UNIVERSITY OF MARYLAND

cisess
Cooperative Institute for Satellite Earth System Studies

NOAA
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
U.S. DEPARTMENT OF COMMERCE

HAFS as a Testbed for Non-Gaussian Data Assimilation Developments for the UFS

Jonathan Poterjoy
University of Maryland

Sponsoring awards: NSF CAREER #AGS1848363, NSF #AGS2136969
NOAA #NA22OAR4590184, #NA20OAR4600281, #NA19NES4320002

Tuesday, 25th July, 2023

Introduction

Primary objective: Implement novel DA methodology that is immediately relevant for HAFS.

Specific topics:

- New developments for fully cycled ensemble DA within HAFS
  - Bias correction for radiance measurements (Knisely and Poterjoy, 2023; UIFCW talk on Monday)
  - Treating sampling error in high-resolution ensembles (Kurosawa and Poterjoy, in progress)
  - Non-Gaussian errors (Poterjoy 2022a,b; Kurosawa and Poterjoy 2021,2023; UIFCW poster)

Broadly relevant to all UFS applications.

Combining particle filters with Var

One objective is to explore implications of replacing the EnKF with LPF for modeling systems that run EnVar.

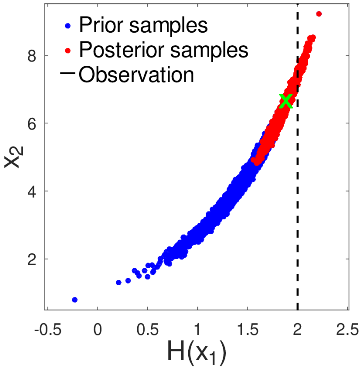
Motivation:

- Most modeling systems run EnVar for practical reasons; e.g., use of a high-resolution deterministic "control."
- EnKF is typically used to update ensemble—to provide future background error covariance for EnVar.
- EnKF members are re-centered on EnVar analysis.

Combining particle filters with Var

One objective is to explore implications of replacing the EnKF with LPF for modeling systems that run EnVar.

Motivation:



i. Posterior tends to be closer to a Gaussian than the prior.

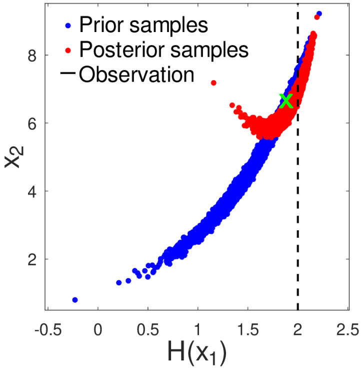
ii. Re-centering posterior ensemble on Var analysis is okay, as long as the distribution is close to Gaussian.

← **Var analysis** alongside **PF members** following assimilation.

Combining particle filters with Var

One objective is to explore implications of replacing the EnKF with LPF for modeling systems that run EnVar.

Motivation:

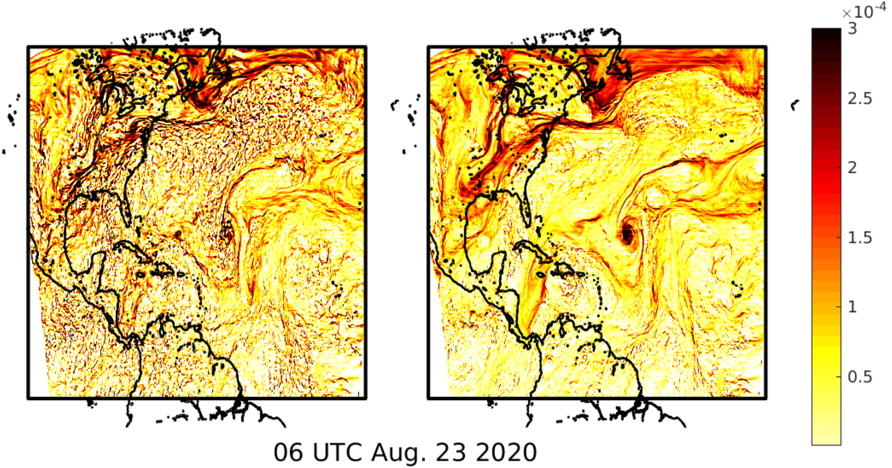


iii. Incremental 3DVar/4DVar can solve moderately nonlinear DA problems through an outer loop (e.g., **x** on left).

iv. Posterior targeted by Var is more consistent with PF than EnKF.

$\leftarrow$ Var analysis alongside **EnKF members**.

