Are Non-Gaussian Kernels Suitable for Ensemble Mixture Model Filtering?

Andrey A. Popov
The Oden Institute for Computational Engineering & Sciences
The University of Texas at Austin
Austin, Texas
andrey.a.popov@utexas.edu

Renato Zanetti
Dept. of Aerospace Engineering and Engineering Mechanics
The University of Texas at Austin
Austin, Texas
renato@utexas.edu

***Abstract*—In the high-dimensional setting, Gaussian mixture kernel density estimates become increasingly suboptimal. In this work we aim to show that it is practical to instead use the optimal multivariate Epanechnikov kernel. We make use of this optimal Epanechnikov mixture kernel density estimate for the sequential filtering scenario through what we term the ensemble Epanechnikov mixture filter (EnEMF). We provide a practical implementation of the EnEMF that is as cost efficient as the comparable ensemble Gaussian mixture filter. We then showcase that the EnEMF has a significant reduction in error per particle on the 40-variable Lorenz '96 system. We answer the titular question, "are non-Gaussian kernels suitable for ensemble mixture model filtering?" in the affirmative.**

***Index Terms*—Non-linear Estimation, High-dimensional filtering, Kernel Density Estimation, Epanechnikov Kernel**

I. Introduction

State estimation [1] and data assimilation [2], [3] methods restricted to propagating a single mean and a single covariance are limited to dealing with near-linear near-Gaussian scenarios. When dealing with highly non-linear dynamics and measurements, these methods do not have robust convergence guarantees and, in the worst case, actively fight against the goal of estimating the uncertainty of the dynamics of interest.

The Gaussian sum filter (GSF) [4], [5] is capable of fully representing almost all useful probability density functions through the use of Gaussian mixture models (GMM) and the Gaussian sum update. Fundamentally, it suffers from requiring a near-Gaussian update of the covariances and requires sophisticated splitting and merging techniques [6] in order to avoid weight collapse in the components. This means that the GSF has the potential to not behave well in the highly non-linear setting.

On the other hand, particle filters [7] are capable of representing any probability density through a collection of samples. A sample-based representation of a probability density is simpler than a GMM, though requires significantly more particles to represent many probability densities, particularly in the high-dimensional setting. The dominant roadblock to the use of particle filters is that they require significant effort to avoid weight collapse or filter collapse by employing either sophisticated resampling or particle flow techniques [8], [9].

A filter that bridges the gap between the two methodologies is the ensemble Gaussian mixture filter (EnGMF) [10], [11], [12], [13]. The EnGMF relies on a kernel density estimate (KDE) to build a GMM representation of our prior uncertainty from a particle-based propagation step, incorporates measurement information into the posterior through the use of the Gaussian sum update. GMMs are attractive for building KDEs state estimation because of well-known "nice" properties of the Gaussian distribution [14].

In the high-dimensional setting, however, GMMs for KDE become progressively less 'efficient' as the dimension grows. Kernel density estimates based on the Epanechnikov ([jəpɐnʲˈetɕnʲɪkɐf]) kernel minimize the error with respect to the underlying distribution of the particles. This work presents an ensemble Epanechnikov mixture filter (EnEMF) that takes advantage of the efficiency of the Epanechnikov kernel in order to perform high-dimensional particle filtering.

This paper is organized as follows: section II provides background on ensemble mixture model filtering, kernel density estimation, and motivation as to the superiority of the Epanechnikov kernel. The EnEMF is described in section III, with a practical implementation that makes use of the Gaussian sum update. A numerical experiment on the 40-variable Lorenz '96 system is provided in section IV. Finally section V provides closing remarks about why we believe the Epanechnikov kernel is suitable for ensemble mixture model filtering.

II. Background and Motivation

Ensemble mixture modeling filtering, visually described in fig. 1, combines four techniques for state estimation: particle propagation, kernel density estimation, the mixture model update, and resampling. Each one providing an essential component that allows for accurate representations of the "true" [15] posterior with a finite number of samples.

Particle propagation [3] allows for long-term propagation of our knowledge about the state of a particular system through highly non-linear dynamics with arbitrary precision. Formally,

This work is based on reaserch that is in part sponsored by the Air Force Office of Scientific Research (AFOSR) under grant FA9550-23-1-0646

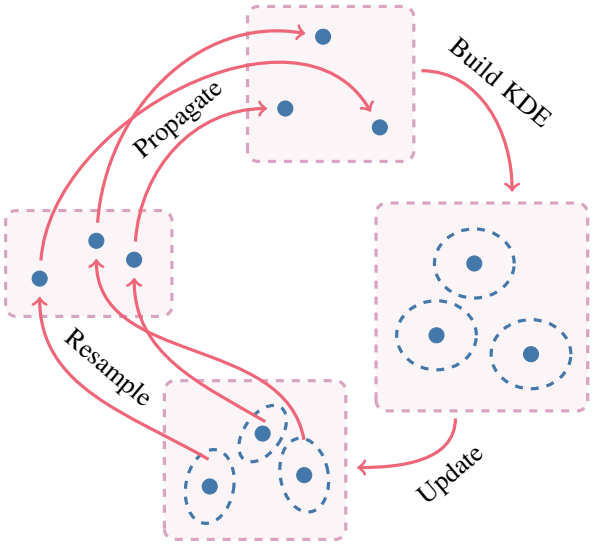


Fig. 1. Ensemble mixture model filtering diagram. Clock wise from the left-most rectangle: given a collection of particles from the previous time, (i) propagate to the current time, (ii) build a mixture model representing the prior from the particles through kernel density estimation, (iii) update the mixture model by making use of the measurements to create a mixture model representing the posterior, and (iv) resample to create a new collection of particles. This process is repeated until a desired time is reached or ad infinitum.

assume that we are given a collection of N independently and identically distributed samples,

$$\mathbf{X}_{k-1} = [x_1, x_2, \dots x_N], \quad \mathbf{X}_{k-1} \in \mathbb{R}^{n \times N}, \tag{1}$$

from the n-dimensional probability distribution of interest at time index $k-1$. With abuse of notation, we can write their propagation through some (discrete or continuous, stochastic or deterministic) dynamics,

$$\mathbf{X}_k = F(\mathbf{X}_{k-1}), \tag{2}$$

resulting in a collection of N particles at time index k. Generally the dynamics F can incorporate some measure of uncertainty through some type of process noise or measure of model error, though for simplicity and without any real loss of generality we ignore any such notions for the remainder of this work.

