AN ADAPTIVE COVARIANCE PARAMETERIZATION TECHNIQUE FOR THE ENSEMBLE GAUSSIAN MIXTURE FILTER*

ANDREY A POPOV† AND RENATO ZANETTI‡

Abstract. The ensemble Gaussian mixture filter combines the simplicity and power of Gaussian mixture models with the provable convergence and power of particle filters. The quality of the ensemble Gaussian mixture filter heavily depends on the choice of covariance matrix in each Gaussian mixture. This work extends the ensemble Gaussian mixture filter to an adaptive choice of covariance based on the parameterized estimates of the sample covariance matrix. Through the use of the expectation maximization algorithm, optimal choices of the covariance matrix parameters are computed in an online fashion. Numerical experiments on the Lorenz '63 equations show that the proposed methodology converges to classical results known in particle filtering. Further numerical results with more advanced choices of covariance parameterization and the medium-size Lorenz '96 equations show that the proposed approach can perform significantly better than the standard EnGMF, and other classical data assimilation algorithms.

Key words. data assimilation, Gaussian mixture model, particle filtering, expectation maximization

MSC codes. 60G25, 62L12, 62M20, 93E11

1. Introduction.

Sequential data assimilation [5, 37] aims to perform Bayesian inference on the state of some natural process from an inaccurate computational model and sparse and noisy observations. Traditionally, particle filter methods have been viewed as theoretically nice, but practically useless for inference of high dimensional systems. Recent advances in particle filters for high dimensions [43] have challenged this view.

Classical particle filters such as the bootstrap particle filter [37] make and empirical measure assumption on the prior distribution. Conversely filters such as the ensemble Kalman filter [17, 10] make an assumption that the first two moments of the empirical distribution are the only ones relevant to performing the inference, similar to a Gaussian assumption on the distribution. Gaussian mixture models (GMM) can extend the idea of an empirical measure approximation of the prior to a larger set of possible prior distributions, that combines the best of both worlds: it is able to represent non-Gaussian distributions while still assuming that the only two moments that matter to each mixture mode are its mean and covariance. The Gaussian sum filter [41] takes advantage of nice properties of GMMs, suffers from needing to propagate the covariance of each mode, and from requiring the need of many heuristics to ensure that the modes do not degenerate [36].

The ensemble Gaussian mixture filter (EnGMF) [4, 25, 44] and its related cousin the adaptive Gaussian mixture filter (AGMF) [42, 43] are sequential data assimilation algorithms that make use of the GMM approximation to the prior through the use of kernel density estimation (KDE) techniques. The EnGMF is based on the observation that Gaussian mixture models, under linear observation assumptions, are closed under multiplication [3]. The quality of the inference produced by the EnGMF is directly

*Submitted to the editors of SIAM SISC January 13, 2024.

Funding: This work was sponsored in part by DARPA (Defense Advanced Research Projects Agency) under STTR contract number W31P4Q-21-C-0032.

†Oden Institute for Computational Engineering and Sciences, University of Texas at Austin (andrey.a.popov@utexas.edu).

‡Department of Aerospace Engineering & Engineering Mechanics, University of Texas at Austin (renato@utexas.edu).

related to the accuracy of the GMM assumption about the prior distribution. This prior distribution is typically determined in whole by Monte Carlo samples through choices of the means and covariances of the GMM. While the choice of means is readily apparent as that of the Monte Carlo samples, the choice of covariance is less straightforward, and is the subject of much of the research surrounding KDE [40, 9].

The key innovations of this work are as follows: We first provide theoretical results that show the convergence of the EnGMF for a certain class of probabilistic assumption on the bandwidth parameter in the classical EnGMF algorithm. We next generalize the EnGMF by introducing the parameterization of statistical covariance matrix estimates from other ensemble-based filters to the EnGMF. We finally show how the EnGMF machinery could be used to choose the value of these parameters in an optimal adaptive fashion by utilizing the expectation maximization algorithm. Thus the sum total of these results is the adaptive Gaussian mixture filter (AEnGMF) which utilizes all this machinery for inference.

This work is organized as follows: we first introduce the data assimilation problem and the EnGMF in section 2. We next present the adaptive ensemble Gaussian mixture filter in section 3 along with the expectation maximization algorithm in subsection 3.1. Numerical experiments are provided in section 4, and concluding remarks in section 5.

2. Background. Assume that we are given a model that evolves a natural process of interest from time index $i-1$ to time index i,

$$x_i^t = \mathcal{M}(x_{i-1}^t) + \xi_i, \tag{2.1}$$

with model error ξ_i. For the remainder of this paper, we assume that the model error ξ_i is always zero, and thus the model (2.1) is exact.

The goal is to estimate the true state x^t of said process given some non-linear observation,

$$y_i = \mathcal{H}(x_i^t) + \eta_i, \tag{2.2}$$

with observation operator $\mathcal{H}$ and an additive error term η_i. Denote with Y_i all the observations up to and including time index i,

$$Y_i = \{y_1,\ y_2,\ \ldots\ y_i\}. \tag{2.3}$$

Given a prior at time index i, namely $x_i^b = x_i^t|Y_{i-1}$, we aim to perform Bayesian inference on these two sources of information,

$$p(x_i^b|y_i) \propto p(x_i^b)\ p(y_i|x_i^b) \tag{2.4}$$

resulting in the 'analysis', $x_i^a = x_i^b|y_i = x_i^t|Y_i$.

Remark 2.1 (Model error). All of the derivations and algorithms presented in this work do not require the model error in (2.1) to be zero. This assumption is merely made for convenience in this work.

We next describe how a solution to (2.4) can be achieved using Monte Carlo sampling and the EnGMF.

2.1. The Ensemble Gaussian Mixture Filter. Assume that we have a collection of N particles at time index i that is represented as $\mathbf{X}_i^b = [x_{i,1}^b, x_{i,2}^b, \ldots, x_{i,N}^b]$ and is composed of weighted samples from the prior distribution $p(x_i^b)$ with weights

$\{u_{i,j}\}_{j=1}^N$. Given non-linear observations of the truth (2.2), our aim is to find a collection of N particles $\mathbf{X}_i^a = [x_{i,1}^a, x_{i,2}^a, \dots x_{i,N}^a]$ containing samples from the posterior distribution, such that the posterior is the prior conditioned by the observations,

$$x_i^a = x_i^b | y_i, \tag{2.5}$$

solving the Bayesian inference problem (2.4).

As the prior distribution of the particles is unknown, our prior knowledge can be used to construct an approximation thereto. From kernel density estimation theory, the ensemble Gaussian mixture filter (EnGMF) assumes that the distribution of the prior state at time index i, x_i^b, is given by the Gaussian mixture,

$$x_i^b \sim \sum_{j=1}^{N} u_{i,j} \mathcal{N}\left(\bar{x}_{i,j}^b, \mathbf{B}_{i,j}^b\right), \tag{2.6}$$

where each mean exactly corresponds to one of the particles in the ensemble,

$$\bar{x}_{i,j}^b \coloneqq x_{i,j}^b. \tag{2.7}$$

The observation distribution at time index i is given by the Gaussian mixture,

$$y_i | x_i^t \sim \sum_{k=1}^{M} v_{i,k} \mathcal{N}\left(\bar{y}_{i,k}, \mathbf{R}_{i,k}\right), \tag{2.8}$$

which is a generalization of the typical Gaussian assumptions on the observation error made in data assimilation literature.

The posterior distribution at time index i is defined [4] by the Gaussian mixture,

$$x_i^a \sim \sum_{j=1}^{N} \sum_{k=1}^{M} w_{i,j,k} \mathcal{N}\left(\bar{x}_{i,j,k}^a, \mathbf{B}_{i,j,k}^a\right), \tag{2.9}$$

with the following set of definitions,

$$\begin{aligned}
\bar{x}_{i,j,k}^a &= \bar{x}_{i,j}^b - \mathbf{G}_{i,j,k}\left(\mathcal{H}(\bar{x}_{i,j}^b) - \bar{y}_{i,k}\right), \\
\mathbf{B}_{i,j,k}^a &= \left(\mathbf{I} - \mathbf{G}_{i,j,k}\mathbf{H}_{i,j}^T\right)\mathbf{B}_{i,j}^b, \\
\mathbf{G}_{i,j,k} &= \mathbf{B}_{i,j}^b \mathbf{H}_{i,j}^T \left(\mathbf{H}_{i,j}\mathbf{B}_{i,j}^b\mathbf{H}_{i,j}^T + \mathbf{R}_{i,k}\right)^{-1}, \\
w_{i,j,k} &\propto u_{i,j} v_{i,k} \mathcal{N}\left(\bar{y}_{i,k} \,\middle|\, \mathcal{H}(\bar{x}_{i,j}^b), \mathbf{H}_{i,j}\mathbf{B}_{i,j}^b\mathbf{H}_{i,j}^T + \mathbf{R}_{i,k}\right), \\
\mathbf{H}_{i,j} &= \left.\frac{d\mathcal{H}}{dx}\right|_{x=\bar{x}_{i,j}^b}.
\end{aligned} \tag{2.10}$$

where $\bar{x}_{i,j,k}^a$, are the analysis Gaussian mixture means, $\mathbf{B}_{i,j,k}^a$ are the analysis Gaussian mixture covariances, $\mathbf{G}_{i,j,k}$ is similar to a gain matrix, $w_{i,j,k}$ are the Gaussian mixture weights, and $\mathbf{H}_{i,j}$ is the linearization of the observation operator around $\bar{x}_{i,j}^b$. When the observation operator is linear, $\mathcal{H}(x) = \mathbf{H}x$, the posterior GMM (2.9) is exactly the posterior corresponding to the assumed prior (2.6) and the observation (2.8) distributions.

