Stability of nonlinear filters - numerical explorations of particle and ensemble Kalman filters

Pinak Mandal[1,*], Shashank Kumar Roy,[2,*] and Amit Apte[3,*]

***Abstract*— Particle filters and ensemble Kalman filters are widely used in data assimilation but in the case of deterministic systems, which are quite commonly used in earth science applications, only a few theoretical results for their stability are available. Current numerical literature explores stability in terms of RMSE which, although practical, can not represent the distance between probability measures, convergence of which is what defines filter stability. In this study, we explore the distance between filtering distributions starting from different initial distributions as a function of time using Wasserstein metric, thus directly assessing the stability of these filters. These experiments are conducted on the chaotic Lorenz-63 and Lorenz-96 models for various initial distributions for particle and ensemble Kalman filters. We show that even in cases when both these filters are stable, the filtering distributions given by each of them may be distinct.**

I. Introduction

One of the challenging problems in earth sciences is to incorporate the vast quantities of data that are constantly being collected world-wide into dynamical models for these systems, and is called the problem of data assimilation (DA). DA is a crucial ingredient for making meaningful real time predictions such as weather forecasts, hurricane tracking, and possibly even climate predictions. [1], [2] The Bayesian formulation of DA naturally leads to the problem of nonlinear filtering, which studies the conditional distribution, called the *filter* or the *posterior* distribution, of the state at any time conditioned on observations up to that time. [3], [4], [5]

A natural question is about the stability of the filter with respect to the initial condition, which is the probability distribution of the state at the initial time. This question has been studied extensively, e.g., [6], [7], but mostly in the context of stochastic systems. In many applications in the earth sciences, the models used are deterministic and only a few results about filter stability are known, see [8], [9] and the references therein. The main focus of this paper is to illustrate numerically the stability of two commonly used filtering algorithms, namely the particle filter and the ensemble Kalman filter.

The main contributions of this paper are as follows. The most commonly used method for assessing the filter stability is using the root-mean square error (RMSE) as the distance between the filter mean and the true trajectory. But such a measure at best implies stability of the mean. In this paper, we explicitly calculate distances between filtering distributions starting with different initial conditions, thus assessing stability directly (see the definition II.1). For this study, we apply the recently developed algorithms [10], [11], [12] for calculating optimal transport distances between distributions relevant to the filtering problem.

#This work was supported by the Department of Atomic Energy, Government of India, under project nos. 12-R&D-TFR-5.10-1100 and 12-R&D-TFR-5.01-0520.

*International Centre for Theoretical Sciences TIFR, Bangalore, India
[1]pinak.mandal@icts.res.in
[2]shashank.roy@icts.res.in
[3]apte@icts.res.in

II. Problem statement

A. The nonlinear filtering problem

The filtering problem studied in this paper can be stated as follows. The model state $x_k \in \mathbb{R}^d$ satisfies a discrete-time deterministic dynamical system $f : \mathbb{R}^d \to \mathbb{R}^d$ and measurement $y_k \in \mathbb{R}^q$ is related to the model state by the observation operator (linear throughout this paper) $H : \mathbb{R}^d \to \mathbb{R}^q$ for $k = 0, 1, \ldots$, as follows:

$$x_{k+1} = f(x_k), \quad x_0 \sim \mu, \tag{1}$$

$$y_k = Hx_k + \eta_k, \tag{2}$$

where μ is the initial distribution of the model state x_0 at time 0, and $\eta_k \sim \mathcal{N}(0_q, \sigma^2 I_q)$ are *iid* Gaussian errors in the observation, and are assumed independent of μ. Given observations $y_0, y_1, \cdots, y_n$, the goal of filtering is to estimate the conditional distribution of the model state at time n conditioned on observations up to that time, $\pi_n(\mu) := p(x_n|y_{0:n})$. We'll use different numerical algorithms to obtain an estimate, denoted by $\hat{\pi}_n(\mu)$, for the filtering distribution.

B. Filter stability

In practice we often do not know the initial distribution μ. In such a case, one obtains a different distribution, denoted by $\hat{\pi}_n(\nu)$, by using the same set of observations and using the same algorithm, but starting with a different initial condition ν. A measure of robustness of a filtering algorithm is how well it is able to "forget" the initial distribution, which motivates the following definition.

Definition II.1 (Stability)**.** *A numerical filter is said to be stable if given two different initial distributions ν_1, ν_2 for x_0 in the filtering problem, the following holds*

$$\lim_{n\to\infty} \mathbb{E}[D(\hat{\pi}_n(\nu_1), \hat{\pi}_n(\nu_2))] = 0 \tag{3}$$

where D is a distance on $P(\mathbb{R}^d)$, the space of probability measures on $\mathbb{R}^d$.

Note that the expectation above is taken with respect to observational noise since $\hat{\pi}_n$ is a random measure whose realizations correspond to observation realizations.

