathbf{Q}$, defined as in (3.25). We denote this hybrid filter by NETF-ESRF. The numerical results can also be found in Figure 7.3. Overall, we find that a second-order corrected hybrid ETPF-ESRF and the NETF-ESRF with optimally chosen rotation matrix perform quite comparable in terms of their RMS errors. The same holds true for the associated CRPS (not displayed).

7.2. Lorenz-96. We now implement the spatially-extended Lorenz-96 system [18] with the standard parameter setting of $p = 40$ grid points and forcing $F = 8$. We observe every second grid point in observation intervals of $\Delta t_{\text{obs}} = 0.11$ with observation error variance $R = 8$. A total of $K = 50,000$ assimilation steps are performed. Contrary to the Lorenz-63 experiments, no particle rejuvenation is applied, i.e., $\beta = 0$ in (7.1). We also apply localization [22] with the localization radius r_{loc} set equal to four grid points and compute separate transformation matrices $\mathbf{D}(x_k)$ for each grid point $x_k = k$, $k = 1, \ldots, 40$. Localization is necessary for this test problem as the ensemble sizes, $M \in \{20, 25, 30\}$, are smaller than the number of grid points, $p = 40$. This specific DA setting has already been used in [5] and [6].

We compare two hybrid methods based on a combination of second-order accurate filters and the local ensemble transform Kalman filter (LETKF) [13]. All filters use R-localization [13, 22] and the transportation cost at each grid point $x_k = k$, $k = 1, \ldots, 40$, is given by

$$J(\mathbf{D}(x_k)) = \sum_{i,j=1}^{M} d_{ij}(x_k) |u_i^{\text{f}}(x_k) - u_j^{\text{f}}(x_k)|^2, \tag{7.2}$$

where $u_i^{\text{f}}(x_k) \in \mathbb{R}$ denotes the forecast value of the ensemble member $\mathbf{z}_i^{\text{f}} \in \mathbb{R}^{40}$ at grid point x_k and $\mathbf{D}(x_k) = \{d_{ij}(x_k)\} \in \mathbb{R}^{M \times M}$.

The results for the hybrid NETF-LETKF filter and the hybrid second-order corrected ETPF-LETKF can be found in Figure 7.4. The hybrid second-order corrected ETPF-LETKF is implemented using the Sinkhorn approximation with $\lambda = 10$ and leads to significant improvements over the hybrid NETF-LETKF and also over the hybrid ETPF-LETKF of [6]. The CRPS leads to a qualitatively similar assessment.

7.3. Estimating parameters for a dynamic scene viewing model. The scene-viewing model *SceneWalk*, as recently proposed by [9], provides a relatively simple mathematical model for a sequence of eye fixations during scene viewing. The model dynamically evolves a two-dimensional array of probabilities, $\pi_{ij}(t)$, for the next fixation target, which is conditioned on past fixations. More specifically, the model consists of two sets of ordinary differential equations

$$\frac{\mathrm{d}A_{ij}(t)}{\mathrm{d}t} = -\omega_A A_{ij}(t) + \omega_A \frac{S_{ij} \cdot G_A(x_i, y_j; x_f, y_f)}{\sum_{kl} S_{kl} \cdot G_A(x_k, y_l; x_f, y_f)} \tag{7.3}$$

$$\frac{\mathrm{d}F_{ij}(t)}{\mathrm{d}t} = -\omega_F F_{ij}(t) + \omega_F \frac{G_F(x_i, y_j; x_f, y_f)}{\sum_{kl} G_F(x_k, y_l; x_f, y_f)} \tag{7.4}$$

for the spatial attention and fixation, respectively, together with a set of transformation rules

$$u_{ij}(t) = \frac{[A_{ij}(t)]^\lambda}{\sum_{kl}[A_{kl}(t)]^\lambda} - c_{inhib}\frac{[F_{ij}(t)]^\gamma}{\sum_{kl}[F_{kl}(t)]^\gamma} \ , \tag{7.5}$$

$$u^*(u) = \begin{cases} u & u > \eta \\ \eta e^{\frac{u-\eta}{\eta}} & u \leq \eta \end{cases} \ , \tag{7.6}$$

which finally produce the desired array of fixation probabilities

$$\pi_{ij}(t) = (1-\zeta)\frac{u^*_{ij}(t)}{\sum_{kl} u^*_{kl}(t)} + \zeta \frac{1}{\sum_{kl} 1} \ . \tag{7.7}$$