Each Gaussian distribution in (2.6) has the following probability density function,

$$\mathcal{N}(x|\bar{x}_{i,j}^b, \mathbf{B}_{i,j}^b) = \left|2\pi\mathbf{B}_{i,j}^b\right|^{-\frac{1}{2}} e^{-\frac{1}{2}\left(x-\bar{x}_{i,j}^b\right)^T \mathbf{B}_{i,j}^{b,-1}\left(x-\bar{x}_{i,j}^b\right)}, \tag{2.11}$$

with the other Gaussian distributions in (2.8)–(2.10) having a similar form.

Remark 2.2 (Normalization Factors). Note that when either the observation operator $\mathcal{H}$ is non-linear, or the kernel covariance matrices $\mathbf{B}_{i,j}$ are not identical, extra care must be taken when computing the weights $w_{i,j,k}$ in (2.10), as the covariances $\mathbf{H}_{i,j}\mathbf{B}^b_{i,j}\mathbf{H}^T_{i,j} + \mathbf{R}_{i,k}$ are not necessarily equal. This means that, the normalization factors in each Gaussian term (similar to (2.11)),

$$\left|2\pi\left(\mathbf{H}_{i,j}\mathbf{B}^b_{i,j}\mathbf{H}^T_{i,j} + \mathbf{R}_{i,k}\right)\right|^{-\frac{1}{2}}, \tag{2.12}$$

are required to be computed. This can be performed in a computationally efficient manner through the use of the Cholesky decomposition and the log-sum-exp trick [7].

While the transformation of the distribution in (2.9) results in an estimate of the posterior distribution, the means $\bar{x}^a_{i,j}$ of this distribution are not actually samples from this distribution, thus it is not the case that the posterior samples are equivalent to these means,

$$x^a_{i,j} \neq \bar{x}^a_{i,j}. \tag{2.13}$$

A resampling procedure is therefore required in order to obtain independently and identically distributed (iid) samples from (2.9). What follows is one such procedure.

Procedure 2.3 (EnGMF resampling). Given the final posterior Gaussian mixture distribution in (2.9), it is possible to resample S samples from the posterior GMM through the following procedure:

1. for $s = 1, \ldots, S$, sample the random variable ℓ from the discrete distribution defined by the weights $\{w_{i,j,k}\}_{j,k}$,
2. sample $\mathbf{X}^a_{i,s}$ from the Gaussian $\mathcal{N}\left(x|\bar{x}^a_{i,\ell}, \mathbf{B}^a_{i,\ell}\right)$,

enabling samples to be generated from the posterior.

Remark 2.4 (Arbitrary Sampling of the Posterior). Note, that using Procedure 2.3 we are able to arbitrarily sample from the posterior distribution. This means that the number of posterior samples S could be significantly larger or significantly smaller than the original number of samples N used to generate said posterior.

Remark 2.5 (Independent and Identically Distributed Samples). Note that while we make the convenient assumption that the samples generated by Procedure 2.3 are iid, this is not actually the case. The parameters of each mode of the GMM are actually functions of the prior samples, and are themselves random variables, and thus introduce a dependence if two samples come from the same mode. This hidden dependence of the particles means that they are merely conditionally independent, making them exchangeable, but not independent in general.

Remark 2.6 (Prior Uniform Weights). If Procedure 2.3 is utilized to re-sample the particles at every step of the assimilation, then the prior distribution weights in (2.6) are all uniform $u_{i,j} = \frac{1}{N}$ under the assumption of a uniform transition density.

Remark 2.7 (Differences Between the EnGMF and the AGMF). Unlike the EnGMF, in the AGMF (see [42, 43]), instead of resampling like in Procedure 2.3, the weights are scaled by a defensive factor towards uniformity,

$$w_{i,j,k} = \alpha_i w_{i,j,k} + (1 - \alpha_i)\frac{1}{N}, \tag{2.14}$$

with a 'defensive factor', α_i, which ensures that the weights do not degenerate. The new particles are taken to be the means of the candidate posterior distribution (2.9). This idea is not explored in this work.

The posterior GMM (2.9) is only a good representation of the exact posterior if the prior GMM assumption (2.6) is a good approximation of the distribution that originated the particles $\mathbf{X}_i^b$. For a finite ensemble N it is possible that the EnGMF analysis is a poor representation of the truth. The prior GMM assumption (2.6), in the particle limit $N \to \infty$, converges to the exact prior under certain assumption on the covariances, $\mathbf{B}_{i,j}$, which is a known result from kernel density estimation literature [40]. We now show that under some assumptions on the covariances, $\mathbf{B}_{i,j}$, the posterior GMM (2.9) produced by the EnGMF procedure also converges, in the particle limit, $N \to \infty$, to a distribution obtained from performing exact Bayesian inference.

THEOREM 2.8 (EnGMF convergence). *Assuming the observation distribution is exact (2.8), if the means $\bar{x}_{i,j}^b$ in the estimated prior distribution GMM (2.6) are samples from the underlying exact distribution with weights $u_{i,j}$, and the prior Kernel covariance matrices tend to zero in the limit of ensemble size, $\lim_{N\to\infty} \mathbf{B}_{i,j}^b = 0$, then the EnGMF with the resampling procedure Procedure 2.3 converges to a filter in the class of sequential importance resampling (SIR) filters.*

Proof. Given the assumptions above, in the limit of ensemble size, $N \to \infty$, the prior distribution GMM converges to the empirical distribution,

$$p(x_i^b) = \frac{1}{N} \sum_{j=1}^{N} u_{i,j} \delta_{x_i^b - \bar{x}_{i,j}^b}, \tag{2.15}$$

which converges weakly to the underlying prior distribution. Then as the prior Kernel covariance tends towards zero, the posterior GMM estimate defined by (2.9) and (2.10) converges to the empirical measure,

$$p(x_i^a) = \frac{1}{N} \sum_{j=1}^{N} \left(\sum_{k=1}^{M} w_{i,j,k} \right) \delta_{x_i^a - \bar{x}_{i,j}^b}, \tag{2.16}$$

which converges weakly to the exact posterior distribution. Then, the EnGMF resampling in Procedure 2.3 makes the EnGMF converge to an SIR filter. ∎

2.2. EnGMF rate of convergence for scalar parameterization. We now motivate the importance of choosing a good parameterization of the covariance matrix is more cost effective than simply increasing the number of particles N. In the following, we show that the rate of convergence of the prior GMM estimate of the EnGMF is sub-linear under a scalar parameterization. This makes the choice of parameter the dominant factor that determines the goodness-of-fit of the distribution.

The most common paramterization of the prior GMM is a scalar parameterization that modifies the scaling of the covariance in a way that is guaranteed to degenerate in the limit of ensemble size. In the case of this scalar bandwidth parameterization, the prior covariance estimates become

$$\mathbf{B}_{i,j}^b(\beta_{i,N}^2) = \beta_{i,N}^2 \mathbf{P}_{i,N}^b, \quad 1 \le j \le N, \tag{2.17}$$

where $\mathbf{P}_i^b$ is the known (or approximated) covariance of the prior distribution, and β_i^2 is a scaling factor yet to be determined. Following the derivations in [40], we provide formulations of the error and optimal density for the covariance parameterization in (2.17).

Given the unknown true prior distribution $p_{x_i^b}$, and its GMM approximation $\widetilde{p}_{x_i^b}$, the most common metric for determining the convergence of the latter to the former is the mean integral squared error (MISE) at time index i,

$$\text{MISE}_i(p_{x_i^b}, \widetilde{p}_{x_i^b}) = \mathbb{E}_{\mathbf{X}_i^b}\left[\int_{\Omega_x}\left(p_{x_i^b}(x) - \widetilde{p}_{x_i^b}(x)\right)^2 \mathrm{d}x\right], \tag{2.18}$$

Observe that the convergence of the prior GMM estimate is not the same as convergence of the posterior GMM estimate. Nevertheless convergence of the prior estimate implies convergence of the posterior estimate. It is therefore the case that the rate of convergence of the prior estimate is directly related to the rate of convergence of the prior estimate.

The MISE is typically approximated using the dominant terms of its expansion into the approximated mean integrated square error (AMISE) given by,

$$\text{AMISE}(p_{x_i^b}, \widetilde{p}_{x_i^b}) = \frac{1}{4}\beta_i^4\alpha^2\gamma_i + N^{-1}\beta_i^{-n}\delta, \tag{2.19}$$

where for the GMM approximation of a distribution α and δ,

$$\alpha = 1, \quad \delta = \left(2\sqrt{\pi}\right)^{-n}, \tag{2.20}$$

are known constants that depend on the dimension n, and γ_i,

$$\gamma_i = \int_{\Omega_x} \text{tr}^2\left(\nabla_x^2 p_{x_i^b}(x)\right)\mathrm{d}x, \tag{2.21}$$

is dependent on the true prior distribution at time index i. As the true prior is unknown, a known reference distribution ϕ can be used to compute an approximation for (2.21),

$$\widetilde{\gamma}_i = \int_{\Omega_x} \text{tr}^2\left(\nabla_x^2 \phi(x)\right)\mathrm{d}x, \tag{2.22}$$

where ϕ is often taken to be the standard Gaussian distribution.

If our assumed estimate of the parameter $\widetilde{\gamma}_i$ is correct, the optimal bandwidth that minimizes the AMISE in (2.19) is,

$$\beta_i^2 = \left(\frac{\delta n}{\alpha^2\widetilde{\gamma}_i N}\right)^{\frac{2}{n+4}}, \tag{2.23}$$

by satisfying the first order optimality conditions of (2.19).

Plugging (2.23) back into (2.19), the error can now be written as,

$$\text{AMISE}(p_{x_i^b}, \widetilde{p}_{x_i^b}) = \underbrace{\frac{\delta}{4}\left(\frac{n\delta}{\alpha^2}\right)^{-\frac{n}{n+4}}}_{\text{const.}} \underbrace{\left(4\widetilde{\gamma}_i^{\frac{n}{n+4}} + n\gamma_i\widetilde{\gamma}_i^{-\frac{4}{n+4}}\right)}_{\text{reference mismatch}} \underbrace{N^{-\frac{4}{n+4}}}_{\text{conv. rate}} \tag{2.24}$$

where the first term is a constant and can be ignored, the second term is dependent on the mismatch between the true (2.21) and approximated (2.22) γ_i terms, and the third term determines the rate of convergence in N. Note that rate of coverenge is sublinear but close to linear for small state-space dimensions n, and is purely sub-linear for even a modestly small n.