The main aim of this paper is to study the stability of two popular filtering algorithms, namely the particle filter described in III-B and the ensemble Kalman filter described in III-C, by studying the limit in (3), where we choose the Wasserstein metric W_2 as our distance D on $P(\mathbb{R}^d)$, for chaotic deterministic dynamical systems which we now describe in III-A.

III. Methodology

A. Models

We use two chaotic models in this paper: (i) Lorenz-63 [13, Chapter 14] with parameters $\rho = 28, \sigma = 10, \beta = \frac{8}{3}$ and (ii) 10 and 40-dimensional Lorenz-96 [14], [15] with forcing constant $F = 10$ and $F = 8$ respectively. We observe the system every 0.1 units of time which fixes the evolution function f. We observe alternate coordinates starting from the first coordinate, so

$$y_{k,j} = x_{k,2j-1} + \eta_{k,j} \tag{4}$$

for $j = 1, 2, \cdots, q = \lceil \frac{d}{2} \rceil$ and $\eta_{k,j} \sim \mathcal{N}(0, \sigma^2)$. Throughout the paper, we choose $\sigma^2 = 0.1$ or $\sigma^2 = 1.0$.

1) Data Generation: After selecting a model we find a point on the corresponding attractor by randomly generating an initial point and evolving it according to f for 10^5 iterations. Starting from this point x_0^{true} on the attractor, we generate a true trajectory according to (1) and then generate 10 different observation realizations for the same trajectory according to (4).

2) Initial distributions: We use three initial conditions:

$$\begin{aligned} \mu_1 &= \mathcal{N}(x_0^{\text{true}}, 0.1 \times I_d)\,, \\ \mu_2 &= \mathcal{N}(x_0^{\text{true}} + 2 \times 1_d, 0.5 \times I_d)\,, \\ \mu_3 &= \mathcal{N}(x_0^{\text{true}} + 4 \times 1_d, I_d)\,, \end{aligned} \tag{5}$$

where 1_d is a d-dimensional vector with all entries 1.

B. Particle Filter

We use the bootstrap particle filter (BPF), algorithm 1, resampling at every assimilation step. In order to avoid weight degeneracy in BPF for deterministic models, we incorporate post-regularization [16] in resampling steps by using an offspring based resampling strategy. We find the significant particles in the same manner as systematic re-sampling [17] and then generate offspring for each of the significant particles by placing a Gaussian distribution with covariance $\tilde{\sigma}^2 I_d$ around them. Number of offspring is set to be proportional to the weight of the particle.

In algorithm 1, we use the convention that $\sum_{l=1}^{0} w_k^l = 0$. It should be noted that larger values of $\tilde{\sigma}$ are needed to prevent filter collapse when working with fewer number of particles. In our experiments for Lorenz 63, $\tilde{\sigma}^2 = 0.1$ and for Lorenz 96, $\tilde{\sigma}^2 = 0.5$.

Algorithm 1: BPF with offspring-based resampling

Initialize N particles $\{x_0^i\}_{i=1}^N$ according to the initial distribution with equal weights $\left\{w_0^i = \frac{1}{N}\right\}_{i=1}^N$. Set $\tilde{\sigma}$. Below $S[i]$ denotes i-th element of S.

for $k = 0, \cdots, n$ **do**
- **if** $k > 0$ **then**
  - **for** $i = 1, \cdots, N$ **do**
    - $x_k^i \leftarrow f(S[i])$
- Sample $u \sim \mathcal{U}\left(0, \frac{1}{N}\right)$
- **for** $i = 1, \cdots, N$ **do**
  - $w_k^i \leftarrow p(y_k | x_k^i)$
  - $U_i \leftarrow u + \frac{i-1}{N}$
- $W \leftarrow \sum_{i=1}^N w_k^i$
- **for** $i = 1, \cdots, N$ **do**
  - **if** $|\{U_j : \sum_{l=1}^{i-1} w_k^l \le W U_j \le \sum_{l=1}^{i} w_k^l\}| > 0$ **then**
    - tag x_k^i as significant
- Set $S \leftarrow \{x_k^{i_1}, x_k^{i_2}, \cdots, x_k^{i_m}\}$ as the set of significant particles and compute $N_j \propto w_k^{i_j} : \sum_{j=1}^m N_j = N$.
- **for** $j = 1, \cdots, m$ **do**
  - $S \leftarrow S \cup \{N_j - 1$ samples from $\mathcal{N}(x_k^{i_j}, \tilde{\sigma}^2 I_d)\}$
- $\hat{\pi}_k \leftarrow \frac{1}{N} \sum_{i=1}^N \delta_{S[i]}$