The resulting particles can be used to represent our knowledge about the state of a system through the use of sample statistic, and, for the purposes of this work, can be used to construct a representation of the distribution of interest at time index k through the use of kernel density estimation techniques [16]. If the distribution of interest is denoted by f, the kernel density estimate of the distribution of interest described by the ensemble eq. (1) is the mixture model,

$$\widetilde{f}_{\mathbf{X}_N}(x) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{K}(x\,;\, x_i, \mathbf{C}_i), \tag{3}$$

where $\mathcal{K}$ is a probability distribution known as a kernel that is parameterized by the sample x_i and by some matrix factor $\mathbf{C}_i$ that represents a measure of local covariance that is defined later. We assume the kernel $\mathcal{K}$ satisfies the following properties,

$$\begin{aligned} \mathcal{K}(x) &\geq 0 \quad \forall x \in \mathbb{R}^n \\ \int_{\mathbb{R}^n} \mathcal{K}(x)\mathrm{d}x &= 1, \\ \int_{\mathbb{R}^n} x\mathcal{K}(x)\mathrm{d}x &= 0, \\ \int_{\mathbb{R}^n} xx^{\mathsf{T}}\mathcal{K}(x)\mathrm{d}x &< \infty, \\ \mathcal{K}(x;\mu,\mathbf{H}) &:= \frac{1}{\sqrt{|\mathbf{H}|}}\mathcal{K}\left(\mathbf{H}^{1/2}(x-\mu)\right), \quad \forall\, \mathbf{H} > 0 \\ \|x\| = \|y\| &\implies \mathcal{K}(x) = \mathcal{K}(y), \quad \forall x, y \in \mathbb{R}^n, \end{aligned} \tag{4}$$

where the first and second properties properties ensure that the Kernel is a probability density, the third property ensures that the kernel is zero mean, the fourth property ensures that the kernel has a finite (co-)variance, the fifth property defines arbitrary shifting by mean and re-scaling by symmetric positive definite matrices, and the sixth property ensures that the kernel is radially symmetric.

Assume that we are given a nonlinear measurement of the state,

$$y = h(x) + \eta, \tag{5}$$

through some non-linear function h, and with unbiased, $\mathbb{E}[\eta] = 0$, additive error that has covariance $\mathrm{Cov}[\eta] = \mathbf{R}$. For the remainder of this work we make the common, but not required assumption that the measurement likelihood is Gaussian,

$$p(y|x) = \mathcal{N}(y\,;\, h(x), \mathbf{R}), \tag{6}$$

though this assumption can be relaxed and generalized in various different ways that are outside the scope of this work.

We want to leverage this general framework to perform Bayesian inference, thus our goal is to find a representation of the posterior distribution,

$$p(x|y) \propto p(y|x)p(x), \tag{7}$$

which in the case of the prior mixture model eq. (3) and measurement likelihood eq. (6) is exactly described by the mixture model,

$$p(x|y) = \sum_{i=1}^{N} w_i\, \widetilde{\mathcal{K}}(x\,;\, x_i, \mathbf{C}_i, y, \mathbf{R}), \tag{8}$$

where the new mixture components are given by

$$\widetilde{\mathcal{K}}(x\,;\, x_i, \mathbf{C}_i, y, \mathbf{R}) = \frac{\mathcal{K}(x\,;\, x_i, \mathbf{C}_i)\mathcal{N}(y\,;\, h(x), \mathbf{R})}{\int_{\mathbb{R}^n} \mathcal{K}(x\,;\, x_i, \mathbf{C}_i)\mathcal{N}(y\,;\, h(x), \mathbf{R})\,\mathrm{d}x}, \tag{9}$$

with the weights,

$$w_i \propto \int_{\mathbb{R}^n} \mathcal{K}(x\,;\, x_i, \mathbf{C}_i)\mathcal{N}(y\,;\, h(x), \mathbf{R})\,\mathrm{d}x, \tag{10}$$

which are derived by simple convolution of distributions.

If our probability of interest f has known finite covariance,

$$\mathrm{Cov}(f) = \mathbf{\Sigma}_f < \infty, \tag{11}$$

then the standard scalar parameterization [16] of the covariances, $\mathbf{C}_i$, in eq. (3), is given by,

$$\mathbf{C}_i = h^2 \mathbf{\Sigma}_f, \quad i = 1, \ldots, N, \tag{12}$$

where h is known as the bandwidth parameter. In this work we make use of the scalar parameterization to restrict the parameters of our kernel density estimate to two choices: the kernel $\mathcal{K}$ and the bandwidth h.

The most common type of kernel that is used is the (zero-mean, unit-covariance-scaled) Gaussian kernel,

$$\mathcal{N}(x) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}x^\mathsf{T} x}, \tag{13}$$

which has many nice properties that make it simple to use and reason about. These very same properties make it extremely attractive to the practitioner, and make its choice a thought-terminating cliché—the mere invocation of the Gaussian kernel assumption terminates reasoning without significant push-back. We now show why the Gaussian kernel assumption needs to be questioned—especially in the high-dimensional setting.

A. Minimizing KDE Error

Let's now put our focus on the goodness-of-fit of the kernel density estimate eq. (3). The most common metric that describes how well the KDE estimate approximates the target distribution is the mean integral squared error (MISE),

$$\mathrm{MISE}(f, \widetilde{f}) = \mathbb{E}_{\mathbf{X}_N}\left[\int_{\mathbb{R}^n} \left(f(x) - \widetilde{f}(x)\right)^2 \mathbf{d}x\right] \tag{14}$$

which describes the squared error over all the support of the target distribution f averaged over all possible realizations of N samples $\mathbf{X}_N$. Dealing with the MISE directly is intractable for most kernels and distributions of interest, thus the approximated MISE (AMISE),

$$\mathrm{AMISE}(f, \widetilde{f}) = \frac{1}{4} h^4 \alpha^2 \gamma + N^{-1} h^{-n} \beta, \tag{15}$$