As the rate of convergence in (2.24) is sub-linear, as determined by the third term, the multiplicative terms in front of the rate of converegence play a very dominant role. The first term is constant, and thus can be ignored. The second term,

$$4\widetilde{\gamma}_i^{\frac{n}{n+4}} + n\gamma_i\widetilde{\gamma}_i^{-\frac{4}{n+4}}, \tag{2.25}$$

therefore largely determines the error. It is trivial to see that (2.25) is minimized when $\widetilde{\gamma}_i = \gamma_i$, therefore the error is only minimized when the reference distribution ϕ in (2.22) matches the true distribution in (2.21).

Thus, when there is a large discrepancy between the reference distribution ϕ and the true distribution, an adaptive choice of the bandwidth parameter (2.23) could produce a much more significant decrease in error than simply increasing the ensemble size N.

3. Adaptive ensemble Gaussian Mixture Filter. Following the observations provided by Theorem 2.8 and by the discussion in subsection 2.2, we want to choose covariance matrices $\mathbf{B}_{i,j}$ found in the prior GMM assumption (2.6) in an intelligent and adaptive manner such that the convergence properties are satisfied. We additionally attempt to fulfill a desire useful to the practitioner: that practical convergence is achieved with as small as possible number of particles.

To that end, in this work we explore arbitrary parameterized covariance matrices in the prior GMM (2.6),

$$p(x_i|\theta_i) = \sum_{j=1}^{N} u_{i,j}\,\mathcal{N}\left(x_i|x_{i,j}^b, \mathbf{B}_{i,j}^b(\theta_i)\right), \tag{3.1}$$

where each covariance $\mathbf{B}_{i,j}^b(\theta_i)$ is a matrix function of some (small number of) parameters θ_i.

The aim of the parameterization in (3.1) is to find a set of parameters θ_i that can both be chosen adaptively at each step, and can ensure that the EnGMF does not violate the assumptions of Theorem 2.8 and, additionally, possibly attempts to minimize the error presented in subsection 2.2.

We now provide a way by which we can solve for the optimal parameters θ_i in (3.1) through the expectation maximization algorithm.

3.1. Expectation Maximization. The expectation maximization (EM) algorithm [8, 6] finds the set of the parameters θ_i that maximize $p(\theta_i|y_i)$ which is the conditional distribution of the parameters given the observations , at time index i.

Given some initial set of parameters $\theta_i^{(0)}$, the expectation maximization algorithm proceeds in an iterative fashion in two steps:

The *expectation step*,

$$\mathbb{E}_{x_i^b|y_i,\theta_i^{(m)}} \log p(x_i^b, y_i, \theta_i) \tag{3.2}$$

constructs the function representing the expectation of the joint distribution of the prior state, the observations, and the parameters. The joint distribution in (3.2) can be written in terms of the prior (2.6), observation (2.8), and parameter distributions as,

$$p(x_i^b, y_i, \theta_i) = p(y_i|x_i^b, \theta_i)p(x_i^b|\theta_i)p(\theta_i), \tag{3.3}$$

and, as the observation GMM (2.8) is not dependent on the parameters, (3.3) can be simplified to,

$$p(x_i^b, y_i, \theta_i) = p(y_i|x_i^b)p(x_i^b|\theta_i)p(\theta_i), \tag{3.4}$$

where the prior distribution (3.1) is parameterized in terms of its covariance (3.1), and the parameter distribution

$$p(\theta_i), \tag{3.5}$$

is determined on a case-by-case basis.

The *maximization step* aims to find the value of the parameters θ_i, that maximize the log joint distribution (3.4),

$$\theta_i^{(m+1)} = \arg\max_{\theta_i} \mathbb{E}_{x_i^b|y_i,\theta_i^{(m)}} \left[\log p(x_i^b|\theta_i) + \log p(\theta_i)\right], \tag{3.6}$$

where $\theta_i^{(m)}$ are the parameters from the previous step, $x_i^b|y_i, \theta_i^{(m)}$ are samples from the posterior distribution (2.9) given the previous set of parameters $\theta_i^{(m)}$, and the term $\log p(y_i|x_i^b)$ is constant and thus can be safely ignored due to the fact that it does not influence the optimization problem. Recall Remark 2.4 that in the EnGMF, it is possible to generate an unlimited number of i.i.d. samples from the posterior distribution, thus the maximization step (3.6) can be computed to an arbitrary level of accuracy, given some reasonable assumptions on the distribution of the sample mean.

Remark 3.1 (Invertible Covariances). Note that the prior covariance $p(x_i^b|\theta_i)$ in (3.6) requires that the covariance matrices $\mathbf{B}_{i,j}^b(\theta_i)$ in (3.1) are invertible, which is not necessarily required by the standard EnGMF.

Remark 3.2 (Meaningful Representation of the Prior). Note that the effect of the expectation maximization algorithm is to pick a parameterization of the prior estimate that best matches the posterior. For arbitrary parameterizations this would simply produce another copy of the posterior. The parameterization in (3.1) does not allow this to happen, as only the covariance is modified, and the mixture weights and means are not. This ensures that for (almost) all choices of the parameters θ, the prior estimate is still a useful representation of the prior. This means that the EM algorithm merely chooses the prior estimate that is most useful in subsequently representing the posterior.

3.1.1. Stochastic Optimization. The maximization step (3.6) requires the solution of a stochastic optimization problem. Much of the recent literature on stochastic optimization has been focused on machine learning applications [1]. As the number of parameters in θ_i is small, it is possible to take advantage of methods that are built for the small parameter size case and that differ from typical machine learning optimization methods. Thus, in this work we utilize a variant of Newton's method.

We can write the loss in the maximization step (3.6) as,

$$\mathcal{L}(x_i, \theta_i) = \mathbb{E}_{x_i^b|y_i,\theta_i^{(m)}} \left[\log p(x_i|\theta_i) + \log p(\theta_i)\right], \tag{3.7}$$

where the posterior can be written as the following,

$$x_i^{a,(m)} = x_i^b|y_i, \theta_i^{(m)}, \tag{3.8}$$

which is useful shorthand for the following derivations.

One algorithm for finding the maximum of the loss function (3.7) is Newton's method,

$$\theta_i^{(m+1,p+1)} = \theta^{(m+1,p)} + \alpha_m \left(\mathbb{E}_{x_i^{a,(m)}}[\nabla_\theta^2 \mathcal{L}(x, \theta_i^{(m+1,p)})] \right)^{-1} \mathbb{E}_{x_i^{a,(m)}}[\nabla_\theta \mathcal{L}(x, \theta_i^{(m+1,p)}))], \tag{3.9}$$

where α_m is the step-size (also known as the learning rate in the machine learning community), and the initial parameter value for the algorithm is $\theta_i^{(m+1,1)} := \theta_i^{(m)}$, which is the parameter from the previous maximization step (3.6). As it is challenging to compute the expected values in (3.9) analytically, some sort of approximation procedure is required.

As this is a stochastic optimization procedure, the two expected values in (3.9) cannot be calculated with the same random samples, as that would introduce unintended bias and variance into the update. In this work we utilize the sub-sampled version of Newton's method [39] built specifically to handle this scenario. In sub-sampled Newton's method, independent samples of $x_i^{a,(m)}$ are used to approximate the Hessian $\mathbb{E}_{x_i^{a,(m)}}[\nabla_\theta^2 \mathcal{L}(x, \theta_i^{(m+1,p)})]$ and the gradient $\mathbb{E}_{x_i^{a,(m)}}[\nabla_\theta \mathcal{L}(x, \theta_i^{(m+1,p)})]$. If the number of samples used is identical, then the number of samples required is double that of the stochastic gradient descent (SGD) algorithm which only requires the computation of $\mathbb{E}_{x_i^{a,(m)}}[\nabla_\theta \mathcal{L}(x, \theta^{(m+1,p)})]$. As Newton's method achieves faster convergence than SGD, it is the authors' belief that for this particular scenario the benefits of this approach outweight the additional costs.

Remark 3.3 (Quasi-Newton Methods). Instead of computing an estimate to the Hessian $\mathbb{E}_{x_i^{a,(m)}}[\nabla_\theta^2 \mathcal{L}(x, \theta_i^{(m+1,p)})]$ at every step, it is possible to only compute the Hessian at the initial step $\mathbb{E}_{x_i^{a,(m)}}[\nabla_\theta^2 \mathcal{L}(x, \theta_i^{(m+1,1)})]$ and use this approximation for all subsequent steps. This type of computationally efficient computation is a type of Quasi-Newton method [30] that is often used in practical applications.

Remark 3.4 (Alternative Optimization Algorithms). Alternative stochastic optimization algorithms could also be utilized. The classic stochastic gradient descent algorithm [38] is an alternative which would require a smaller step-size α_m. Another alternative is ADAM [23] which would require to keep track of separate momentum and velocity terms.

Remark 3.5 (Incremental Expectation Maximization). If the expectation maximization algorithm is performed online in sequential data assimilation, it is not necessary to perform many steps of either the expectation maximization algorithm, or sub-sampled Newton's method (3.9). In this work we initialize the parameters expectation maximization algorithm subsection 3.1 from the previous time step of the data assimilation algorithm. This can be weakly justified as a type of incremental expectation maximization [28], and in the authors' experience significantly increases the utility of the proposed approach.

We now discuss several different strategies for parameterizing the kernel covariance (3.1).

3.2. Bandwidth-based covariance. In kernel density estimation, choosing the optimal covariance matrices has had considerable research interest [40]. And, as discussed in section 2 has a considerable impact on the efficacy of the EnGMF algorithm.