Algorithm 2: EnKF with covariance localization

Initialize N particles $\{x_0^i\}_{i=1}^N$ according to the initial distribution and set $x_0^{i,a} = x_0^i$

for $k = 1, \cdots, n$ **do**
- **for** $i = 1, \cdots, N$ **do**
  - $x_k^{i,f} \leftarrow f(x_{k-1}^{i,a})$
- $m_k^f \leftarrow \frac{1}{N} \sum_i x_k^{i,f}$
- $P_k^f \leftarrow \rho \circ \frac{\sum_i \left(x_k^{i,f} - m_k^f\right)\left(x_k^{i,f} - m_k^f\right)^\top}{N-1}$
- $K \leftarrow P_k^f H^T \left[H P_k^f H^T + R_k\right]^{-1}$
- **for** $i = 1, \cdots, N$ **do**
  - Sample $\eta_k^i \sim \mathcal{N}(0_q, \sigma^2 I_q)$
  - $y_k^i \leftarrow y_k + \eta_k^i$
  - $x_k^{i,a} \leftarrow x_k^{i,f} + K\left[y_k^i - H x_k^{i,f}\right]$
- $\hat{\pi}_k \leftarrow \frac{1}{N} \sum_{i=1}^N \delta_{x_k^{i,a}}$

C. Ensemble Kalman filter (EnKF)

We also use EnKF [18] with observation perturbation, a Monte-Carlo method based on Kalman filter, where the mean and covariance are estimated from the ensemble. Covariance localization is implemented by constructing localization matrix (ρ in algorithm 2) using Gaspari-Cohn localization function with localization radius 4 to prevent filter divergence in high dimensions for small ensemble sizes [16].

Algorithm 3: Computation of S_ε

Note the definition, $\text{LSE}_{k=1}^{L} V_k \stackrel{\text{def}}{=} \log \sum_{k=1}^{L} \exp(V_k)$.
Initialize $a_i \leftarrow 0 \; \forall \; i = 1, \cdots, N$ and
$b_j \leftarrow 0, \; \forall \; j = 1, \cdots, M$. Set T.
iteration $\leftarrow 0$
while L_1 *relative error in* $a > 0.1\%$ *and iteration* $< T$ **do**
  for $i = 1, \cdots, N$ **do**
    $a_i \leftarrow -\varepsilon \text{LSE}_{k=1}^{M} \left(\log \nu_k + \frac{1}{\varepsilon} b_k - \frac{1}{\varepsilon} \|x_i - y_k\|_2^2 \right)$
  for $j = 1, \cdots, M$ **do**
    $b_j \leftarrow -\varepsilon \text{LSE}_{k=1}^{N} \left(\log \mu_k + \frac{1}{\varepsilon} a_k - \frac{1}{\varepsilon} \|x_k - y_j\|_2^2 \right)$
  iteration $\leftarrow$ iteration + 1
$\text{OT}_{\mu,\nu} \leftarrow \sum_{i=1}^{N} \mu_i a_i + \sum_{j=1}^{M} \nu_j b_j$
Initialize $a_i \leftarrow 0 \; \forall \; i = 1, \cdots, N$ and
$b_j \leftarrow 0, \; \forall \; j = 1, \cdots, M$.
while L_1 *relative error in* $a > 0.1\%$ **do**
  for $i = 1, \cdots, N$ **do**
    $a_i \leftarrow \frac{1}{2} \left[a_i - \varepsilon \text{LSE}_{k=1}^{N} \left(\log \mu_k + \frac{1}{\varepsilon} a_k - \frac{1}{\varepsilon} \|x_i - x_k\|_2^2 \right) \right]$
while L_1 *relative error in* $b > 0.1\%$ **do**
  for $j = 1, \cdots, M$ **do**
    $b_j \leftarrow \frac{1}{2} \left[b_j - \varepsilon \text{LSE}_{k=1}^{M} \left(\log \nu_k + \frac{1}{\varepsilon} b_k - \frac{1}{\varepsilon} \|y_j - y_k\|_2^2 \right) \right]$
$S_\varepsilon \leftarrow \text{OT}_{\mu,\nu} - \sum_{i=1}^{N} \mu_i a_i - \sum_{j=1}^{M} \nu_j b_j$

D. Sinkhorn divergence

We choose the distance on $P(\mathbb{R}^d)$ in (3) to be the Wasserstein metric W_2. We approximate W_2 [see (8) later] by the Sinkhorn divergence S_ε defined in (7), which in turn needs optimal transport distance OT_ε defined in (6).

Computation of optimal transport distances between probability measures has been a challenging task until recently. But entropy regularization has made solving the dual to the optimal transport problem tractable for sampling distributions through the use of Sinkhorn algorithm [11, and references therein]. To estimate W_2 we use Sinkhorn divergence S_ε defined as follows [10],

$$\text{OT}_\varepsilon(\mu, \nu) \stackrel{\text{def}}{=} \min_{\pi \in \mathbb{S}} \left[\int \|x - y\|_2^2 \, d\pi(x, y) + \varepsilon \text{KL}(\pi | \mu \otimes \nu) \right] \tag{6}$$

