Unbiased Fully Nonlinear Data Assimilation: the Stochastic Particle Flow Smoother

(29th IOAS-AOLS Poster Session II, #383)

105th AMS Annual Meeting
Tue., 14 Jan. 2025; 15:00 - 16:30 CT

Hao-Lun Yeh and **Peter Jan van Leeuwen**

Department of Atmospheric Science, Colorado State University, Colorado, USA

Email : haolun.Yeh@colostate.edu

CADRE Consortium for Advanced Data Assimilation Research and Education

1. Introduction

Background

The significance of **nonlinearities** in geosciences numerical models and in observation operators demands more advanced data assimilation techniques. **The Particle Flow Filter (PFF)** is a fully nonlinear and efficient sequential Monte Carlo filter that removes the weight degeneracy in particle filters. The **deterministic** version of the PPF has been successfully applied to high-dimensional systems. With a small number of particles, however, it can exhibit **low-bias ensemble spread**, particularly in observed parts of the state space.

Research Goals

To address this challenge, we extend the PFF in two ways. First, we developed a **stochastic** version incorporating Gaussian noise into the Stein Variational Gradient Descent dynamics, ensuring **unbiased ensemble spread** for any ensemble size. Second, we extended the filter to a **smoother**, resulting in the Stochastic Particle Flow Smoother (SPFS), allowing it to solve the fully nonlinear smoother problem.

We first evaluate the performance of the PFF, its stochastic version and smoother version in a high-dimensional Lorenz 96 model. The goal is to apply this filter (or smoother) in the real-world ocean model MOM6.

2. Methodology

A. The Deterministic Particle Flow Filter (PFF)

(Pulido and van Leeuwen, 2019; Hu and van Leeuwen, 2021)

- The PFF utilizes the idea of particle flow that pushes forward an ensemble of N_p model states $\{\mathbf{x}^i\}_{i=1}^{N_p}$ (so called particles) in a series of discrete steps from the prior $p(\mathbf{x})$ to the samples from the posterior $p(\mathbf{x}|\mathbf{y})$, given the observation vector $\mathbf{y}$ (See Fig. 1). The PFF updates each particle via a deterministic flow and can be described by a set of ordinary coupled differential equation evolving over pseudo-time s

$$d\mathbf{x}_s^i = f(\mathbf{x}_s^i)ds$$

$$f(\mathbf{x}_s^i) = \frac{1}{N_p}\sum_{j=1}^{N_p}\left[\mathbf{K}(\mathbf{x}_s^j,\mathbf{x}_s^i)\nabla_{\mathbf{x}_s^j}\log p(\mathbf{x}_s^j|\mathbf{y}) + \nabla_{\mathbf{x}_s^j}\cdot\mathbf{K}(\mathbf{x}_s^j,\mathbf{x}_s^i)\right]$$

(Attracting term; Repelling term)

where $\mathbf{K}$ is a kernel matrix. The kernel matrix characterizes the distance between any two particles. A diagonal and isotropic **scalar Gaussian kernel** is adopted

$$\mathbf{K}(\mathbf{x}_s^j,\mathbf{x}_s^i) = K(\mathbf{x}_s^j,\mathbf{x}_s^i)\mathbf{I}_{n_x}$$

$$K(\mathbf{x}_s^j,\mathbf{x}_s^i) = \exp\left(-\frac{1}{2}(\mathbf{x}_s^j-\mathbf{x}_s^i)^T(\alpha\mathbf{B})^{-1}(\mathbf{x}_s^j-\mathbf{x}_s^i)\right)$$

α : Kernel width
$\mathbf{B}$: Background error covariance
$\mathbf{I}_{n_x}$: Identity matrix of size $n_x \times n_x$

- The PFF is not limited by any constraints in the prior distribution and linearity assumptions. It corresponds with ensemble variational DA methods, but **particles can interact with each other** through the kernel matrix during the update.

B. The Stochastic Particle Flow Filter (SPFF)

- Gallego and Insua (2020) prove that PFF with **Gaussian noise term** can be viewed as a multiple-chain **Stochastic Gradient Markov Chain Monte Carlo** (SG-MCMC) sampler, ensuring the algorithm asymptotically converges to the truly "correct" stationary posterior distribution → **No bias**.
- The evolution of SPFF can be described as multiple continuous-time Markov processes generated by a stochastic differential equation (SDE)

$$d\mathbf{x}_s^i = f(\mathbf{x}_s^i)ds + (\sqrt{2}\mathbf{P}^T\mathbf{L}_{\mathbf{K_r}}N(0,\mathbf{I}))^i\sqrt{ds}$$

(Noise term)

where $\mathbf{P}$ is a permutation matrix, $N(0,\mathbf{I})$ represents a normal distribution with a mean of 0 and covariance $\mathbf{I}$ (identity matrix), and $\mathbf{L}_{\mathbf{K_r}}$ is a block-diagonal matrix where each block is the lower triangular Cholesky decomposition of the kernel Gram matrix, $\bar{\mathbf{k}}$, repeated n_x times along the diagonal. Here, $\bar{\mathbf{k}}(i,j) = K(\mathbf{x}_s^i,\mathbf{x}_s^j)$.