From subsection 2.2, recall the covariance parameterization,

$$\mathbf{B}^b_{i,j}(\beta^2_{i,N}) = \beta^2_{i,N}\mathbf{P}^b_{i,N}, \quad 1 \leq j \leq N, \tag{3.10}$$

where $\beta^2_{i,N}$ is known as the bandwidth parameter [40]. This is one particular case of the covariance in the prior GMM (3.1) that is a focus of this work. The samples covariance in (3.10)

$$\mathbf{P}^b_{i,N} = \frac{1}{N-1}\mathbf{X}^b_i\left(I_N - \frac{1}{N}11^T\right)\mathbf{X}^{b,T}_i \approx \mathbb{E}\left[\left(x^b_i - \mathbb{E}[(x^b_i]\right)\left(x^b_i - \mathbb{E}[(x^b_i]\right)^T\right], \tag{3.11}$$

is known as the empirical covariance, and is an estimate of the covariance matrix of the prior state x^b_i. In (3.10), the only parameter is the bandwidth estimate, $\theta_i = \beta^2_{i,N}$.

Remark 3.6 (Stochastic Newton's for the Bandwidth Parameter). When the sub-sampled version of the stochastic Newton's method (3.9) is applied to the covariance parameterized by the bandwidth parameter (3.10), then both the stochastic estimate of the gradient and the stochastic estimate of the Hessian are scalars. This enables the computation of the maximization step (3.6) to be performed with minimal linear system solves.

The prior kernel covariance estimate in (3.10) takes advantage of the underlying covariance of the data, and is thus a type of online estimate however, the resulting accuracy of the density estimate is still highly dependent on the bandwidth parameter $\beta^2_{i,N}$.

It is known from [40] that if the underlying exact prior distribution of x^b in (2.6) is Gaussian, that the optimal choice of bandwidth parameter β^2 in (3.10) that minimizes the mean integrated square error is,

$$\beta^2_{i,N,\text{Gaussian}} = \left(\frac{4}{N(n+2)}\right)^{\frac{2}{n+4}}, \tag{3.12}$$

which is also known as Silverman's rule of thumb.

In practice most probability distributions of interest are not Gaussian, and (3.12) can result in a very poor approximation of the underlying density [40], thus a more refined choice of the bandwidth parameter is required.

Theorem 2.8 showed that a sufficient condition for the convergence of the EnGMF is that the covariance estimate tends towards zero as $N \to \infty$. We now show a condition on the bandwidth parameter that is sufficient for the EnGMF to converge.

LEMMA 3.7. *Given the sequence of parameters* $\{\beta^2_{i,N}\}_{N=1}^{\infty}$ *parameterized by the particle amount* N*, a sufficient condition for the covariance estimate* (3.10)*,*

$$\mathbf{B}_{i,N} = \beta^2_{i,N}\mathbf{P}^b_{i,N}, \tag{3.13}$$

to degenerate in the limit of particle number,

$$\mathbf{B}_{i,N} \xrightarrow{D} \delta_{\mathbf{0}}, \tag{3.14}$$

is that the bandwidth parameter tends towards zero,

$$\lim_{N\to\infty} \beta^2_{i,N} = 0, \tag{3.15}$$

in the limit of particle number N*.*

Proof. Observe that,

$$\begin{aligned}\mathbb{E}[\mathbf{B}_{i,N}] &= \beta_{i,N}^2 \mathbb{E}[\mathbf{P}_{i,N}^b], \\ \mathrm{Cov}[\mathbf{B}_{i,N}] &= \beta_{i,N}^4 \,\mathrm{Cov}[\mathbf{P}_{i,N}^b],\end{aligned} \tag{3.16}$$

meaning that both the mean and covariance of the random variable tend towards zero as $N \to \infty$, as required. ∎

COROLLARY 3.8. *The sequence* $\{\beta_{i,N,Gaussian}^2\}_{N=1}^{\infty}$ *of bandwidth parameters defined by Silverman's rule of thumb* (3.12) *satisfies the conditions of Lemma* 3.7.

Lemma 3.7 showed that when $\beta_{i,N}^2$ is a constant, and converges to zero in the limit of particle number $N \to \infty$, the EnGMF converges. However, as we have uncertainty about the bandwidth parameter, it is natural to think about it as a random variable with some distribution. Thus an important choice is that of the distribution of the bandwidth parameter. Care must be taken to ensure that this choice is sufficient to make the resulting algorithm converge.

We provide a sufficient condition on the distribution of the bandwidth parameter $\beta_{i,N}^2$ from (3.10), as an extension of Theorem 2.8. We therefore extend Lemma 3.7 to bandwidth parameters that are random variables with some prior distribution in the expectation maximization algorithm.

THEOREM 3.9. *Given the sequence of random variables* $\{\beta_{i,N}^2\}_{N=1}^{\infty}$ *with a sequence of distributions* $\{p(\beta_{i,N}^2)\}_{N=1}^{\infty}$ *parameterized by the particle amount* N*, a sufficient condition for the covariance estimate* (3.10)*,*

$$\mathbf{B}_N = \beta_{i,N}^2 \mathbf{P}_N^b, \tag{3.17}$$

to tend towards zero in distribution in the limit of particle number,

$$\mathbf{B}_N \xrightarrow{D} \mathbf{0}, \tag{3.18}$$

is that the distribution of the bandwidth parameter tends towards the delta distribution around zero,

$$\lim_{N\to\infty} p(\beta_{i,N}^2) = \delta_0, \tag{3.19}$$

ensuring that $\beta_{i,N}^2$ *almost surely becomes* 0*.*

Proof. If the distribution of $\beta_{i,N}^2$ converges to δ_0, then the solution to the maximization step in the EM algorithm (3.6) almost surely becomes a constant, namely that $\beta_{i,N}^2 \xrightarrow{\text{a.s.}} 0$, as required. ∎

3.2.1. Choosing the bandwidth distribution. One way in which the conditions of Theorem 3.9 could be satisfied is through an intelligent choice of the probability distribution of the bandwidth parameter $p(\beta_{i,N}^2)$.

A common choice in the literature, the principal of maximum entropy (PME) [22] could be used to find a good candidate for this distribution. If we assume that the expected value of the bandwidth, $\beta_{i,N}^2$, is Silverman's rule of thumb (3.12), and we have no other information available, then the distribution that satisfies the PME is the exponential distribution,

$$p(\beta_{i,N}^2) = \beta_{i,N,\mathrm{Gaussian}}^{-2} e^{-\beta_{i,N,\mathrm{Gaussian}}^{-2} \beta_{i,N}^2}, \tag{3.20}$$

this distribution, however, always has a single mode at zero, thus, from the authors' experience, is ill-suited for use in expectation maximization.

It is possible to perform some slight-of-hand in order to make this assumption more tractable. It is more efficient to look at $\beta_{i,N}$, the square-root of the bandwidth parameter. If (3.20) is the distribution of $\beta_{i,N}^2$, then $\beta_{i,N}$ is distributed according to

$$p(\beta_{i,N}) = 2\beta_{i,N,\text{Gaussian}}^{-2}\beta_{i,N}e^{-\beta_{i,N,\text{Gaussian}}^{-2}\beta_{i,N}^2} \tag{3.21}$$

which is the Rayleigh distribution [31] with known mode $2^{-1/2}\beta_{i,N,\text{Gaussian}}$. We assume this Rayleigh distribution (3.21) on the bandwidth parameter for the remainder of this paper.

It is also possible to assume a more general distribution around $\beta_{i,N}^2$, such as a gamma distribution, though this choice would introduce another free parameter into the algorithm; an undesirable outcome.

While the parameterized covariance in (3.10) is well-studied, it has a few limitations that prevent it from being used in high-dimensional inference, chief among those being the fact that the covariance estimate in (3.11) can potentially be low-rank, and thus generate a covariance that is not invertible Remark 3.1, thus we can introduce covariance matrix estimates that have extra parameters in order to mitigate this issue.

Remark 3.10. It is important to note that the optimal bandwidth is deterministic, but unknown. The uncertainty that transforms our knowledge about the bandwidth into a random variable is purely from the point of view of the agent performing the inference. For a more in-depth discussion about choosing distributions for parameters see [22].

3.3. Covariance Shrinkage Estimates. Covariance shrinkage [13, 12, 14, 11, 24] aims to use extra prior information about the covariance of x_i^b in (2.6) in order to have a more accurate covariance estimate in the case when the number of samples is smaller than the dimension of the dynamical system $N < n$. Covariance shrinkage methods have previously been employed for ensemble data assimilation [29, 34] and for regularization in particle filtering [35].

Assume that we have prior information about the covariance structure of x_i^b in the form of a 'target' covariance matrix $\mathbf{T}_i$. The covariance shrinkage estimate of the covariance, scaled by the bandwidth, is given by,

$$\mathbf{B}_{i,j}^b = \beta_{i,N}^2\left[\gamma_i\mu_i\mathbf{T}_i + (1-\gamma_i)\mathbf{P}_i^b\right] \tag{3.22}$$

where,

$$\mu_i = n^{-1}\operatorname{tr}\mathbf{C}_i, \quad \mathbf{C}_i = \mathbf{T}_i^{-\frac{1}{2}}\mathbf{P}_i^b\mathbf{T}_i^{-\frac{1}{2}} \tag{3.23}$$

is a rescaling factor, and γ_i is the shrinkage factor, which we treat as a parameter.

Under Gaussian assumptions on the samples, $x_{i,j}^b$, a good known shrinkage factor is,

$$\begin{aligned}\gamma_{i,\text{RBLW}} &= \min\left[\frac{N-2}{N(N+2)} + \frac{(n+1)N-2}{N(N+2)(n-1)\hat{U}_i}, 1\right], \\ \hat{U}_i &= \frac{1}{n-1}\left(\frac{n\operatorname{tr}\mathbf{C}_i^2}{\operatorname{tr}^2\mathbf{C}_i} - 1\right)\end{aligned} \tag{3.24}$$

called the Rao-Blackwell Ledoit-Wolf estimator [13].