$$\text{S}_\varepsilon \stackrel{\text{def}}{=} \text{OT}_\varepsilon(\mu, \nu) - \frac{1}{2}\text{OT}_\varepsilon(\mu, \mu) - \frac{1}{2}\text{OT}_\varepsilon(\nu, \nu) \tag{7}$$

where the minimisation is over the set $\mathbb{S}$ of distributions π with the first and second marginals being μ and ν respectively and KL is the Kullback–Leibler divergence. We compute $S_\varepsilon(\mu, \nu)$ for probability measures of form $\mu = \sum_{i=1}^{N} \mu_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{M} \nu_j \delta_{y_j}$ using algorithm 3 from [10]. In our experiments $\mu_i = \frac{1}{N}$ and $\nu_j = \frac{1}{M} \; \forall \; i, j$. We use $\varepsilon = 0.01$ and $T = 200$ for all our experiments. Note that

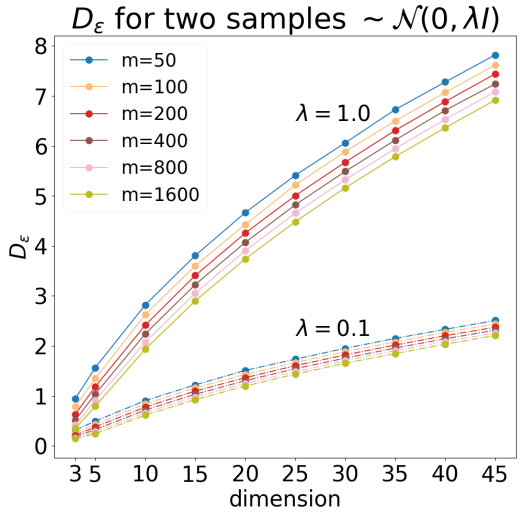

Fig. 1: Average $D_\varepsilon(\alpha_m^d, \beta_m^d)$ (over 20 realizations) where α_m^d, β_m^d are two different sampling distributions with the same sample size m for the same underlying d-dimensional Gaussian $\mathcal{N}(0_d, \lambda I_d)$

[11], [19]

$$\lim_{\varepsilon \to 0} \sqrt{S_\varepsilon(\mu, \nu)} = W_2(\mu, \nu) \,, \tag{8}$$

and $\varepsilon = 0.01$ is a decent practical choice for estimating W_2. We use the notation $D_\varepsilon := \sqrt{S_\varepsilon}$ in the following sections. Although KL-divergence can be estimated with k-nearest neighbors based algorithms [20], this requires choosing the parameter k. D_ε also requires choosing ε but in this case convergence of D_ε to W_2 as $\varepsilon \to 0$ and the appearance of division by ε in the Sinkhorn algorithm are two opposing forces that help us come up with a reasonable choice. In absence of these constraints for KL-divergence, choosing k becomes unmotivated, not to mention the estimated KL-divergence is more sensitive to the choice of k than D_ε is to the choice of ε. Combined with this practical nuance, the nice geometric properties of W_2 e.g. metrizing the convergence in law [10], make it preferable to KL-divergence.

IV. Main Results

We now discuss the stability of PF and EnKF by calculating $\mathbb{E}[D_\varepsilon(\hat{\pi}_n(\mu_i), \hat{\pi}_n(\mu_j))], \; i \neq j$ as a function of time n for initial conditions μ_i from (5), with the expectation taken by averaging over 10 observation realizations. For clarity, the above quantity is shown at every 4-th assimilation step in figures 2–4.

A. Zero of the Sinkhorn algorithm

In order to understand the convergence to 0 of the above quantity [see (3)], we first discuss how close to zero D_ε can approach numerically. In figure 1 we see the average $D_\varepsilon(\alpha_m^d, \beta_m^d)$ where $\alpha_m^d = \frac{1}{m} \sum_{i=1}^{m} \delta_{x_i^{m,d}}$ and $\beta_m^d = \frac{1}{m} \sum_{i=1}^{m} \delta_{y_i^{m,d}}$, with $\{x_i^{m,d}\}$ and $\{y_i^{m,d}\}$ both samples from the same underlying d-dimensional Gaussian distribution $\mathcal{N}_d^\lambda := \mathcal{N}(0_d, \lambda I_d)$. For 'small' λ, we can expect D_ε to

behave in a similar fashion as if $\mathcal{N}_d^\lambda$ were supported on a compact set. With that in mind, we relate the numerical results shown in figure 1 to the results VI.1–VI.3 in the appendix VI by noting the following key points:

1) Drop with increase in sample size: Theorem VI.3 explains the monotone drop in average D_ε for a fixed dimension while increasing the sample size.

2) Rise with increase in dimension: As the dimension increases, larger sample sizes are required to accurately estimate $\mathcal{N}_d^\lambda$. Consequently, $D_\varepsilon(\alpha_m^d, \beta_m^d)$ grows with d for fixed m since α_m^d, β_m^d become poorer estimators of $\mathcal{N}_d^\lambda$ as d increases.