- With a finite number of particles and a limited update time, the SPFF remains unbiased due to its direct link with the unbiased Langevin equation.

C. The Stochastic Particle Flow Smoother (SPFS)

- The SPFF can be further extended to a smoother that:
  1. Enhances the accuracy of historical state estimates, and
  2. Has smaller error inside an assimilation window compared to the filter
- The update of SPFS is the same as SPFF, with only difference in the gradient of the logarithm of posterior: ($\mathbf{R}$: observation error covariance; H: observation operator)

For Filter: $\mathbf{H}^{j^T}\mathbf{R}^{-1}(\mathbf{y}-H(\mathbf{x}_s^j)) - \mathbf{B}^{-1}(\mathbf{x}_s^j-\bar{\mathbf{x}}_\mathbf{b})$

For Smoother: $\sum_{t=1}^{T}\left\{\tilde{\mathbf{H}}_t^{j^T}\mathbf{R}^{-1}(\mathbf{y}_t-H(\mathbf{x}_{s,t}^j))\right\} - \mathbf{B}^{-1}(\mathbf{x}_s^j-\bar{\mathbf{x}}_\mathbf{b})$

- SPFS adopts a **strong-constraint** perfect model assumption, so $\tilde{\mathbf{H}}_t^{j^T} = \mathbf{M}_{0\to t}{}^T\mathbf{H}^{j^T}$, where $\mathbf{M}_{0\to t}{}^T$ is an **adjoint model** and $\mathbf{H}$ is a linearized observation operator.
- The SPFS has computational similarities with **the Ensemble of 4DVars** during minimization but utilizes interacting particle systems.

3. Lorenz 96 Model Results

1000-dimension system, with 20 ensemble members

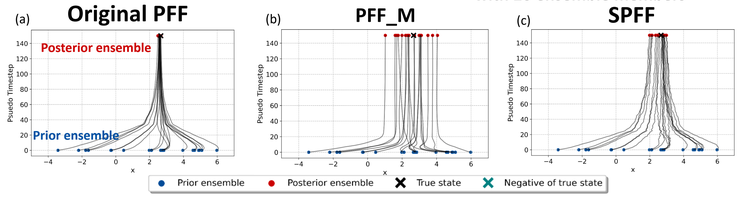


Fig. 1. Evolution of particles during PFF update. (a) PFF using a scalar Kernel. (b) PFF using a matrix-valued Kernel. (c) SPFF using a scalar Kernel.

- In Original PFF, particles may collapse to the mode of the posterior, while the PFF with a matrix-valued kernel (PFF_M) avoids particle collapse in high-dimensional systems (Hu and van Leeuwen, 2021) but its ensemble spread can be biased-high with finite ensemble size (Fig. 1a, 1b).
- Though using the same scalar kernel as Original PFF, SPFF (Fig. 1c) can prevent particles collapse by adding small random noises in each iteration and maintain the spread of the particles.

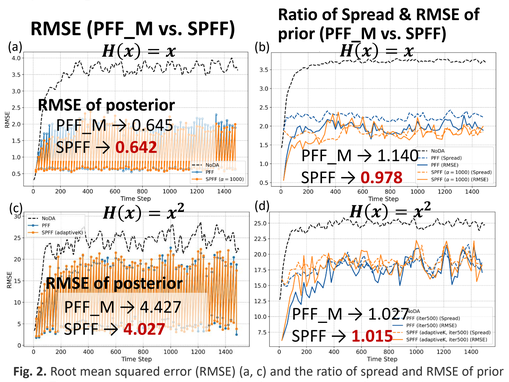


Fig. 2. Root mean squared error (RMSE) (a, c) and the ratio of spread and RMSE of prior (b, d). The upper panel shows results using linear operator, while the bottom shows the ones using square operator (Blue: PFF_M; Orange: SPFF)

- The SPFF has a comparable performance with PFF_M, but with **less bias** in spread (Fig. 2a, 2b).
- For the nonlinear (square) operator, both PFF_M and SPFF can capture the **bimodal distribution**, but SPFF has less biased spread and RMSE (due to more efficient sampling) (Fig. 2c, 2d, 3b, 3c).
- Compared with PFF_M and SPFF, the SPFS reduces the error throughout the time interval (Fig. 3a).
- For nonlinear operator, SPFS not only successfully captures bimodal distribution, but better simulates evolution of the system with smooth changes (Fig. 3b, 3c, 3d).