One possible choice of the target matrix $\mathbf{T}_i$ that does not require any prior knowledge is the diagonal of the empirical covariance (3.11),

$$\mathbf{T}_i = \boxslash \mathbf{P}_i^b, \tag{3.25}$$

where the notation of $\boxslash$ is introduced to signify the matrix consisting of only the diagonal of the subsequent term. Observe that,

$$\operatorname{tr}\left[\left(\boxslash \mathbf{P}_i^b\right)^{-\frac{1}{2}} \mathbf{P}_i^b \left(\boxslash \mathbf{P}_i^b\right)^{-\frac{1}{2}}\right] = n, \tag{3.26}$$

meaning that when the target matrix is defined by (3.25), the scaling factor $\mu_i = 1$. This means that direct calculation the matrix $\mathbf{C}_i$, from (3.23), is only required for the calculation of $\gamma_{i,\mathrm{RBLW}}$ in (3.24).

Remark 3.11 (On $p(\gamma_i)$). A commonly made assumption is that parameters are independently distributed, therefore the distribution of $p(\beta^2_{i,N})$ can be chosen independently of the distribution $p(\gamma_i)$. As there are no requirements that the optimal γ_i is dependent on ensemble size, it is a natural choice to assume a uniform likelihood,

$$p(\gamma_i) \propto 1, \tag{3.27}$$

which is a typical assumption in parameter estimation [8].

3.4. Covariance Localization. In the geosciences, states usually have some sort of innate spatial structure. State variables that are spatially far apart are generally more weakly correlated than states that are closer together. Taking advantage of this fact, covariance localization [5] is a matrix tapering technique which aims to reduce the impact of spurious correlations caused by undersampled ($N \ll n$) covariance matrix estimates.

In this work we focus on what is known as the B-localization methodology, and combine it with the bandwidth scaling (3.10) in the following manner,

$$\mathbf{B}_{i,j}^b = \beta_i^2 \left(\rho(r_i) \circ \mathbf{P}_i^b\right), \tag{3.28}$$

where the matrix $\rho(r_i)$ contains a set of decorrelation variables parameterized by the localization radius r_i, and $\circ$ is the element-wise Schur product.

A common choice for ρ is known as Gaussian localization,

$$\rho(r_i)_{\ell,q} = e^{-\frac{1}{2}\frac{d(\ell,q)^2}{r_i^2}}, \tag{3.29}$$

where $d(\ell, q)$ represents the spatial distance between the variables at index ℓ and index q.

Remark 3.12 (Choice of Localization Radius r_i). The choice of localization radius r_i in (3.28) can be informed by the temporal covariance of the model of interest if the model of interest is Ergodic, however in practice, the best localization radius is almost always determined empirically.

Adaptive-in-time choices for r_i have been explored for the ensemble Kalman filter in [33].

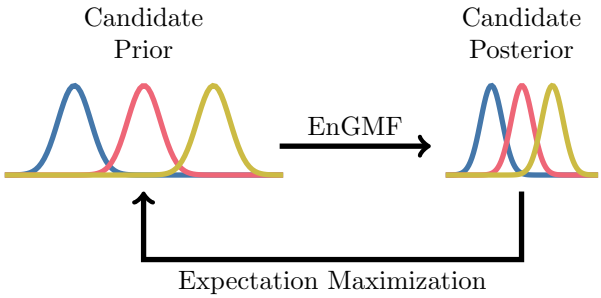

FIG. 1. *An illustration of the outer loop of the AEnGMF algorithm. In this example the candidate prior is made up of three distinct Gaussian modes which are plotted separately. The candidate prior is transformed into the candidate posterior through the standard EnGMF update (2.10). Next the expectation maximization algorithm (3.2) and (3.6) is performed, and a new candidate prior is obtained. This procedure repeats until some desired level of convergence is achieved.*

3.5. Practical Implementation of the AEnGMF. We are now able to combine all the elements presented in this section to fully describe the inner-workings of the adaptive ensemble Gaussian mixture filter (AEnGMF). The AEnGMF operates as follows. First a choice of parameterized covariance is made by the user. This choice determines the parameters that are optimized for. Next a choice of parameter distribution is required. In this work the bandwidth parameter is assumed to be distributed according to the Rayleigh distribution subsection 3.2.1, and the rest of the parameters are assumed to be proportional to one, thus of no additional consequence.

At each step of the algorithm, the previous choice of covariance parameters is carried over, $\theta_i^{(1,1)} := \theta_{i-1}$. This choice from Remark 3.5 is motivated by incremental approaches to expectation maximization, and lends itself particularly well to parameterized covariances that do not depend on their parameters changing a lot from step to step.

Next, M iterations of the expectation-maximization algorithm are performed. The expectation maximization algorithm can be treated as the 'outer-loop' [30] in this optimization procedure.

The cost function is solved using P loops of sub-sampled Newton's method (3.9) with a constant learning-rate α making this the 'inner-loop' algorithm. As it is possible to sample from the posterior arbitrarily Remark 2.4, the gradient and Hessian calculations can be performed using a different number of samples, S, than that of the number of particles N. Specifically, the gradient is computed using S samples from the candidate posterior, and the Hessian is computed using S separate samples from the candidate posterior, for a total of $2S$ samples.

As the AEnGMF is a particle filter, resampling of N particles is performed at the end of the algorithm with the EnGMF resampling procedure Procedure 2.3.

The outer loop of the algorithm is illustrated in Figure 1, and a detailed step-by-step look at the algorithm can be seen in Algorithm 3.1.

Remark 3.13 (Choosing M, P, α, and S). In the authors' experience, it is much more advantageous to perform multiple iterations of the expectation maximization algorithm than that of sub-sampled Newton's method, thus it is advantageous to take $M \geq P$. It is also advantageous to oversample the gradient and Hessian, thus $S \geq N$. By far the hardest choice to make is that of the learning-rate α. A learning rate that is too large ($\alpha \approx 1$) could cause the algorithm to become unstable, thus

Algorithm 3.1 The Adaptive Ensemble Gaussian Mixture Filter

Input Initial ensemble $\mathbf{X}_0^a$, initial estimate for the parameters θ_0, outer loop iteration count M, inner loop iteration count P, learning rate α, and number of internal samples S.

for $i = 1, \ldots$ **do**
  % Propagate the ensemble forward in time through the model
  $\mathbf{X}_i^b = \mathcal{M}(\mathbf{X}_{i-1}^a)$
  % Initialize the parameters θ to the parameters from the previous step
  $\theta_i^{(1,1)} \coloneqq \theta_{i-1}$
  % Perform the expectation maximization loop M times
  for $m = 1, \ldots, M$ **do**
    % Construct loss function
    $\mathcal{L}(x, \theta) \coloneqq \mathbb{E}_{x^b|y,\theta_i^{(m)}}[\log p(x|\theta) + \log p(\theta)]$
    % Initialize the inner loop θ parameter
    $\theta_i^{(m+1,1)} \coloneqq \theta_i^{(m,P+1)}$
    % Perform P steps of subsampled Newton's method.
    for $p = 1, \ldots, P$ **do**
      % Sample S particles from the candidate posterior.
      $\mathbf{X}^a \sim_{(S)} \pi(x|y, \theta_i^{(m)})$
      % Compute the loss gradient
      $g \coloneqq \mathbb{E}_{\mathbf{X}^a}[\nabla_\theta \mathcal{L}(X, \theta_i^{(m+1,p)})]$
      % Similarly, compute sample Hessian using different samples
      $\mathbf{X}^a \sim_{(S)} \pi(x|y, \theta_i^{(m)})$
      $H \coloneqq \mathbb{E}_{\mathbf{X}^a}[\nabla_\theta^2 \mathcal{L}(X, \theta_i^{(m+1,p)})]$
      % Compute new estimate of the parameters
      $\theta_i^{(m+1,p+1)} \coloneqq \theta_i + \alpha H^{-1} g$
    end for
    % Set the current θ parameter
    $\theta_i^{(m+1)} \coloneqq \theta_i^{(m+1,P+1)}$
  end for
  % Set the θ parameter for the current time index
  $\theta_i \coloneqq \theta_i^{(M+1)}$
  % Sample a new ensemble of N particles with new parameters θ_i
  $\mathbf{X}_i^a \sim_{(N)} \pi(x|y, \theta_i)$
end for

increasing the cost instead of decreasing it, and thus make it choose parameters θ that are worse than the original choices. A learning rate that is too small ($\alpha \to 0$) could lead to parameters that react poorly to the changing conditions of the states. From the practitioner's point of view, this is by far the most important parameter to choose correctly. Ideally, the parameter α can be chosen through some type of line search technique [30] that ensures that steps are always taken in a direction that decreases the error, though this is not explored in this work.

Remark 3.14 (Considerations for the High-dimensional Setting). There are many considerations to be made for getting the AEnGMF to work in the high-dimensional setting. First is that the computation of the covariance cannot be made explicitly.

This can be resolved by utilizing a covariance estimate that does not have to be explicitly computed like that of the shrinkage estimate in (3.22). Matrix inverse vector products of (3.22) can also be computed without explicit computation of the entire matrix. The normalizing factor issue in Remark 2.2 can also be mitigated by this covariance matrix estimate. Another problem is the resampling procedure in Procedure 2.3, which requires the computation of matrix square root vector products. Methods such as those proposed in [2, 19, 15] could be utilized to solve this issue, though this is still an open problem. A final consideration is that the intermediate results or approximations of the densities could represent non-physical states, thus special consideration must be given to those problems, especially in the high-dimensional setting.

Remark 3.15 (Computational complexity of the AEnGMF). Take n to be the dimension of the full state, m to be the dimension of the observations, q to be the dimensions of the parameters θ, and N to be the ensemble size. The dominant terms of the EnGMF update involve constructing the covariance, computing the gain matrix and updating the covariance, which in the worst case has computational complexity $\mathcal{O}(m^3N + m^2nN + n^2N)$. The resampling procedure Procedure 2.3 has complexity $\mathcal{O}(n^3N)$. For the AEnGMF, the cost of computing the Hessian in the worst case, where the gradients need to be computed by repeat evaluation of the cost function, the complexity becomes $\mathcal{O}(q^2n^3N)$. Thus, the total complexity of the AEnGMF is

$$\mathcal{O}\left(MP[C_{\text{EnGMF}} + q^2n^3N] + C_{\text{EnGMF}}\right), \tag{3.30}$$

where C_{EnGMF} is the complexity of the EnGMF. Thus, the cost of the algorithm has to be weighted against the cost of propagating more particles through the forward model dynamics.