3) Drop with decrease in covariance: Decreasing the covariance λ has the opposite effect since, for fixed dimension d and sample size m, smaller covariance leads to a better estimation of the underlying distribution, i.e., α_m^d, β_m^d become better estimators of $\mathcal{N}_d^\lambda$ as λ decreases.

4) Support of our distributions: Since the true trajectories for both systems (L63, L96) lie on bounded attractors, we can assume that true filtering distributions are supported on a compact set. Consequently, in the filtering experiments shown later, the zero of the Sinkhorn algorithm shows qualitatively similar behavior (e.g., in figure 2) with respect to dimension as seen in figure 1.

B. Particle Filter

Here we use the notation $\pi_n^{P,N}$ for $\hat{\pi}_n$ obtained by alogithm 1 with N particles (omitting N for brevity when value of N is clear from context). Figure 2 shows $\mathbb{E}[D_\varepsilon(\hat{\pi}_n(\mu_i), \hat{\pi}_n(\mu_j))],\ i \neq j$ as a function of n. We note some important conclusions.

1) BPF quickly forgets the initial distribution: From the insets in figure 2 we can see that for every pair (μ_i, μ_j) of initial distributions, $\mathbb{E}\left[D_\varepsilon(\pi_n^P(\mu_i), \pi_n^P(\mu_j))\right]$ stabilizes in the first few assimilation steps. In fact, this behavior is consistent with exponential stability of particle filters [21].

2) Dependence on the number of particles: $\mathbb{E}\left[D_\varepsilon(\pi_n^{P,N}(\mu_i), \pi_n^{P,N}(\mu_j))\right]$ for a fixed n decreases monotonically with increasing N for both L63 and L96 and for both observation covariances for all pairs $i \neq j$.

3) Stability: Suppose the best possible filtering distribution that can be computed by the particle filter is $\pi_n^{P,*} = \lim_{N\to\infty} \pi_n^{P,N}$. Figure 2 is consistent with the condition

$$\lim_{n\to\infty} \liminf_{N\to\infty} \mathbb{E}[D_\varepsilon(\pi_n^{P,N}(\mu_i), \pi_n^{P,N}(\mu_j))] = 0 \ \forall\ i \neq j$$

since fixing n and increasing N results in a steady drop in D_ε averaged over observation realizations. By theorem VI.4 this condition is sufficient for concluding

$$\lim_{n\to\infty} \mathbb{E}[D_\varepsilon(\pi_n^{P,*}(\mu_i), \pi_n^{P,*}(\mu_j))] = 0 \ \forall\ i \neq j$$

4) Dependence on observation covariance: All plots in figure 2 correspond to observation covariance $\sigma^2 = 0.1$. The other case $\sigma^2 = 1.0$ mentioned in III-A results in plots that are qualitatively similar to the ones in figure 2 and we omit those plots here. In our experiments, stability of particle filter was not seen to be affected by observation covariance.

C. EnKF

Here we use the notation $\pi_n^{E,N}$ for $\hat{\pi}_n$ obtained by alogithm 2 with ensemble size N. We might omit N for brevity.

1) Drop in D_ε over time: From figure 3, we see that for every pair (μ_i, μ_j) of initial distributions $D_\varepsilon(\pi_n^E(\mu_i), \pi_n^E(\mu_j))$ decreases with time rapidly within the first 50 assimilation steps and beyond 100 assimilation steps, the observation average of D_ε for filters with different pairs initial distributions are similar and have very little variance.

2) Variation with respect to observation realization: We see that the variation of D_ε for different observation realizations (shown by the shaded bands in left two panels in figure 3 and in top row in figure 2) is larger for the case of EnKF when compared to the particle filter for initial times (e.g. $n < 100$ for the 10-dimensional L96 model). On the other hand, for larger times (approx. $n > 100$), the variation for EnKF is significantly smaller than the particle filter.

3) Effect of localization: EnKF with small ensemble size needs localization which, however, is an ad-hoc procedure to prevent filter divergence and may not approximate the true filter. Figure 3 for 10-dimensional L96 (left panel) and for 40-dimensional L96 (right panel) shows that for $N = 50$ with localization length 4, the EnKF is stable, whereas the middle panel shows the stability (with the same configuration as the left panel) for 10-dimensional L96 without localization, but with larger ensemble size $N = 200$. This indicates that that localization does not affect EnKF's stability properties.

D. BPF vs EnKF

We now compare the BPF and EnKF for the case of 10-dimensional L96 with $\sigma^2 = 1.0$ with the same true trajectory and observation realizations. This is shown in figure 4. In the following discussion we assume BPF with 2000 particles to be a decent approximation for the true filter and refer to them interchangably. We note a few important points.