The functions $G_{A/F}$ in (7.3)-(7.4) are Gaussians given by

$$G_{A/F}(x, y; x_f, y_f) = \frac{1}{2\pi\sigma^2_{A/F}} \exp\left(-\frac{(x-x_f)^2 + (y-y_f)^2}{2\sigma^2_{A/F}}\right) \tag{7.8}$$

and $\{S_{ij}\}$ is a static saliency map. See [9, 23] for a detailed description of the model. The *SceneWalk* model contains 9 parameters, which have been estimated in [23] using maximum likelihood estimates. Here we estimate σ_F in (7.8) and ω_F in (7.4) with the remaining seven parameter values taken from [23]. We start from a uniform prior over the interval $[1, 5]$ for the first variable and a uniform prior over the interval $[8, 16]$ for the second variable, respectively.

Our experiments consist of first computing the importance weights for each sample from the prior under given pool of fixation paths and then using an LETF to transform those samples into equally weighted samples from the posterior parameter distribution. We wish to demonstrate the impact of different LETFs in terms statistical consistency and distribution of their posterior samples.

The importance weights resulting from a given pool of scan paths and $M = 500$ samples from the prior distribution can be found in Figure 7.5. The effective sample size is about ninety.

We implement the NETF method with $\mathbf{Q} = \mathbf{I}$ (symmetric NETF), the NETF with the optimal $\mathbf{Q}$ (optimal NETF), the ETPF, and the second-order accurate ETPF. The distribution of transformed versus prior sample values for each of the two parameters separately can be found in Figure 7.6. While the optimal NETF leads to a nearly linear relation between the prior and transformed samples, the symmetric NETF leads to a rather non-regular structure. At the same time we find that the second-order accurate ETPF leads to large fluctuations in the transformed samples with some samples leaving the prior range. Since this behavior is violating Bayes' law, it must be seen as a undesirable effect of enforcing strict second-order accuracy. The associated two-dimensional scatter plots of the prior and transformed samples can be found in Figure 7.7. These plots show even more clearly that second-order accurate methods can lead to transformed samples, which violate Bayes' law. Nevertheless, all methods consider qualitatively capture the posterior distribution.

8. Conclusions. We have proposed and tested second-order variants of the ETPF. These modifications are computationally attractive since it allows one to replace the computationally expensive solution of an optimal transport problem by its Sinkhorn approximation. Furthermore, if the regularization parameter, λ, in the Sinkhorn approximation is set to zero, then we recover the NETF [25] with an optimally chosen orthogonal matrix $\mathbf{Q}_{\text{opt}}$ in (3.25, while $\lambda \to \infty$ leads formally back to the optimal transport implementation of the ETPF.

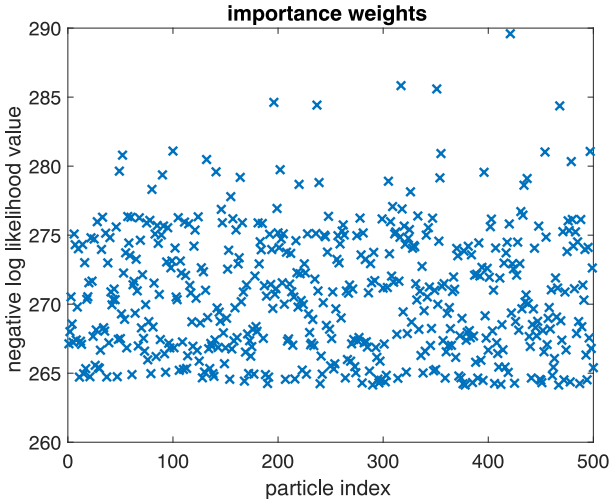

FIG. 7.5. *Importance weights for $M = 500$ samples in two-dimensional parameter space for the Scene Walk model. The effective sample size is $M_{\text{eff}} \approx 90$.*