is frequently used instead. The new parameters of eq. (15) are given by,

$$\begin{aligned} \alpha &= \frac{1}{n} \operatorname{tr}\left(\int_{\mathbb{R}^n} x^\mathsf{T} x\, \mathcal{K}(x) \mathrm{d}x\right) \\ \beta &= \int_{\mathbb{R}^n} \mathcal{K}(x)^2 \mathrm{d}x, \\ \gamma &= \int_{\mathbb{R}^n} \operatorname{tr}^2\left[\nabla_x^2 \widehat{f}(x)\right] \mathrm{d}x, \end{aligned} \tag{16}$$

where $\widehat{f}$ is the scaling of f by $\mathbf{\Sigma}_f$ that is ameanable to representation by the unscaled kernel K. The derivation of the above can be found in [16]. If the target distribution $\widehat{f}$ in eq. (16) is the unit-Covariance Gaussian, then the last parameter simplifies to,

$$\gamma = \frac{1}{2^n \sqrt{\pi}^n} \left(\frac{1}{2} n + \frac{1}{4} n^2\right), \tag{17}$$

which—while not required to be defined for any of the subsequent derivations—is the value that is used for the rest of this work, as exploring alternatives is outside the current scope.

Given an arbitrary kernel $\mathcal{K}$ we want create a kernel density estimate that minimizes the AMISE eq. (15). In effect, this means that we want choose the bandwidth parameter in eq. (12) that is optimal in terms of the error. The following result about the optimal bandwidth is from [16]:

Theorem II.1. *The optimal bandwidth in eq.* (12) *that minimizes the AMISE in eq.* (16)*, is given by,*

$$h = \left[\frac{\beta n}{\alpha^2 \gamma N}\right]^{\frac{1}{n+4}}, \tag{18}$$

where n is the dimension of the system, N is the number of samples and the rest of the parameters are defined by eq. (16)*.*

We now go the other direction. Fixing the optimal bandwidth to be eq. (18), we want to find the kernel that minimizes the AMISE. Substituting the optimal bandwidth eq. (18) into the AMISE error metric eq. (15) we get,

$$\mathrm{AMISE}(f, \widetilde{f}) = \underbrace{\frac{(n+4)}{4} \gamma^{\frac{n}{n+4}}}_{\text{reference dist.}} \underbrace{\beta \left(\frac{n\beta}{\alpha^2}\right)^{-\frac{n}{n+4}}}_{C(\mathcal{K})} \underbrace{N^{-\frac{4}{n+4}}}_{\text{conv. rate}}, \tag{19}$$

where the first term is purely a function of the target distribution (if known), or is a function of the mismatch between the true distribution and a reference used to compute the term γ. The third term in eq. (19) is the rate of converengece in the number of samples N, which becomes slower and slower as the dimension n increases. The term that we are interested in is the second term, $C(\mathcal{K})$, which is purely dependent on the choice of kernel $\mathcal{K}$. As the rate of convergence is sub-linear, the scaling term $C(\mathcal{K})$ plays a significant role in the convergence of the kernel density estimation method. This means that the choice of kernel is the sole choice that fully determines the error of the KDE.

The kernel that minimizes $C(\mathcal{K})$, is given by [17], [18] and is the unit-covariance-scaled Epanechnikov distribution,

$$\mathcal{E}(x) = \frac{n+2}{2c_n (n+4)^{\frac{n+2}{2}}} (n + 4 - x^\mathsf{T} x), \quad x^\mathsf{T} x < n + 4, \tag{20}$$

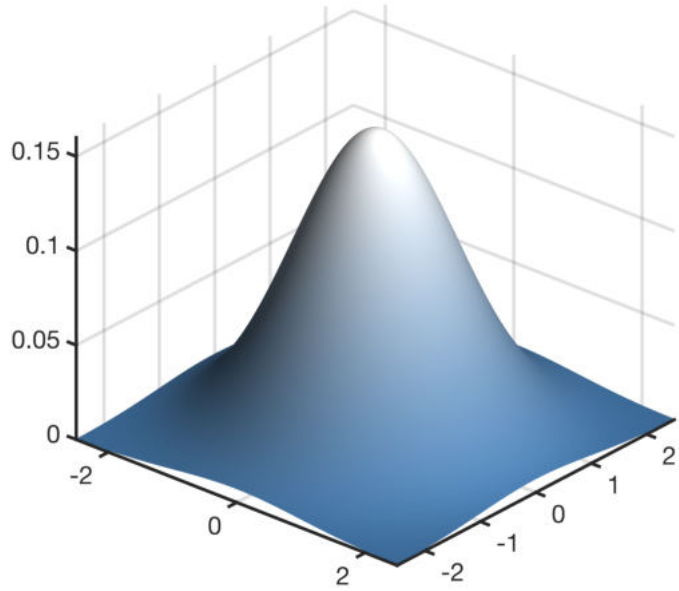
where,

$$c_n = \frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2} + 1\right)}, \tag{21}$$

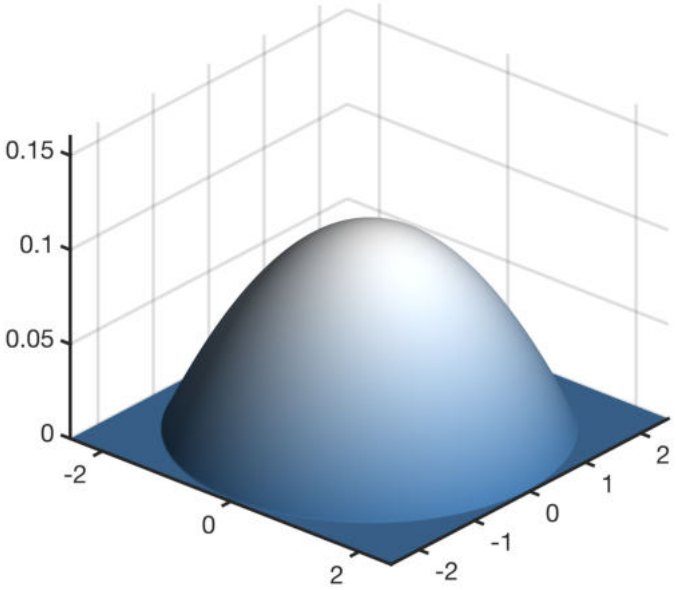
is the volume of a unit sphere in n dimensions. A visualization of the Epanechnikov distribution eq. (20) compared to the Gaussian distribution eq. (13) is presented in fig. 2.