1) Poor approximation of the true filter by EnKF: The three lines towards the top show the distance between EnKF and BPF, for three different initial conditions. We see that EnKF produces distributions that are significantly different from the true filter, for all the initial distributions. But recall that for this setup, the EnKF is stable (as is the BPF too), i.e., the distance between EnKF with different initial conditions is smaller (as seen in figures 2-3) than the distance between the BPF and EnKF

2) BPF is closer to the true filter than is EnKF: The three lines in the middle show the distance between BPF with $N = 250$ and $N = 2000$ (putatively true filter). We see that in comparison with EnKF with $N = 200$ particles, BPF with similar ensemble size (250 particles) is much closer to the true filter.

3) EnKF with different ensemble size are very similar: The bottom three lines in figure 4 show D_ε between EnKF with different ensemble size, which shows that the EnKF is quite stable with respect to changes in the ensemble size, even though it is not very close to the true filter – thus EnKF is stable but biased, whereas BPF is stable and unbiased. A

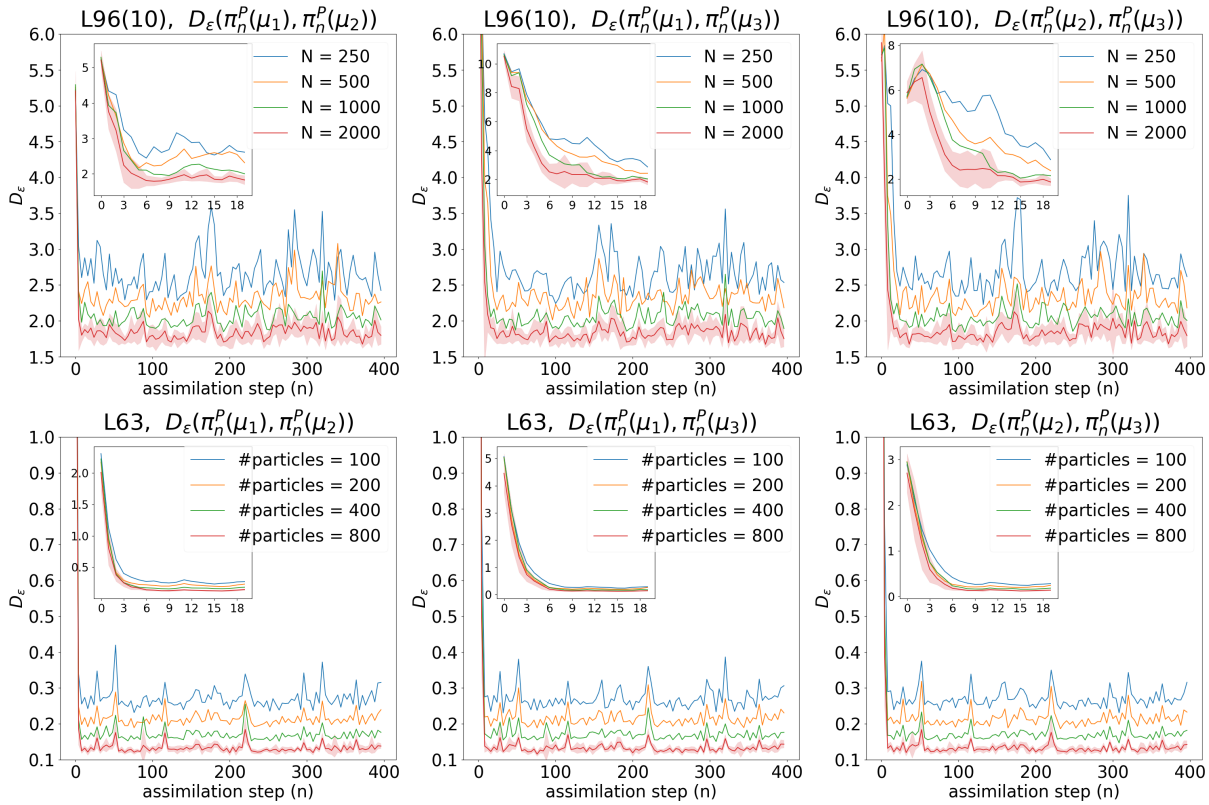

Fig. 2: D_ε (averaged over 10 observation realizations) for BPF for 10-dimensional $L96$ (row 1) and $L63$ (row 2) systems with observation covariance $\sigma^2 = 0.1$, for pairs of initial distributions in (5), with varying sample size. The line for $N = 2000$ has a band showing one standard deviation. The inset shows the drop in average D_ε during the first few assimilation steps.