4. Numerical Experiments.

The aim of the numerical experiments is first to demonstrate the viability and convergence of the proposed AEnGMF on small-scale problem, and secondly to demonstrate the more complicated covariance parameterization approaches on a larger-scale problem.

4.1. Lorenz '63.

With the first set of experiments we aim to look at a highly non-linear system with a non-linear observation operator. We focus on the standard EnGMF case of the Kernel covariance parameterized by the bandwidth parameter (3.10).

We take the the 3-variable Lorenz '63 equations [26],

$$\begin{aligned} x' &= \sigma(y - x), \\ y' &= x(\rho - z), \\ z' &= xy - \beta z, \end{aligned} \tag{4.1}$$

with canonical parameters $\sigma = 10$, $\rho = 28$, and $\beta = \frac{8}{3}$. The time between assimilations is taken to be $\Delta t = 0.5$, which allows the system enough time to evolve in a highly non-linear manner. The non-linear dynamics are propagated through time with an adaptive Runge-Kutta method [16] with absolute and relative tolerances of 10^{-11} in order to simulate a costly forward model calculation.

It is known [20] that the system (4.1) has three critical points, one at the origin,

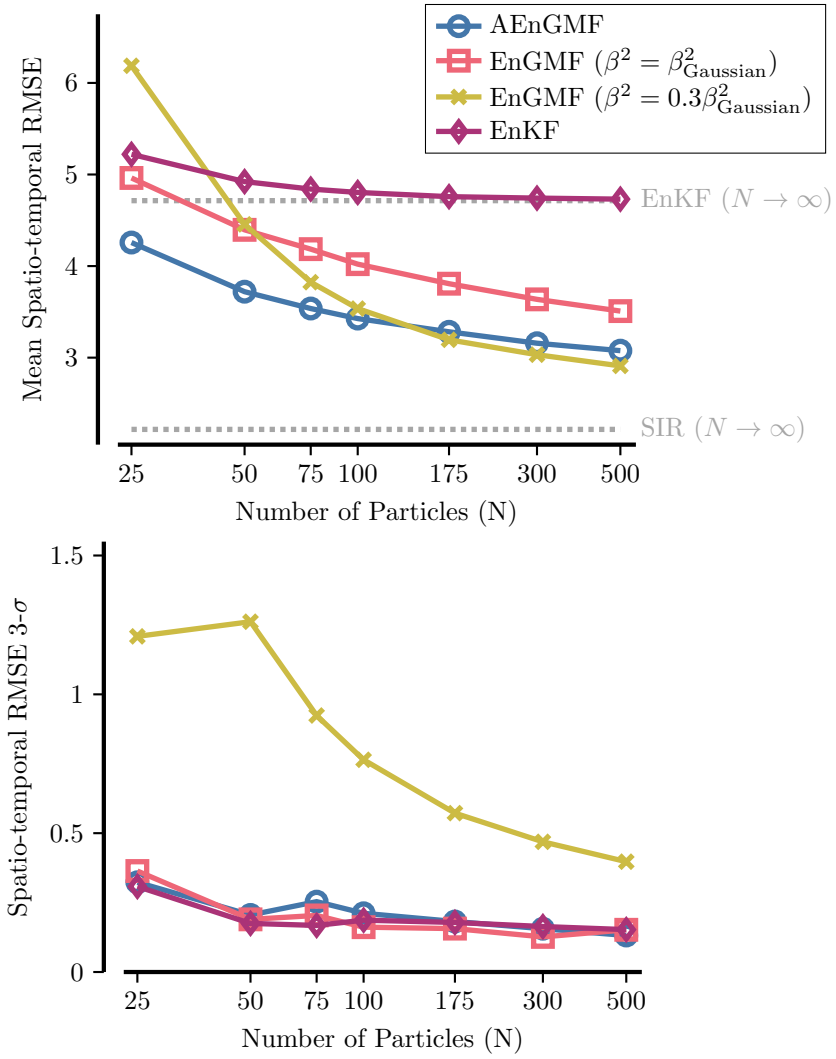

FIG. 2. *RMSE simulation results for the Lorenz '63 equations for four different data assimilation algorithms. The top plot represents the mean RMSE value across all runs, while the bottom plot represents three standard deviations of error around the mean. The blue line with circular marks represents the AEnGMF, the red line with square marks represents the canonical EnGMF with Silverman's rule of thumb, the yellow line with x marks represents the EnGMF with a scaled Silverman's rule of thumb, and the Raspberry line with diamond marks represents the EnKF. Two baseline lines, running the EnKF and a particle filter (SIR) for a high particle number are also provided to provide theoretical bounds.*

and in the center of each of the butterfly wings. The first non-zero critical point,

$$\begin{aligned} x_c &= \sqrt{\beta(\rho-1)}, \\ y_c &= \sqrt{\beta(\rho-1)}, \\ z_c &= \rho - 1, \end{aligned} \tag{4.2}$$

defines the center of one of the wings of the butterfly, with the other center being $(-x_c, -y_c, z_c)$ and the origin being $(0, 0, 0)$. For the non-linear observation operator

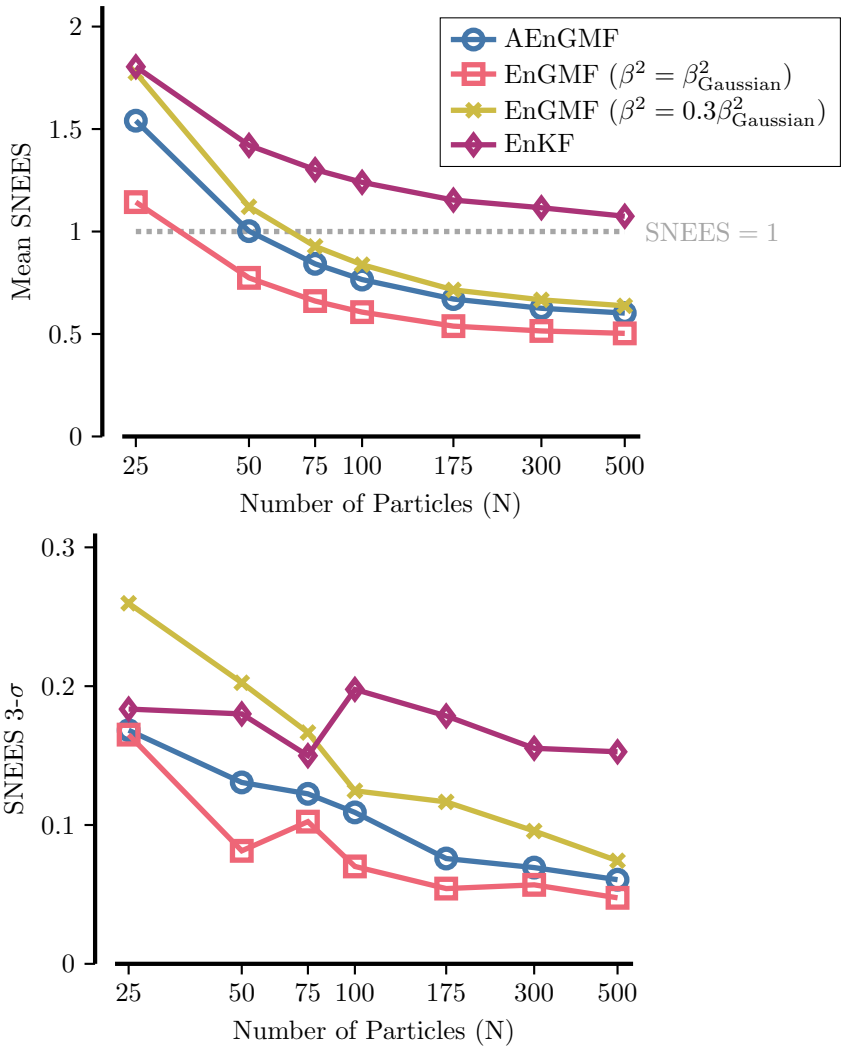

FIG. 3. *SNEES simulation results for the Lorenz '63 equations for four different data assimilation algorithms. The top plot represents the mean SNEES value across all runs, while the bottom plot represents three standard deviations of error around the mean. The lines representing the algorithms are identical to those in Figure 2. The dashes lines around the main lines represent three standard deviations over the samples. The ideal SNEES of one is represented by the constant gray dashed line.*

we measure the distance from the critical point (4.2) to the point being measured,

$$\mathcal{H}(x, y, z) = \sqrt{(x - x_c)^2 + (y - y_c)^2 + (z - z_c)^2}, \tag{4.3}$$

as a scalar observation, with Gaussian error with an error variance of $\mathbf{R} = 1$.

The goal of this experiment is to show that the various variants of the EnGMF are superior to the ensemble Kalman filter (EnKF) and converge to exact Bayesian inference in the limit of particle number ($N \to \infty$). We therefore calculate two reference boundaries for this problem, one using the EnKF for a large ensemble size (N = 1000) and for the sequential importance resampling (SIR) particle filter with a

large number of particles (N = 1000), specifically the variant found in [37].

It is known in the literature [40, 21] that Silverman's rule-of-thumb is usually an over-estimate of the optimal bandwidth term. We thus attempt to find a scaling factor $0 < s < 1$ such that the bandwidth parameter defined by the product,

$$\beta^2_{i,N} = s\beta^2_{\text{Gaussian}}, \tag{4.4}$$

would produce the minimal error for our choice of number of particles. From a quick parameter sweep, it was determined that $s = 0.3$ provides a good factor, that is optimal for a high number of particles.