As eq. (20) is the optimal kernel that minimizes the AMISE eq. (19) given the optimal bandwidth, it is reasonable to ask, just how much accuracy are we sacrificing when we choose a suboptimal kernel? We can quantify the how effective a given kernel is relative to the Epanechnikov kernel by calculating the the efficiency [16] of an arbitrary kernel $\mathcal{K}$,

$$\mathrm{eff}(\mathcal{K}) = \left(\frac{C(\mathcal{E})}{C(\mathcal{K})}\right)^{\frac{n+4}{4}}, \tag{22}$$



(a) Gaussian distribution



(b) Epanechnikov distribution

Fig. 2. (a) Gaussian and (b) Epanechnikov distribution for $n = 2$ spatial dimensions, mean of zero and identity covariance.

which describes a scaling of the effective ensemble size of the kernel $\mathcal{K}$ with respect to the optimal Epanechnikov distribution $\mathcal{E}$. The power $(n+4)/4$ represents the inverse of the rate of convergence in eq. (19). In other words, this means that the error using N samples and the kernel $\mathcal{K}$ is equivalent to using $N \operatorname{eff}(\mathcal{K})$ samples and the kernel $\mathcal{E}$. Conversely this also means that the error of using the Epanechnikov kernel $\mathcal{E}$ with N samples is equivalent to using the kernel $\mathcal{K}$ with $N/\operatorname{eff}(\mathcal{K})$ samples. Note that the efficiency eq. (22) is highly dependent on the dimension n, thus in the worst-case the efficiency could effectively be zero for a large enough n.

Before we give a closed form expression for the efficiency

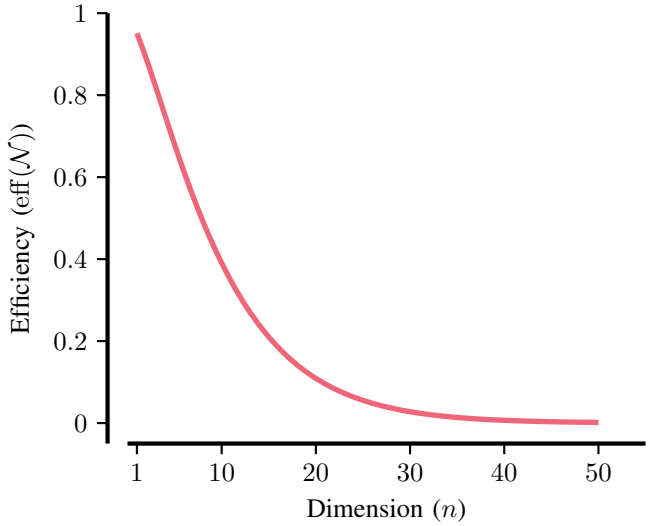


Fig. 3. Efficiency of the Gaussian kernel relative to the dimension n. Values of the efficiency for non-integer dimensions are plotted for completeness.

of the Gaussian kernel, we have to derive a few constants for the Gaussian eq. (13) and Epanechnikov eq. (20) kernels.

Lemma II.2. *By trivial applications on basic calculus, for the Gaussian kernel,*

$$\alpha_{\mathcal{N}} = 1, \quad \beta_{\mathcal{N}} = \frac{1}{(2\sqrt{\pi})^n}, \tag{23}$$

are the constants defined by eq. (16)*.*

Lemma II.3. *Again, by trivial applications on basic calculus, for the Epanechnikov kernel,*

$$\alpha_{\mathcal{E}} = 1, \quad \beta_{\mathcal{E}} = \frac{2}{c_n}(n+2)(n+4)^{-\frac{n}{2}-1}, \tag{24}$$

are the constants defined by eq. (16)*.*

We now have the tools necessary to derive the efficiency of the Gaussian kernel eq. (13) for n-dimensional KDE.

Theorem II.4. *The efficiency eq.* (22) *of the Gaussian kernel is,*

$$\operatorname{eff}(\mathcal{N}) = \frac{2^{n+2}}{(n+4)^{\frac{n}{2}+1}}\Gamma\left(\frac{n}{2}+2\right), \tag{25}$$

Proof. This is a direct application of the constants in lemma II.2 and lemma II.3 to eq. (22). □

Corollary II.4.1. *By simple application of elementary calculus rules, the efficiency of the Gaussian kernel in theorem II.4 tends towards zero as the dimension of the system n tends towards infinity,*

$$\lim_{n\to\infty} \operatorname{eff}(\mathcal{N}) = 0. \tag{26}$$

The efficiency of the Gaussian kernel is plotted in fig. 3. Notice that the rate of decay of the efficiency is exponential, and that for $n = 40$ the efficiency is well below 1%. This means that the Gaussian kernel is highly inefficient in higher

dimensions, and that attempting to use the Epanechnikov kernel might be cost-effective.

B. Sampling from a Epanechnikov distribution

Being able to sample from the Epanechnikov distribution is important step for sampling from the approximated posterior. Similar to eq. (4), we can write the Epanechnikov distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ as

$$\mathcal{E}(x; \mu, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} \mathcal{E}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}(x-\mu)\right). \tag{27}$$

It is known from [19] that the random variable,

$$\begin{aligned} \varepsilon &= \boldsymbol{\mu} + \boldsymbol{\Sigma}^{\frac{1}{2}} \sqrt{(n+4)\eta}\, \mathbf{T}, \\ \mathbf{T} &\sim \mathcal{U}(\mathcal{S}^{n-1}), \quad \eta \sim \beta\left(\frac{n}{2}, 2\right), \end{aligned} \tag{28}$$

is a sample from the Epanechnikov distribution eq. (27) with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. We slightly modify eq. (28) in order to accomodate further modification later on. A sample from the Epanechnikov distribution can be created using the following procedure:

1) Sample a random s from the unit Gaussian distribution $\mathcal{N}(\mathbf{0}_n, \mathbf{I}_{n\times n})$,
2) project s onto the shell with radius $\sqrt{n+4}$, to get $\hat{s} = \frac{\sqrt{n+4}}{\|s\|} s$,
3) sample the beta-distributed, $\kappa \sim \beta(n/2, 2)$,
4) combine to get a sample $\varepsilon = \mu + \boldsymbol{\Sigma}^{\frac{1}{2}} \kappa \hat{s}$.