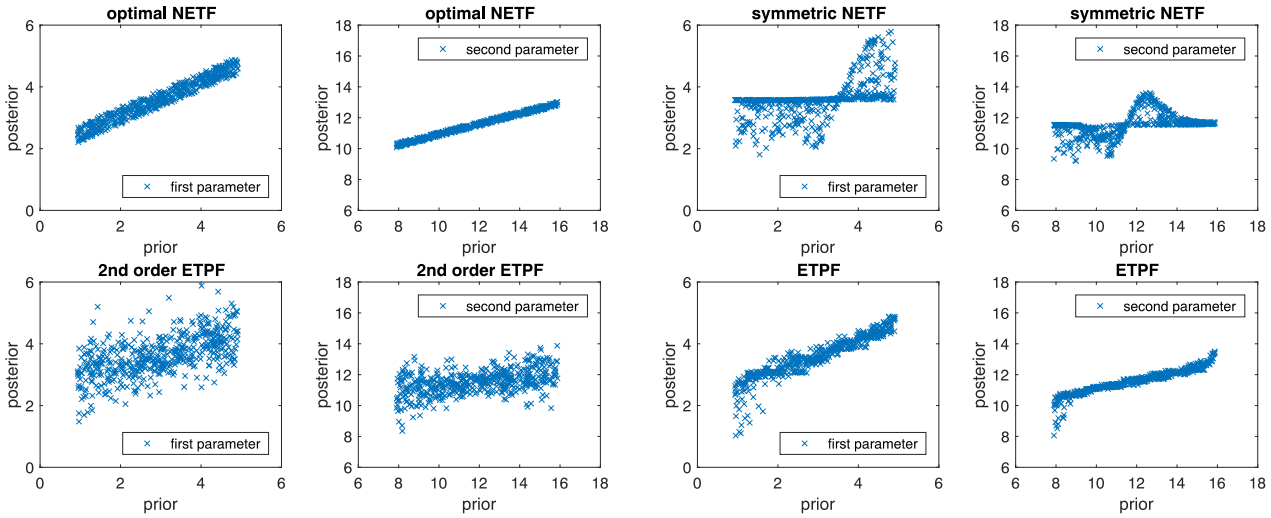

FIG. 7.6. *Prior vs posterior samples for Scene Walk model: optimal NEFT (left panel, top row), symmetric NETF (right panel, top row). The two panels also show the ETPF (right panel, bottom row) and the 2nd order corrected ETPF (left panel, bottom row) for comparison. Both the optimal NETF and the ETPF lead to relatively concentrated sample sets, following nearly liner relationships.*

As a byproduct, we also found that the NETF with an optimally chosen orthogonal matrix, $\mathbf{Q}$, leads to smaller RMSEs compared to a random choice, as suggested in [25].

The second-order accurate ETPF can be put into the hybrid ensemble transform particle framework of [6] and can be combined with localization as necessary for spatially extended evolution equation [10, 22, 5] such as the Lorenz-96 model.

The numerical findings for the Lorenz-63 and Lorenz-96 models confirm that the methodology proposed in this paper together with the hybrid approach of [6] provides a powerful framework for performing sequential data assimilation. We mention that all methods considered in this paper can be combined with alternative proposal densities, which lead to more balanced importance weights (2.7) [26].

It should be noted though, that second-order accuracy comes at a price, i.e., the entries of the transformation matrix $\widehat{\mathbf{D}}$ are not necessarily non-negative, as it is the case for the ETPF transformation matrix $\mathbf{D}$. Hence the analysis ensemble is not necessarily contained in the convex hull spanned by the forecast ensemble. This can cause non-physical states if, for example, the states should only take values in a bounded interval or semi-interval, as has been demonstrated for the *SceneWalk* model.