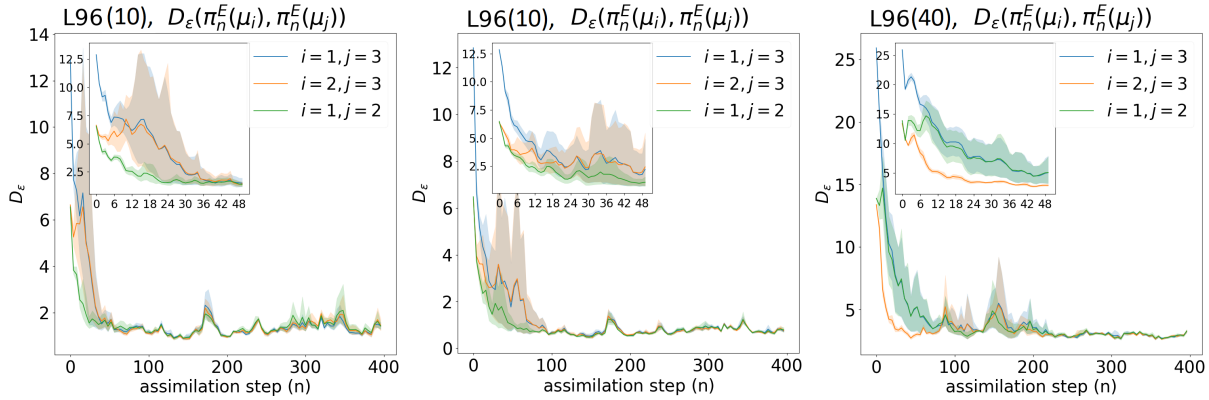

Fig. 3: D_ε (averaged over 10 observation realizations, with one standard deviation confidence band) for EnKF for 10-dimensional L96 with $N = 50$ with localization (left), $N = 200$ without localization (middle) for observation covariance $\sigma^2 = 0.1$, and for 40-dimensional L96 with $N = 50$ with localization (right) with observation covariance $\sigma^2 = 1.0$ for pairs of initial distributions in 5. The inset shows the drop in D_ε for the first 50 assimilation steps.

more detailed study of the reasons for this behaviour will be taken up in the future.

V. Discussion

The main focus of this study was to develop a novel methodology to assess nonlinear filter stability using recently introduced techniques for calculating Wasserstein distances between distributions. We show that the particle filter and ensemble Kalman filter are stable when applied to a wide range of chaotic, deterministic dynamical systems, but the EnKF fails to capture the true filtering distribution.

We expect that the use of numerical algorithms for computing distances between the distributions will be a powerful tool for understanding nonlinear filters, leading to several avenues for further exploration. One direction of particular interest in earth sciences is to examine the relation between

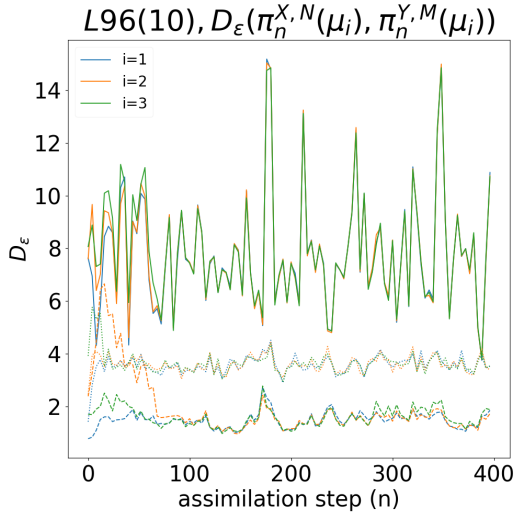

Fig. 4: Comparison between filters for 10-dimensional L96 and $\sigma^2 = 1.0$. The solid lines on the top show average D_ε between EnKF without localization with $N = 200$ and BPF with $M = 2000$. The dotted lines in the middle show average D_ε between BPF with $N = 250$ and BPF with $M = 2000$. The dashed lines at the bottom show average D_ε between EnKF without localization with $N = 200$ and EnKF with localization with $M = 50$. In each case, different colors are for different initial conditions from (5).

filter stability and the dynamical properties of the systems being observed.

VI. Appendix: Properties of Sinkhorn Divergence

Lemma VI.1. *If α, β, α_m are probability measures on a compact set $\chi \subset \mathbb{R}^d$ then*

$$0 = S_\varepsilon(\beta, \beta) \leq S_\varepsilon(\alpha, \beta) \tag{9}$$

$$\alpha = \beta \iff S_\varepsilon(\alpha, \beta) = 0 \tag{10}$$

$$\alpha_m \xrightarrow{weak^*} \alpha \iff S_\varepsilon(\alpha_m, \alpha) \to 0 \tag{11}$$

Proof: See Theorem 1 in [10].

Lemma VI.2. *If $\alpha_m, \beta_m, \alpha, \beta$ are probability measures supported on a compact set $\chi \subset \mathbb{R}^d$ such that $\alpha_m \xrightarrow{weak^*} \alpha$ and $\beta_m \xrightarrow{weak^*} \beta$ then $\lim_{m\to\infty} S_\varepsilon(\alpha_m, \beta_m) = S_\varepsilon(\alpha, \beta)$.*

Proof: Direct consequence of proposition 13 in [10].