C. Ensemble Gaussian Mixture Filter

We now present the Ensemble Gaussian mixture filter, which we subsequently generalize to make use of the optimal Epanechnikov mixture.

Given an ensemble of N samples from some unknown prior distribution,

$$\mathbf{X}^- = \left[x_1^-, x_2^-, \ldots, x_N^-\right], \tag{29}$$

first approximate the covariance of the distribution,

$$\widetilde{\boldsymbol{\Sigma}}^- = \frac{1}{N-1} \mathbf{X}^- \left(\mathbf{I}_{N\times N} - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T\right) \mathbf{X}^{-,T}, \tag{30}$$

as a proxy to the known covariance in eq. (11).

Second, build KDE estimate of the distribution as in eq. (3),

$$\widetilde{f}_{\mathbf{X}^-}(x) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{N}(x; x_i^-, h_{\mathcal{N}}^2 \widetilde{\boldsymbol{\Sigma}}^-), \tag{31}$$

by making use of the covariance approximation in eq. (30) and the optimal bandwidth $h_{\mathcal{N}}$ defined by eq. (18) with constants defined for the normal distribution in lemma II.2.

We can take advantage of the Gaussian mixture structure, to perform an Gaussian sum update on each one of the components to get an approximation of the posterior. In this work we make use of the extended Kalman filter to perform said update from the prior Gaussian mixture eq. (31) to an approximation of the posterior,

$$\begin{aligned} x_i^{\sim} &= x_i^- - \mathbf{G}_i \left(h(x_i^-) - y\right), \\ h_{\mathcal{N}}^2 \widetilde{\boldsymbol{\Sigma}}_i^{\sim} &= \left(\mathbf{I} - \mathbf{G}_i \mathbf{H}_i^T\right) h_{\mathcal{N}}^2 \widetilde{\boldsymbol{\Sigma}}_i^-, \\ \mathbf{G}_i &= h_{\mathcal{N}}^2 \widetilde{\Sigma}_i^- \mathbf{H}_i^T \left(\mathbf{H}_i h_{\mathcal{N}}^2 \widetilde{\Sigma}_i^- \mathbf{H}_i^T + \mathbf{R}_i\right)^{-1}, \\ w_i &\propto \mathcal{N}\left(y\,;\, h(x_i^-),\, \mathbf{H}_i h_{\mathcal{N}}^2 \widetilde{\boldsymbol{\Sigma}}_i^- \mathbf{H}_i^T + \mathbf{R}\right), \\ \mathbf{H}_i &= \left.\frac{dh}{dx}\right|_{x=x_i^-}, \end{aligned} \tag{32}$$

with $\widetilde{x}_i$ representing the mean of the ith posterior component, $h_{\mathcal{N}}^2 \widetilde{\boldsymbol{\Sigma}}_i^{\sim}$ representing its covariance, and w_i its weight. Thus, the Gaussian mixture,

$$\widetilde{f}_{\mathbf{X}^+}(x) = \sum_{i=1}^{N} w_i \mathcal{N}(x; x_i^{\sim}, h_{\mathcal{N}}^2 \widetilde{\boldsymbol{\Sigma}}_i^{\sim}), \tag{33}$$

is an approximation to the posterior. Note that as the GMM update eq. (32) is exact when the measurement h in eq. (5) is linear, the measurement error η is Gaussian, and when the prior mixture model is a Gaussian mixture.

The final step of the EnGMF is to resample from the posterior approximation eq. (33) through the use of standard techniques such as:

1) First sample from the probability mass function defined by the weights $\{w_i\}_{i=1}^N$, to get a mode j,
2) then sample from the normal distribution defined by the mode, $\mathcal{N}(x_j^{\sim}, h_{\mathcal{N}}^2 \widetilde{\boldsymbol{\Sigma}}_i^{\sim})$,
3) repeat as many times as samples are required.

We make use of all of the machinery of the EnGMF to generalize to a Epanechnikov kernel mixture.

III. Ensemble Epanechnikov Mixture Filter

We now present the ensemble Epanechnikov mixture filter as a generalization of the ensemble Gaussian mixture filter to the more efficient Epanechnikov mixture model.

Just like the EnGMF, the first step of the EnEMF is to find the statistical covariance eq. (30). The second step is to perform kernel density estimation,

$$\widetilde{f}_{\mathbf{X}^-}(x) = \sum_{i=1}^{N} \frac{1}{N} \mathcal{E}(x; x_i^-, h_{\mathcal{E}}^2 \widetilde{\boldsymbol{\Sigma}}_i^-) \tag{34}$$

but this time using the Epanechnikov distribution eq. (27) and the Epanechnikov bandwidth from lemma II.3.

Given the eq. (34) the exact posterior distribution is given by

$$\widetilde{f}_{\mathbf{X}^+}(x) = \sum_{i=1}^{N} w_i\, \widetilde{\mathcal{E}}\left(x; x_i^-, h_{\mathcal{E}}^2 \widetilde{\boldsymbol{\Sigma}}_i^-, y, \mathbf{R}\right), \tag{35}$$

where the distribution above is of the form eq. (9),

$$\begin{aligned} \widetilde{\mathcal{E}}\left(x; x_i^-, h_{\mathcal{E}}^2 \widetilde{\boldsymbol{\Sigma}}_i^-, y, \mathbf{R}\right) \propto & \mathcal{E}(x; x_i^-, h_{\mathcal{E}}^2 \widetilde{\boldsymbol{\Sigma}}_i^-) \\ & \cdot \mathcal{N}(y\,;\, h(x_i^-), \mathbf{R}) \end{aligned}, \tag{36}$$