Thus, we run four different algorithms, the EnKF, the AEnGMF, the EnGMF with Silverman's rule of thumb, and the EnGMF with Silverman's rule of thumb scaled by $s = 0.3$, for this model setup for various numbers of particles ranging from $N = 25$ to $N = 500$.

All experiments were run for 48 independent initial ensembles, with the same truth but different observations, for 5500 assimilation steps with the first 500 discarded for spinup, meaning that the first 500 steps do not count into the error calculations to let the filter reach a steady state. The mean of the spatio-temporal RMSE,

$$\text{RMSE}(\bar{x}, x^t) = \sqrt{\frac{1}{nT}\sum_{i=1}^{T}\sum_{l=1}^{n}\left(\bar{x}_{i,l} - x^t_{i,l}\right)^2}, \tag{4.5}$$

is calculated between the statistical mean $\bar{x}$ and the truth x^t, over the 12 runs and is the metric by which the efficacy of the algorithms is determined. In order to check the consistency, the mean of the scaled normalized estimated error squared (SNEES) metric [44] is utilized,

$$\text{SNEES}(\bar{x}, x^t) = \frac{1}{nT}\sum_{i=1}^{T}\left(\bar{x}_i - x^t_i\right)^T \mathbf{P}^{a,-1}_i \left(\bar{x}_i - x^t_i\right), \tag{4.6}$$

where $\mathbf{P}^a_i$ is the estimate of the posterior covariance at time index i. A SNEES of one is considered to be ideal, as that means that the error predicted by the filter is in line with the actual error of the filter. Additionally, if the SNEES is not one, it is better for the filter to be more conservative, meaning the SNEES is less than one, that overconfident, meaning a SNEES greater than one.

For the choices of parameters in Algorithm 3.1, we choose $M = 5$ loops of the expectation maximization algorithm, $P = 1$ loops of sub-sampled Newton's method, sampling $S = N$ exactly as many samples as there are particles, and a high learning rate of $\alpha = 1$. The previous parameter choices were hand-tuned to balance error and time to solution. The Rayleigh distribution with mean of Silverman's rule-of-thumb is chosen for the bandwidth parameter just like in subsection 3.2.1.

The results of the RMSE experiments are visually demonstrated in Figure 2. As can be seen, the AEnGMF is consistently lower in error than the EnGMF with bandwidth equivalent to Silverman's rule-of-thumb, and provides lower error in the particle number range of $N = 25$ to $N = 100$. The EnGMF with scaling factor $s = 0.3$ is the only algorithm to perform worse than the EnKF for $N = 25$, but also produces the lowest error between $N = 300$ and $N = 500$. Crucially, almost all algorithms have the same error bounds, as shown by the 3-σ plot of the RMSE, except for the The EnGMF with scaling factor $s = 0.3$, which has a significantly higher error standard deviation. This means that the AEnGMF lowers the error without sacrificing stability.

For the SNEES of the experiments, demonstrated in Figure 3, the AEnGMF lies between the EnGMF and the EnGMF with scaling facotr $s = 0.3$, both in terms of raw SNEES value and in terms of standard deviation. As both algorithms are conservative for higher particle numbers, this is not a particularly surprising result.

It can be seen that the EnGMF family of methods converges to the SIR limit very slowly, driven by the sublinear rate of convergence. The practical use of the Rayleigh distribution subsection 3.2.1 for achieving this effect can be questioned as the convergence is clearly sub-optimal. A better choice of the distribution of the bandwidth parameter is required.

Finally, we report the practical computational time increase of the AEnGMF over the EnGMF. For the full forecast-analysis loop a 1.2 to 1.7 times increase in computational cost was observed, meaning that when the computational time is dominated by the forward model runs, the cost of the AEnGMF is not an overly significant computational burden.

4.2. Lorenz '96. For the next set of experiments, we look at the case of an undersampled ($N \ll n$) estimate of the prior distribution. We look at two different types of covariance matrix parameterizations: one based on covariance shrinkage (3.22) with the target matrix being the diagonal of the statistical covariance (3.25), and a localization-based covariance matrix estimate (3.28).

For the model of interest we take the 40-variable Lorenz '96 equations [27],

$$x'_k = -x_{k-1}(x_{k-2} - x_{k+1}) - x_k + F \ldots, \quad k = 1, \ldots, 40, \tag{4.7}$$

with cyclic boundary conditions, and the forcing factor $F = 8$. For the time between assimilations we take one day of model time which is equivalent to a $\Delta t = 0.2$, leading to a high level of non-linearity in the system. The non-linear dynamics are propagated through time with an adaptive Runge-Kutta method [16] with absolute and relative tolerances of 10^{-6}.

We want to compare the AEnGMF approach of adaptively choosing the parameters of the Kernel covariance with that of the more classic EnGMF approach where the parameters are determined by a known good heuristic. We also want to compare with a base-line state-of-the-art algorithm, the localized ensemble Kalman filter. To that end, we perform experiments on the following set of filters:

1. the shrinkage-based AEnGMF (Shr-AEnGMF), with parameters of β_i^2 for the bandwidth and $\zeta_i = \tanh^{-1} \gamma_i$, for an unbounded transformation of the shrinkage parameter $0 < \gamma_i < 1$,
2. the shrinkage-based EnGMF (Shr-EnGMF) with bandwidth defined by Silverman's rule-of-thumb (3.12), $\beta_i^2 = \beta_{i,\text{Gaussian}}^2$ and the RBLW (3.24) shrinkage parameter $\gamma_i = \gamma_{i,\text{RBLW}}$,
3. the localized AEnGMF (LAEnGMF) with parameters of β_i^2 for the bandwidth and $\zeta_i = \sqrt{r_i}$ for an unbounded transformation of the localization radius $0 < r_i$,
4. the localized EnGMF (LEnGMF) with bandwidth defined by Silverman's rule-of-thumb (3.12), $\beta_i^2 = \beta_{i,\text{Gaussian}}^2$ and a fixed radius of $r_i = 4$,
5. and the localized EnKF (LEnKF) with fixed radius $r_i = 4$ for a useful comparison with a state-of-the-art filter.

For the non-linear observation operator, we take the point-wise non-linear operator,

$$\mathcal{H}(x_i) = \frac{x_i}{2}\left[1 + \left(\frac{|x_i|}{10}\right)^{\omega - 1}\right], \tag{4.8}$$

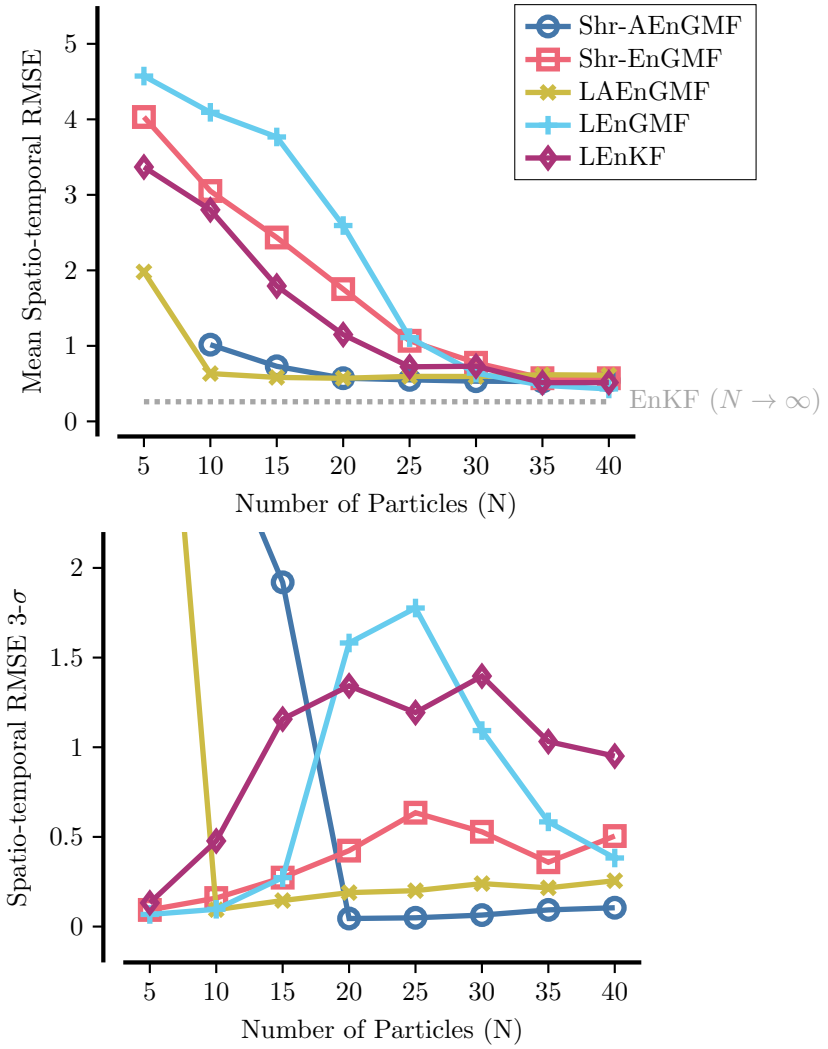

FIG. 4. *Simulation results for the Lorenz '96 equations for five different data assimilation algorithms. The dark blue line with circle marks represents the AEnGMF with a shrinkage-based estimate to the covariance , with the light red line with square represents the standard EnGMF with a shrinkage-based estimate to the covariance, the yellow line with x marks represents a localized AEnGMF, the light-blue line with plus marks represents a localized EnGMF, and the raspberry line with diamond marks represnts the localized EnKF.*

as found in [5], with $\omega = 5$ for a medium level of non-linearity, with the observation covariance matrix being set to $\mathbf{R} = \frac{1}{4}\mathbf{I}_{40}$. The number of particles is taken to range from as little as $N = 5$ to as high as $N = 40$.

All experiments were run for 48 independent initial ensembles, with the same truth but different observations, for 5500 assimilation steps with the first 500 discarded for spinup, meaning that the first 500 steps do not count into the error calculations to let the filter reach a steady state. For our error metric we again take the spatio-temporal RMSE (4.5).