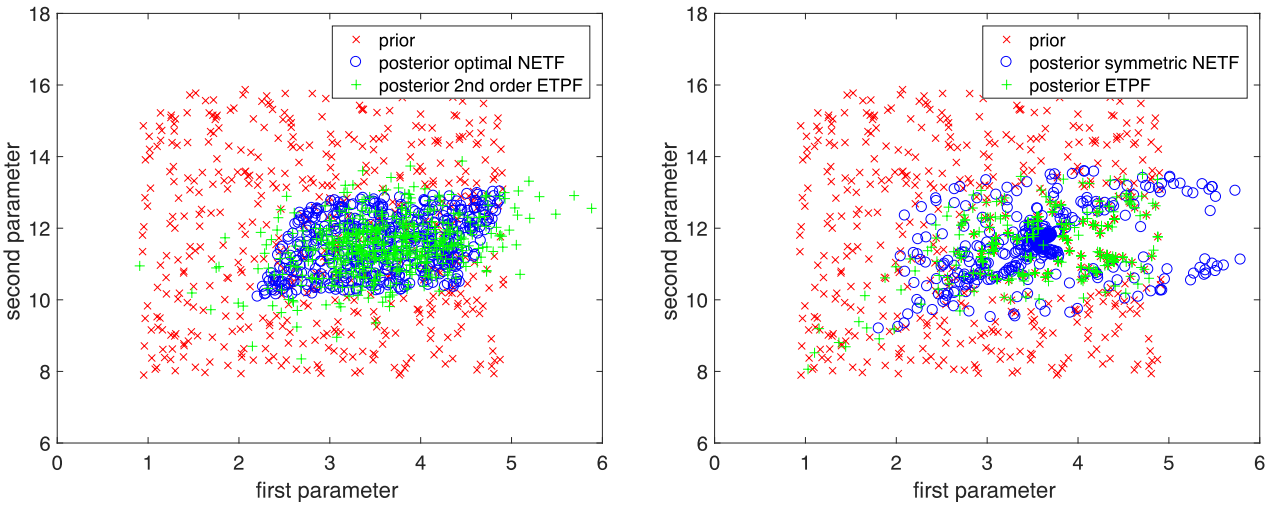

FIG. 7.7. *Prior vs posterior samples for Scene Walk model: optimal NEFT (left panel), symmetric NETF (right panel). Both panels show the ETPF and the 2nd-order corrected ETPF, respectively, for comparison. It can be clearly seen that the symmetric NETF and the 2nd order corrected ETPF lead to posterior samples which are outside the range of the prior samples.*

Acknowledgments. We like to thank Hans-Rudolf Künsch and Sylvain Robert for discussions on second-order corrections to linear ensemble transform filters. We also thank Ralf Engbert and Heiko Schütt for providing the data set used in Section 7.3. This research has been partially funded by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1114 "Scaling Cascades in Complex Systems", Project (A02) "Multiscale data and asymptotic model assimilation for atmospheric flows".