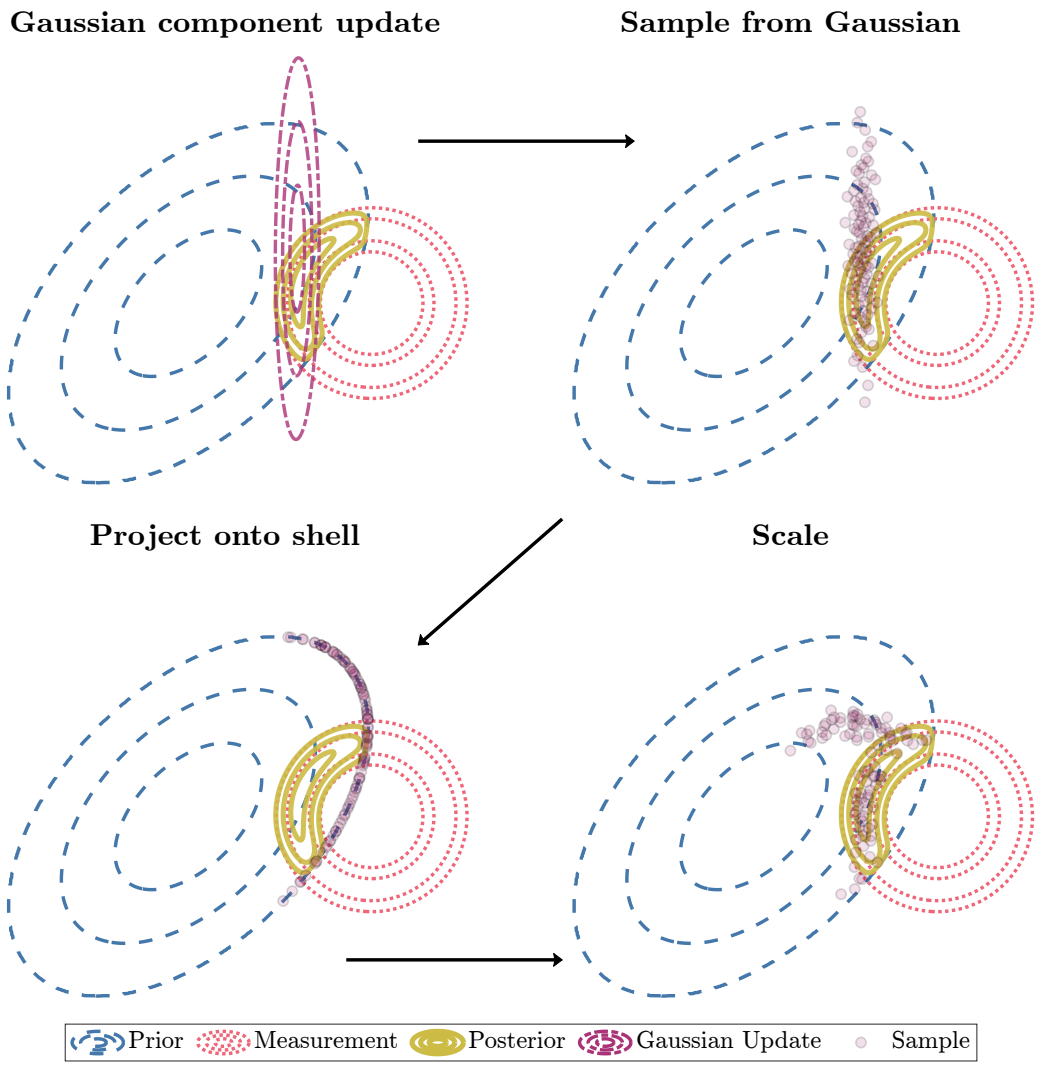


Fig. 4. A visual description of the approximated posterior Epanechnikov sampling procedure for one Epanechnikov mode. First, in the top left panel the Gaussian component update from the prior to the candidate posterior is performed using the measurement likelihood with samples taken therefrom in the top right panel. Next, in the bottom left panel the samples are projected onto the shell of the prior Epanechnikov component, and finally the projected samples are randomly scaled in the bottom left panel.

which does not have a simple closed form representation. We show later that it is possible to sample from eq. (36) by making use of the Gaussian sum update,

$$\begin{aligned} x_i^{\sim} &= x_i^- - \mathbf{G}_i \left(h(x_i^-) - y \right), \\ h_{\mathcal{E}}^2 \widetilde{\mathbf{\Sigma}}_i^{\sim} &= \left(\mathbf{I} - \mathbf{G}_i \mathbf{H}_i^T \right) h_{\mathcal{E}}^2 \widetilde{\mathbf{\Sigma}}_i^-, \\ \mathbf{G}_i &= h_{\mathcal{E}}^2 \widetilde{\Sigma}_i^- \mathbf{H}_i^T \left(\mathbf{H}_i h_{\mathcal{E}}^2 \widetilde{\Sigma}_i^- \mathbf{H}_i^T + \mathbf{R}_i \right)^{-1}, \\ \mathbf{H}_i &= \left. \frac{dh}{dx} \right|_{x = x_i^-}, \end{aligned} \tag{37}$$

where the terms are almost identical to the EnGMF update eq. (32) except for the bandwidth factor $h_{\mathcal{E}}$, and the weights, which require special attention.

For the weights $\{w_i\}_{i=1}^N$ in eq. (35), we can (i) again use the same weights defined in the Gaussian sum update eq. (32), (ii) attempt to exactly compute the integral in eq. (10), or (iii) approximate the integral through a closed form Gaussian solution. In the authors' experience, (i) does not produce accurate results, and (ii) incurs a significant computational cost in the high-dimensional setting, therefore (iii) is the most reasonable option for now.

Observe that though the use of a Taylor series approximation to the logarithm,

$$\begin{aligned} n + 4 - (x-\mu)^T \Sigma^{-1} (x-\mu) = \\ C e^{-\frac{1}{2}(x-\mu)^T \left(\frac{n+4}{2} \Sigma \right)^{-1} (x-\mu)^T + \text{h.o.t.}}, \end{aligned} \tag{38}$$

where C is a constant factor that depends on $n+4$, meaning that it will cancel, and the higher order terms we can choose to discard.

Therefore by using the derivation in eq. (38) for the weights it is possible to approximate the weights of the EnEMF update

as,

$$w_i \propto \mathcal{N}(y\,;\, h(x_i^-), \mathbf{H}_i \frac{s_\mathcal{E} h_\mathcal{E}^2 (n+4)}{2} \widetilde{\Sigma}_i^- \mathbf{H}_i^T + \mathbf{R}), \quad (39)$$

where $s_\mathcal{E}$ is scaling factor to help taper the effect of under-weighting. The hope, is that with an optimally tuned scaling factor $s_\mathcal{E}$, the weight update approximation in eq. (39) should approach the true mixture model weight update in eq. (10). Importantly, the weight update in eq. (39) has the same computational cost as the weight update in the EnGMF eq. (32), and leverages a very similar implementation—the only difference involving a slightly modified covariance. As an additional consideration the weight update in eq. (39) is compatible with convergence of the EnEMF in the limit of particle number, though a proof of this is ancillary to this work.