Theorem VI.3. *If $\alpha_m = \frac{1}{m}\sum_{i=1}^m \delta_{x_i^m}$ and $\beta_m = \frac{1}{m}\sum_{i=1}^m \delta_{y_i^m}$ are sampling distributions for the same underlying probability distribution μ which is supported on a compact set $\chi \subset \mathbb{R}^d$ then $\lim_{m\to\infty} S_\varepsilon(\alpha_m, \beta_m) = 0$.*

Proof: Direct consequence of lemmas VI.1 and VI.2.

Theorem VI.4. *If $\alpha_{m,n}, \beta_{m,n}, \alpha_n, \beta_n$ are random probability measures supported on a compact set $\chi \subset \mathbb{R}^d$ such that $\alpha_{m,n} \xrightarrow{weak^*} \alpha_n$ and $\beta_{m,n} \xrightarrow{weak^*} \beta_n$ as $m \to \infty$ and*

$$\lim_{n\to\infty} \liminf_{m\to\infty} \mathbb{E}[D_\varepsilon(\alpha_{m,n}, \beta_{m,n})] = 0$$

then, $\lim_{n\to\infty} \mathbb{E}[D_\varepsilon(\alpha_n, \beta_n)] = 0$.

Proof:

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}[D_\varepsilon(\alpha_n, \beta_n)] \\
&= \lim_{n\to\infty} \mathbb{E}[\lim_{m\to\infty} D_\varepsilon(\alpha_{m,n}, \beta_{m,n})] \quad \textit{by lemma VI.2} \\
&\leq \lim_{n\to\infty} \liminf_{m\to\infty} \mathbb{E}[D_\varepsilon(\alpha_{m,n}, \beta_{m,n})] = 0 \quad \textit{by Fatou's lemma}
\end{aligned}$$

References

[1] M. Asch, M. Bocquet, and M. Nodet, *Data Assimilation: Methods, Algorithms, and Applications*. SIAM, 2016.

[2] A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen, "Data assimilation in the geosciences: An overview of methods, issues, and perspectives," *Wiley Interdisciplinary Reviews: Climate Change*, vol. 9, no. 5, p. e535, 2018.

[3] A. Apte, M. Hairer, A. Stuart, and J. Voss, "Sampling the posterior: an approach to non-Gaussian data assimilation," *Physica D*, vol. 230, pp. 50–64, 2007.

[4] K. Law, A. Stuart, and K. Zygalakis, *Data Assimilation*. Springer, 2015.

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

[6] A. N. Bishop and P. Del Moral, "On the stability of kalman–bucy diffusion processes," *SIAM Journal on Control and Optimization*, vol. 55, no. 6, pp. 4015–4047, 2017.

[7] P. Chigansky, "Stability of nonlinear filters: A survey," 2006, lecture notes, Petropolis, Brazil.

[8] A. S. Reddy, A. Apte, and S. Vadlamani, "Asymptotic properties of linear filter for noise free dynamical system," *Systems & Control Letters*, vol. 139, p. 104676, 2020.

[9] A. S. Reddy and A. Apte, "Stability of non-linear filter for deterministic dynamics," *Foundations of Data Science*, vol. 3, no. 3, pp. 647–675, 2021.

[10] J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouvé, and G. Peyré, "Interpolating between optimal transport and mmd using sinkhorn divergences," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 2681–2690.

[11] A. Genevay, "Entropy-regularized optimal transport for machine learning," Ph.D. dissertation, Paris Sciences et Lettres, 2019.

[12] A. Thibault, L. Chizat, C. Dossal, and N. Papadakis, "Overrelaxed sinkhorn–knopp algorithm for regularized optimal transport," *Algorithms*, vol. 14, no. 5, p. 143, 2021.

[13] M. W. Hirsch, S. Smale, and R. L. Devaney, *Differential equations, dynamical systems, and an introduction to chaos*. Academic press, 2012.

[14] E. N. Lorenz, "Predictability: a problem partly solved," in *Seminar on Predictability I*. ECMWF, Reading UK, 1995, pp. 1–18.

[15] D. van Kekem, "Dynamics of the Lorenz-96 model: Bifurcations, symmetries and waves," Ph.D. dissertation, University of Groningen, 2018.

[16] A. Farchi and M. Bocquet, "Comparison of local particle filters and new implementations," *Nonlinear Processes in Geophysics*, vol. 25, no. 4, pp. 765–807, 2018.

[17] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later," *Handbook of nonlinear filtering*, vol. 12, no. 656-704, p. 3, 2009.

[18] G. Evensen, *Data assimilation: the ensemble Kalman filter*. Springer, 2007.

[19] G. Carlier, V. Duval, G. Peyré, and B. Schmitzer, "Convergence of entropic schemes for optimal transport and gradient flows," *SIAM Journal on Mathematical Analysis*, vol. 49, no. 2, pp. 1385–1418, 2017.

[20] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation for multidimensional densities via k-nearest-neighbor distances," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2392–2405, 2009.

[21] P. Chigansky, R. Liptser, and R. Van Handel, "Intrinsic methods in filter stability," *Handbook of Nonlinear Filtering*, 2009.