REFERENCES

[1] J. ANDERSON, *A non-Gaussian ensemble filter update for data assimilation*, Monthly Weather Review, 138 (2010), pp. 4186–4198.
[2] T. BENGTSSON, P. BICKEL, AND B. LI, *Curse of dimensionality revisited: Collapse of the particle filter in very large scale systems*, in IMS Lecture Notes - Monograph Series in Probability and Statistics: Essays in Honor of David F. Freedman, vol. 2, Institute of Mathematical Sciences, 2008, pp. 316–334.
[3] J. BRÖCKER, *Evaluating raw ensembles with the continuous ranked probability score*, Q.J.R. Meteor. Soc., 138 (2012), pp. 1611–1617.
[4] R. BUCY AND P. JOSEPH, *Filtering for stochastic processes with applications to guidance*, AMS Chelsea Publishing, Providence, Rhode Island, 2nd ed., 1987.
[5] Y. CHEN AND S. REICH, *Assimilating data into scientific models: An optimal coupling perspective*, in Frontiers in Applied Dynamical Systems: Reviews and Tutorials, vol. 2, Springer-Verlag, New York, 2015, pp. 75–118.
[6] N. CHUSTAGULPROM, S. REICH, AND M. REINHARDT, *A hybrid ensemble transform filter for nonlinear and spatially extended dynamical systems*, SIAM/ASA J. Uncertainty Quantification, 4 (2016), pp. 592–608.
[7] M. CUTURI, *Sinkhorn distances: Lightspeed computation of optimal transport*, in NIPS 2013, 2013.
[8] A. DOUCET, N. DE FREITAS, AND N. G. (EDS.), *Sequential Monte Carlo methods in practice*, Springer-Verlag, Berlin Heidelberg New York, 2001.
[9] R. ENGBERT, H. A. TRUKENBROD, S. BARTHELMÉ, AND F. A. WICHMANN, *Spatial statistics and attentional dynamics in scene viewing*, Journal of Vision, 15 (2015).
[10] G. EVENSEN, *Data assimilation. The ensemble Kalman filter*, Springer-Verlag, New York, 2006.
[11] M. FREI AND H. KÜNSCH, *Bridging the ensemble Kalman and particle filters*, Biometrika, 100 (2013), pp. 781–800.
[12] G. FREILING, V. MEHRMANN, AND H. XU, *Existence, uniqueness, and parametrization of Lagrangian invariant subspaces*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 1045–1069.
[13] B. HUNT, E. KOSTELICH, AND I. SZUNYOGH, *Efficient data assimilation for spatialtemporal chaos: A local ensemble transform Kalman filter*, Physica D, 230 (2007), pp. 112–137.
[14] A. LAUB, *A Schur method for solving algebraic Riccati equations*, IEEE Trans. Automatic Control, 24 (1979), pp. 913–921.
[15] K. LAW, A. STUART, AND K. ZYGALAKIS, *Data Assimilation: A Mathematical Introduction*, Springer-Verlag, New York, 2015.
[16] J. LEI AND P. BICKEL, *A moment matching ensemble filter for nonlinear and non-Gaussian data assimilation*, Mon. Weath. Rev., 139 (2011), pp. 3964–3973.
[17] E. LORENZ, *Deterministic non-periodic flows*, J. Atmos. Sci., 20 (1963), pp. 130–141.
[18] ———, *Predictibility: A problem partly solved*, in Proc. Seminar on Predictibility, vol. 1, ECMWF, Reading, Berkshire, UK, 1996, pp. 1–18.
[19] S. METREF, E. COSME, C. SNYDER, AND P. BRASSEUR, *A non-Gaussian analysis scheme using rank histograms for ensemble data assimilation*, Nonlinear Processes in Geophysics, 21 (2013), pp. 869–885.
[20] I. OLKIN AND F. PUKELSHEIM, *The distance between two random vectors with given dispersion matrices*, Linear Algebra and

its Applications, 48 (1982), pp. 257–263.
[21] S. Reich, *A nonparametric ensemble transform method for Bayesian inference*, SIAM J. Sci. Comput., 35 (2013), pp. A2013–A2024.
[22] S. Reich and C. Cotter, *Probabilistic Forecasting and Bayesian Data Assimilation*, Cambridge University Press, Cambridge, 2015.
[23] H. Schütt, L. Rothkegel, H. Trukenbrod, S. Reich, F. Wichmann, and R. Engbert, *Likelihood-based parameter estimation and comparison of dynamical cognitive models*, Tech. Rep. ArXiv:1606.07309, accepted for publication in Psychological Review, University of Potsdam, 2016.
[24] A. Stordal, H. Karlsen, G. Nævdal, H. Skaug, and B. Vallés, *Bridging the ensemble Kalman filter and particle filters: the adaptive Gaussian mixture filter*, Comput. Geosci., 15 (2011), pp. 293–305.
[25] J. Tödter and B. Ahrens, *A second-order exact ensemble square root filter for nonlinear data assimilation*, Mon. Wea. Rev., 143 (2015), pp. 1347–1367.
[26] P. Van Leeuwen, *Nonlinear data assimilation for high-dimensional systems*, in Frontiers in Applied Dynamical Systems: Reviews and Tutorials, vol. 2, Springer-Verlag, New York, 2015, pp. 1–73.
[27] W. Wonham, *On a matrix Riccati equation of stochastic control*, SIAM J. Contr., 6 (1968), pp. 681–697.
[28] X. Xiong, I. Navon, and B. Uzungoglu, *A note on the particle filter with posterior Gaussian resampling*, Tellus, 85A (2006), pp. 456–460.