The most important step is the resampling procedure, as it is the one that is most heavily modified from that of the EnGMF, and relies on the Epanechnikov sampling procedure described in section II-B. We can perform resampling from the EnEMF posterior eq. (35) in the following way:

1) Generate a sample from the discrete distribution defined by the weights $\{w_i\}_{i=1}^N$, which defines the mode j,
2) generate a sample from the$j$th mode of the Gaussian distribution defined by the Gaussian sum update

$$u \sim \mathcal{N}(x_j^\sim, h_\mathcal{E}^2 \widetilde{\mathbf{\Sigma}}_j^\sim), \quad (40)$$

3) project u onto the unit shell defined by the prior mode $\mathcal{E}(x_j^-, h_\mathcal{E}^2 \widetilde{\mathbf{\Sigma}}_j^-)$,

$$\begin{aligned} s &= \left(h_\mathcal{E}^2 \widetilde{\mathbf{\Sigma}}_j^-\right)^{-\frac{1}{2}} \left(u - x_j^-\right), \\ \hat{s} &= \frac{\sqrt{n+4}}{\|s\|} s, \end{aligned} \quad (41)$$

to find the sample direction $\hat{s}$ relative to the prior mode mean x_j^-, meaning that $x_j^- + \left(h_\mathcal{E}^2 \widetilde{\mathbf{\Sigma}}_j^-\right)^{\frac{1}{2}} \hat{s}$ lies on the boundary of the prior mode, then

4) sample the variable κ from the modified beta distribution in the direction $\hat{s}$

$$\begin{aligned} \kappa \sim \frac{1}{2}(n+2) z^{n-1}(1-z^2) \\ \cdot \mathcal{N}\left(y\,;\, h\left(x_j^- + \left(h_\mathcal{E}^2 \widetilde{\mathbf{\Sigma}}_j^-\right)^{\frac{1}{2}} z\hat{s}\right), \mathbf{R}\right), \end{aligned} \quad (42)$$

in the scalar variable $0 \leq z < 1$, through an inverse CDF method,

5) and finally, combine with the prior mean and covariance to get a sample

$$\varepsilon = x_i + \mathbf{\Sigma}^{\frac{1}{2}} \kappa \hat{s}. \quad (43)$$

A visual interpretation of the resampling procedure can be found in fig. 4.

Conjecture 1. *The resampling procedure defined above is exact when the measurement operator h is linear, and the weights $\{w_i\}_{i=1}^N$ are not approximated.*

A proof of the above is of future interest.

IV. Numerical Experiment

The goal of the numerical experiment is to show that the EnEMF has the potential to be a superior filter to that of the EnGMF in the high-dimensional setting. We thus make use of the 40-variable Lorenz '96 equations,

$$x_k' = -x_{k-1}(x_{k-2} - x_{k+1}) - x_k + F \ldots, \quad k = 1, \ldots, 40, \quad (44)$$

where by the cyclic boundary conditions, $x_0 = x_{40}$, $x_{-1} = x_{39}$, and $x_{41} = x_1$. The forcing is set to $F = 8$ to have a chaotic system with a Kaplan-Yorke dimension of 27.1 and 13 positive Lyapunov exponents [20].

For the non-linear measurement operator we take a magnitude measurement of adjacent variables,

$$[h(x)]_i = \sqrt{x_{2i+1}^2 + x_{2i+2}^2}, \quad i = 1, \ldots, 20, \quad (45)$$

with an error covariance of $\mathbf{R} = \frac{1}{4}\mathbf{I}_{20}$.

For a 40 variable system the efficiency of the Gaussian kernel is about 0.6%, meaning that for a density estimate with $N = 100$ samples with the Epanechnikov kernel eq. (20), a sample size of $N = 14{,}484$ would be needed with the Gaussian kernel eq. (13), under an ideal scenario.

In order to account for the fact that the ensemble covariance eq. (30) is not the true covariance eq. (11), we make use of a standard covariance tapering technique in the data assimilation literature, B-localization [2]. For all the filters we take a Gaussian decorrelation function with a localization radius of $r = 4$. More details about this technique can be found in [2], [20].

As the EnEMF weight update is not exact, we test two different choices of the scaling parameter $s_\mathcal{E}$ in eq. (39), namely $s_\mathcal{E} = 1$, which is the EnEMF with the full approximate weight update, and $s_\mathcal{E} = 1/2$, which should provide a middle ground between the EnEMF and the naive likelihood weights.

We additionally test against the ensemble Kalman filter (EnKF) [21], [22], specifically the linearized Jacobian variant described in [23], [24]. A heristic inflation [25] factor of $\alpha_{\text{inf}} = 1.01$ is applied for stability of the EnKF. As the EnKF makes use of only the first two statistical moments of the ensemble, and provides an almost linear update, it is a useful baseline for highly non-linear non-Gaussian sequential filtering problems.

We run all algorithms for 192 independent Monte Carlo simulations with $\Delta t = 0.2$ time units between measurements, corresponding roughly to a day in model time. The simulations are run for 2200 measurements, discarding the first 200 to account for spinup. The spatio-temporal root mean squared error (RMSE) is used as the error metric for determining the 'best' algorithm for the given problem, given by,

$$\text{RMSE}(\mathbf{x}^+) = \sqrt{\frac{1}{|\mathbf{x}|} \sum_i \|\bar{\mathbf{x}}_i^+ - \mathbf{x}_i^{\text{true}}\|_2^2}, \quad (46)$$

where the $\bar{\mathbf{x}}^+$ is the collection of posterior mean estimates, and $\mathbf{x}^t$ is the collection of true states. The algorithms are all