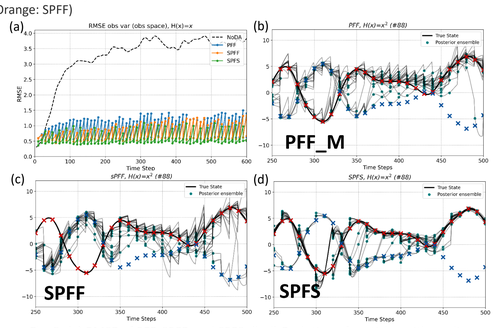


Fig. 3. (a) RMSE of PFF, SPFF, and SPFS. (b-d) Evolution of particles with nonlinear operator from (b) PFF_M, (c) SPFF, and (d) SPFS.

4. The MOM6 Model Results

A. The Model Configurations

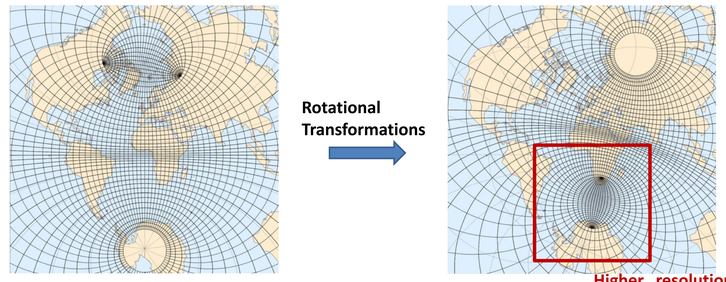


Fig. 4. The MOM6 grid setting

- **Ocean model:** Modular Ocean Model 6 **(MOM6)** with **OM4 1/4-degree grid**
- **Assimilated Observation:** ocean satellite data, XBT and Argo float data, and mooring data
- **Assimilation system:** JEDI-SOCA

B. Preliminary Results

- Due to technical issues, we have not yet been able to run the PFF in JEDI-SOCA system, and all the experiments use the **OM4 1/4-degree grid**.
- We use an LETKF with 15 ensemble members.
- Assimilated Observations: Metop-B Sea Surface Temperature and Cryosat-2 absolute dynamic topography.
- Even with 1/4-degree resolution, the smaller-scale structures (eddies) are captured due to the data assimilation.
- The SSH assimilation leads to a much stronger Agulhas and Somalia Currents and tropical perturbations.

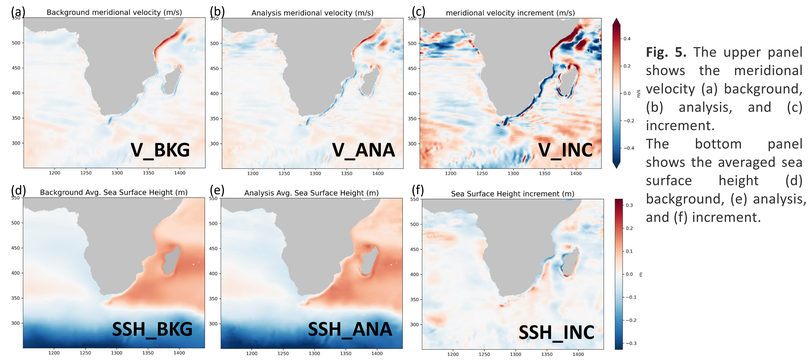


Fig. 5. The upper panel shows the meridional velocity (a) background, (b) analysis, and (c) increment. The bottom panel shows the averaged sea surface height (d) background, (e) analysis, and (f) increment.

5. Summary

- The stochastic PFF (SPFF) can handle high-dimensional nonlinear problems with unbiased sampling, demonstrating its ability to prevent particle collapse, maintain spread, and efficiently sample.
- The smoother version (SPFS) can further reduce analysis errors and better handle the nonlinearity in the model by using adjoint model to propagate the observation information.
- Nonlinear ocean data assimilation will be conducted with MOM6 coupled JEDI-SOCA system, and 5-day averaged MOM6 reanalysis over 30 years with 50 members will be produced in the future.

6. References

1. Pulido, M. and van Leeuwen, P.J. (2019) Sequential Monte Carlo with kernel embedded mappings: the mapping particle filter. *Journal of Computational Physics*, **396**, 400–415.
2. Hu, C.-C. and van Leeuwen, P. J. (2021) A particle flow filter for high-dimensional system applications. *Quarterly Journal of the Royal Meteorological Society*, **147**, 2352–2374.
3. Gallego, V. and Insua, D. R. (2020). Stochastic Gradient MCMC with Repulsive Forces, arXiv: 1812.00071 .

The link for this poster

Acknowledgements: This work is funded by the National Science Foundation