For the choices of parameters in Algorithm 3.1, we choose $M = 1$ loops of the

expectation maximization algorithm, $P = 1$ loops of sub-sampled Newton's method, sampling $S = 100$ to sample in the excess, and a low learning rate of $\alpha = 1e-2$. The Rayleigh distribution with mean of Silverman's rule-of-thumb is chosen for the bandwidth parameter just like in subsection 3.2.1, with both the radius r and shrinkage parameter γ having distributions proportional to one along all of their support as they are not required for convergence.

The results for this round of experiments can be seen in figure Figure 4. At around $N = 30$ particles, all algorithms perform roughly the same, thus the interesting behavior occurs when there are fewer particles. Both versions of the EnGMF without adaptive covariance estimates (Shr-EnGMF and LEnGMF) perform worse than the localized EnKF. The adaptive versions of the same algorithms (Shr-AEnGMF and LAEnGMF) perform significantly better than all other tested algorithms. The the LAEnGMF practically converges for $N = 10$ particles, and the Shr-AEnGMF practically converges for $N = 20$ particles. Additionally the Shr-AEnGMF and LAEnGMF have tighter error bounds than all the other tested algorithms, potentially signifying that the adapative nature of the algorithm is better able to handle outlier scenarios. These results highlight the need and utility of the adaptive covariance estimate approach in the EnGMF presented in this paper.

5. Conclusions. By leveraging parameterized sample covariance estimates and the expectation maximization algorithm, this work introduced the adaptive ensemble Gaussian mixture filter (AEnGMF) as an extension of the ensemble Gaussian mixture filter (EnGMF). Theoretical results about the convergence properties of this filter were derived by making assumptions about the distribution of the kernel bandwidth. Numerical results have verified the theoretical convergence properties of the AEnGMF, and have shown that for a certain set of parameters the AEnGMF has superior convergence to that of the EnGMF.

Future work could extend the AEnGMF to a smoothing [37] framework, a hybrid filtering [18] framework, and to a multifidelity filtering [32] framework. An active research direction is in applying the AEnGMF to a real-world orbit tracking problem [44]. Work exploring practical consideration on choosing the parameters discussed in Remark 3.13 is also of independent interest.

Acknowledgments. The authors would like to thank the two anonymous reviewers for their very detailed and thorough comments that have helped significantly increase the quality of this work.

REFERENCES

[1] C. C. Aggarwal et al., *Neural networks and deep learning*, vol. 10, Springer, 2018.
[2] E. Allen, J. Baglama, and S. Boyd, *Numerical approximation of the product of the square root of a matrix with a vector*, Linear Algebra and its Applications, 310 (2000), pp. 167–181.
[3] B. D. Anderson and J. B. Moore, *Optimal filtering*, Courier Corporation, 2012.
[4] J. L. Anderson and S. L. Anderson, *A monte carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts*, Monthly weather review, 127 (1999), pp. 2741–2758.
[5] M. Asch, M. Bocquet, and M. Nodet, *Data assimilation: methods, algorithms, and applications*, SIAM, 2016.
[6] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4, Springer, 2006.
[7] P. Blanchard, D. J. Higham, and N. J. Higham, *Accurately computing the log-sum-exp and softmax functions*, IMA Journal of Numerical Analysis, 41 (2021), pp. 2311–2330.
[8] M. Bocquet, J. Brajard, A. Carrassi, and L. Bertino, *Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization*,

arXiv preprint arXiv:2001.06270, (2020).
[9] L. Breiman, W. Meisel, and E. Purcell, *Variable kernel estimates of multivariate densities*, Technometrics, 19 (1977), pp. 135–144.
[10] G. Burgers, P. J. Van Leeuwen, and G. Evensen, *Analysis scheme in the ensemble Kalman filter*, Monthly weather review, 126 (1998), pp. 1719–1724.
[11] X. Chen, Z. J. Wang, and M. J. McKeown, *Shrinkage-to-tapering estimation of large covariance matrices*, IEEE Transactions on Signal Processing, 60 (2012), pp. 5640–5656.
[12] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, *Shrinkage algorithms for mmse covariance estimation*, IEEE Transactions on Signal Processing, 58 (2010), pp. 5016–5029.
[13] Y. Chen, A. Wiesel, and A. O. Hero, *Shrinkage estimation of high dimensional covariance matrices*, in 2009 ieee international conference on acoustics, speech and signal processing, IEEE, 2009, pp. 2937–2940.
[14] Y. Chen, A. Wiesel, and A. O. Hero, *Robust shrinkage estimation of high-dimensional covariance matrices*, IEEE Transactions on Signal Processing, 59 (2011), pp. 4097–4107.
[15] E. Chow and Y. Saad, *Preconditioned Krylov subspace methods for sampling multivariate gaussian distributions*, SIAM Journal on Scientific Computing, 36 (2014), pp. A588–A608.
[16] J. R. Dormand and P. J. Prince, *A family of embedded runge-kutta formulae*, Journal of computational and applied mathematics, 6 (1980), pp. 19–26.
[17] G. Evensen, *Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics*, Journal of Geophysical Research: Oceans, 99 (1994), pp. 10143–10162.
[18] M. Frei and H. R. Künsch, *Bridging the ensemble Kalman and particle filters*, Biometrika, 100 (2013), pp. 781–800.
[19] S. Güttel, *Rational Krylov approximation of matrix functions: Numerical methods and optimal pole selection*, GAMM-Mitteilungen, 36 (2013), pp. 8–31.
[20] M. W. Hirsch, S. Smale, and R. L. Devaney, *Differential equations, dynamical systems, and an introduction to chaos*, Academic press, 2012.
[21] P. Janssen, J. S. Marron, N. Veraverbeke, and W. Sarle, *Scale measures for bandwidth selection*, Journal of Nonparametric Statistics, 5 (1995), pp. 359–380, https://doi.org/10.1080/10485259508832654, https://doi.org/10.1080/10485259508832654, https://arxiv.org/abs/https://doi.org/10.1080/10485259508832654.
[22] E. T. Jaynes, *Probability theory: The logic of science*, Cambridge university press, 2003.
[23] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980, (2014).
[24] O. Ledoit and M. Wolf, *A well conditioned estimator for large dimensional covariance matrices*, Journal of Multivariate Analysis, 88 (2004), pp. 365–411, 10.1016/S0047-259X(03)00096-4.
[25] B. Liu, B. Ait-El-Fquih, and I. Hoteit, *Efficient kernel-based ensemble gaussian mixture filtering*, Monthly Weather Review, 144 (2016), pp. 781–800.
[26] E. N. Lorenz, *Deterministic nonperiodic flow*, Journal of atmospheric sciences, 20 (1963), pp. 130–141.
[27] E. N. Lorenz, *Predictability: A problem partly solved*, in Proc. Seminar on predictability, vol. 1, 1996.
[28] R. M. Neal and G. E. Hinton, *A view of the em algorithm that justifies incremental, sparse, and other variants*, in Learning in graphical models, Springer, 1998, pp. 355–368.
[29] E. D. Nino-Ruiz and A. Sandu, *Ensemble Kalman filter implementations based on shrinkage covariance matrix estimation*, Ocean Dynamics, 65 (2015), pp. 1423–1439.
[30] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer, 1999.
[31] A. Papoulis and S. Pilla, *Probability, random variables, and stochastic processes, 583 pp*, 1965.
[32] A. A. Popov, C. Mou, A. Sandu, and T. Iliescu, *A multifidelity ensemble Kalman filter with reduced order control variates*, SIAM Journal on Scientific Computing, 43 (2021), pp. A1134–A1162.
[33] A. A. Popov and A. Sandu, *A bayesian approach to multivariate adaptive localization in ensemble-based data assimilation with time-dependent extensions*, Nonlinear Processes in Geophysics, 26 (2019), pp. 109–122.
[34] A. A. Popov, A. Sandu, E. D. Nino-Ruiz, and G. Evensen, *A stochastic covariance shrinkage approach in ensemble transform Kalman filtering*, arXiv preprint arXiv:2003.00354, (2020).
[35] A. A. Popov, A. N. Subrahmanya, and A. Sandu, *A stochastic covariance shrinkage approach to particle rejuvenation in the ensemble transform particle filter*, Nonlinear Processes in Geophysics, 29 (2022), pp. 241–253.
[36] M. L. Psiaki, *Gaussian mixture nonlinear filtering with resampling for mixand narrowing*,

IEEE Transactions on Signal Processing, 64 (2016), pp. 5499–5512.
[37] S. Reich and C. Cotter, *Probabilistic forecasting and Bayesian data assimilation*, Cambridge University Press, 2015.
[38] H. Robbins and S. Monro, *A stochastic approximation method*, The annals of mathematical statistics, (1951), pp. 400–407.
[39] F. Roosta-Khorasani and M. W. Mahoney, *Sub-sampled newton methods*, Mathematical Programming, 174 (2019), pp. 293–326.
[40] B. W. Silverman, *Density estimation for statistics and data analysis*, Routledge, 2018.
[41] H. W. Sorenson and D. L. Alspach, *Recursive bayesian estimation using Gaussian sums*, Automatica, 7 (1971), pp. 465–479.
[42] A. S. Stordal, H. A. Karlsen, G. Nævdal, H. J. Skaug, and B. Vallès, *Bridging the ensemble Kalman filter and particle filters: the adaptive gaussian mixture filter*, Computational Geosciences, 15 (2011), pp. 293–305.
[43] P. J. Van Leeuwen, H. R. Künsch, L. Nerger, R. Potthast, and S. Reich, *Particle filters for high-dimensional geoscience applications: A review*, Quarterly Journal of the Royal Meteorological Society, 145 (2019), pp. 2335–2365.
[44] S. Yun, R. Zanetti, and B. A. Jones, *Kernel-based ensemble gaussian mixture filtering for orbit determination with sparse data*, Advances in Space Research, 69 (2022), pp. 4179–4197.