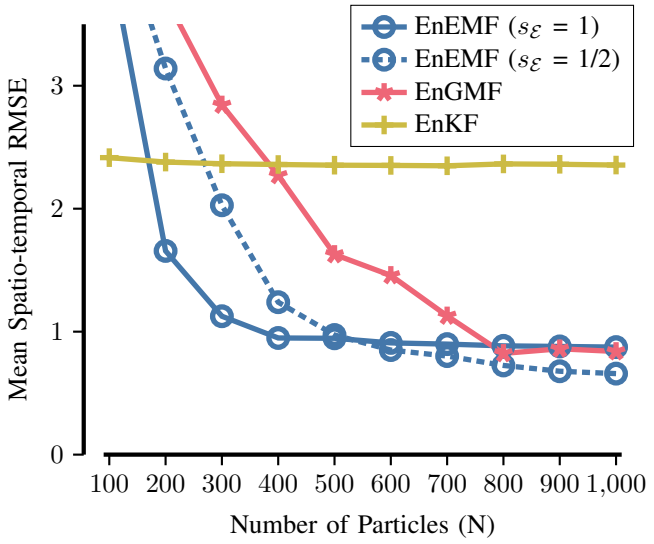


Fig. 5. Number of particles (N) versus spatio-temporal RMSE for the Lorenz '96 problem for the EnEMF (both with $s_\mathcal{E} = 1$ and $s_\mathcal{E} = 1/2$), EnGMF, and the EnKF

run for a varying degree of particle number N in the range of 100 to 1000.

The results of the experiment are provided in fig. 5. As can be seen, the linearized EnKF, which is an (almost) linear filter is at an error plateau for all tested particle numbers. The EnEMF reaches a significantly lower error plateau for $N = 400$ particles, with the EnGMF requiring over about $N = 800$. Additionally the EnEMF with $s_\mathcal{E} = 1/2$ reaches an even lower error plateau than all the algorithms, at a cost of slower convergence.

V. Conclusions

In this work we provide a theoretical derivation of the ensemble Epanechnikov mixture filter, and shown that it should outperform its cousin, the ensemble Gaussian mixture filter in the high-dimensional setting. We derive a practical implementation of the EnEMF that leverages the computational machinery of the EnGMF, such that the EnEMF can be implemented without significant computational overhead.

We additionally show through a numerical experiment that the EnEMF—while not attaining the theoretical error leaps over the EnGMF—still requires half the particles for the same level of error for a 40-variable problem. These results provide a promising path forward for making use of the Epanechnikov kernel in high-dimensional particle filtering applications.

Future work will focus on (i) proving that the EnEMF converges, in distribution, to exact Bayesian inference, and (ii) on providing more robust empirical results for filtering in the high-dimensional setting.

References

[1] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation: theory algorithms and software*. Ney York, NY, USA: Wiley, 2004.

[2] M. Asch, M. Bocquet, and M. Nodet, *Data assimilation: methods, algorithms, and applications*. SIAM, 2016.

[3] S. Reich and C. Cotter, *Probabilistic forecasting and Bayesian data assimilation*. Cambridge University Press, 2015.

[4] H. W. Sorenson and D. L. Alspach, "Recursive bayesian estimation using gaussian sums," *Automatica*, vol. 7, no. 4, pp. 465–479, 1971.

[5] G. Terejanu, P. Singla, T. Singh, and P. D. Scott, "Uncertainty propagation for nonlinear dynamic systems using gaussian mixture models," *Journal of guidance, control, and dynamics*, vol. 31, no. 6, pp. 1623–1633, 2008.

[6] F. Faubel, J. McDonough, and D. Klakow, "The split and merge unscented gaussian mixture filter," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 786–789, 2009.

[7] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, and P.-J. Nordlund, "Particle filters for positioning, navigation, and tracking," *IEEE Transactions on signal processing*, vol. 50, no. 2, pp. 425–437, 2002.

[8] F. Daum, J. Huang, and A. Noushin, "Gromov's method for Bayesian stochastic particle flow: A simple exact formula for q," in *2016 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, 2016, pp. 540–545.

[9] 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*, vol. 145, no. 723, pp. 2335–2365, 2019.

[10] 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*, vol. 127, no. 12, pp. 2741–2758, 1999.

[11] B. Liu, B. Ait-El-Fquih, and I. Hoteit, "Efficient kernel-based ensemble Gaussian mixture filtering," *Monthly Weather Review*, vol. 144, no. 2, pp. 781–800, 2016.

[12] A. A. Popov and R. Zanetti, "An adaptive covariance parameterization technique for the ensemble Gaussian mixture filter," *arXiv preprint arXiv:2212.10323*, 2022.

[13] ——, "Ensemble-localized kernel density estimation with applications to the ensemble gaussian mixture filter," *arXiv preprint arXiv:2308.14143*, 2023.

[14] J. K. Patel and C. B. Read, *Handbook of the normal distribution*. CRC Press, 1996, vol. 150.

[15] E. T. Jaynes, *Probability theory: The logic of science*. Cambridge, UK: Cambridge university press, 2003.

[16] B. W. Silverman, *Density estimation for statistics and data analysis*. London, UK: Chapman&Hall, 1986.

[17] P. Deheuvels, "Estimation non paramétrique de la densité par histogrammes généralisés (ii)," in *Annales de l'ISUP*, vol. 22, 1977, pp. 1–23.

[18] V. A. Epanechnikov, "Non-parametric estimation of a multivariate probability density," *Theory of Probability & Its Applications*, vol. 14, no. 1, pp. 153–158, 1969.

[19] L. Devroye and L. Györfi, *Nonparametric Density Estimation: the L_1 view*. Wiley, 1985.

[20] 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*, vol. 26, no. 2, pp. 109–122, 2019.

[21] G. Evensen, "Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics," *Journal of Geophysical Research: Oceans*, vol. 99, no. C5, pp. 10 143–10 162, 1994.

[22] G. Burgers, P. J. Van Leeuwen, and G. Evensen, "Analysis scheme in the ensemble kalman filter," *Monthly weather review*, vol. 126, no. 6, pp. 1719–1724, 1998.

[23] K. Michaelson, A. A. Popov, and R. Zanetti, "Bayesian recursive update for ensemble kalman filters," *arXiv preprint arXiv:2310.18442*, 2023.

[24] ——, "Ensemble kalman filter with bayesian recursive update," in *2023 26th International Conference on Information Fusion (FUSION)*. IEEE, 2023, pp. 1–6.

[25] A. A. Popov and A. Sandu, "An explicit probabilistic derivation of inflation in a scalar ensemble kalman filter for finite step, finite ensemble convergence," *arXiv preprint arXiv:2003.13162*, 2020.