Real-world impact of assuming Gaussian prior



EnKF
(Gaussian DA)

Particle filter
(Non-parametric DA)

Combining particle filters with Var

DA comparisons:

- "EnKF-Var" $\leftarrow$ HAFS ensemble updated with EnKF and Var
- "PF-Var" $\leftarrow$ HAFS ensemble updated with LPF and Var

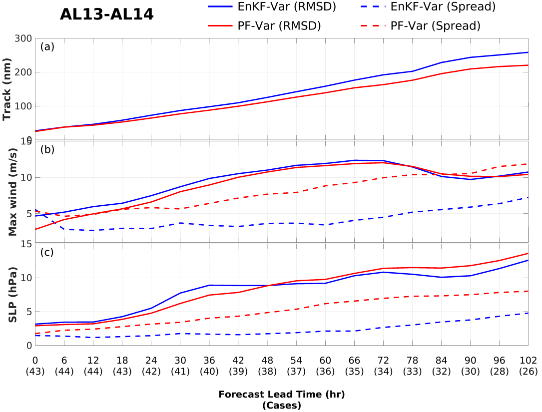
In both experiments, role of EnKF or LPF is to update 40 HAFS ensemble members about a variational analysis.

Verification:

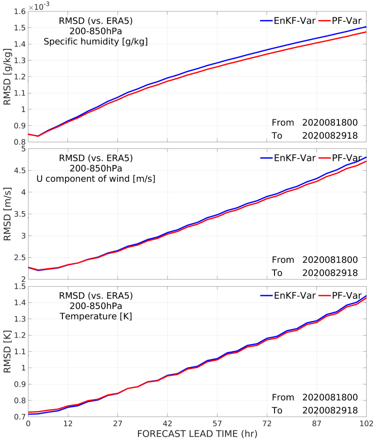
- 10-member forecasts generated every 6 h for 2 weeks
- Storm features verified using NHC Best Track data
- Synoptic scale features verified using ERA5

Verification (2 weeks of forecasts)

Track and intensity RMSEs for Laura and Marco (2020)



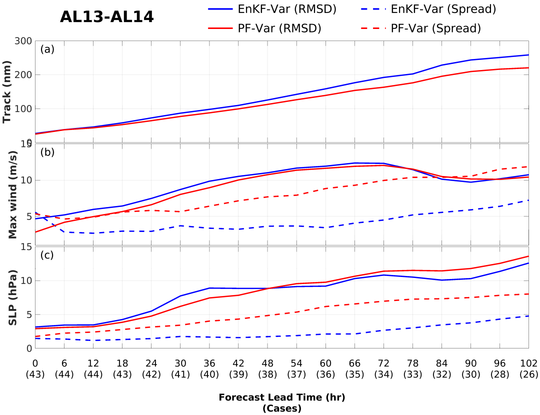
Domain-average RMSEs from ERA5



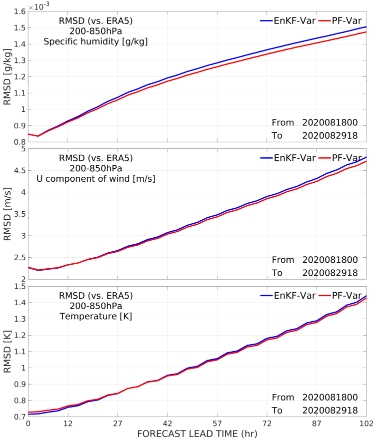
- Currently testing with 2023 HAFS-A and HAFS-B; preliminary results shows similar improvements with LPF-Var.

Verification (2 weeks of forecasts)

Track and intensity RMSEs for Laura and Marco (2020)



Domain-average RMSEs from ERA5



- LPF will soon be applied for hourly-updated GFS (FY23 WPO Innovations for Community Modeling Competition).

Future directions

Flexibility provided by non-Gaussian data assimilation:

$$p(\mathbf{x}_t|\mathbf{y}_{0:t}) \quad \propto \quad p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{0:t-1}),$$

Future directions

Flexibility provided by non-Gaussian data assimilation:

$$
\begin{aligned}
p(\mathbf{x}_t|\mathbf{y}_{0:t}) &\propto p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{0:t-1}),\\
&\approx p(\mathbf{y}_t|\mathbf{x}_t)\frac{1}{N_e}\sum_{n=1}^{N_e}\delta(\mathbf{x}-\mathbf{x}_t^n),
\end{aligned}
$$

Future directions

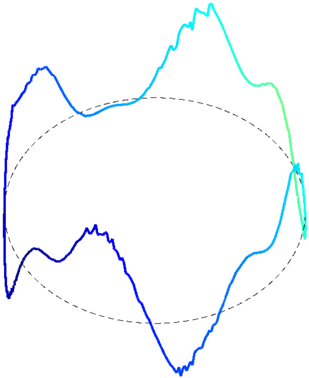
Flexibility provided by non-Gaussian data assimilation:

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{y}_{0:t}) &\propto p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{0:t-1}), \\ &\approx p(\mathbf{y}_t|\mathbf{x}_t)\frac{1}{N_e}\sum_{n=1}^{N_e}\delta(\mathbf{x}-\mathbf{x}_t^n), \\ &\propto \sum_{n=1}^{N_e} p(\mathbf{y}_t|\mathbf{x}_t^n)\delta(\mathbf{x}-\mathbf{x}_t^n). \end{aligned}$$

Large freedom exists in how we specify $p(\mathbf{y}_t|\mathbf{x}_t^n)$.

Revisiting error assumptions for measurements

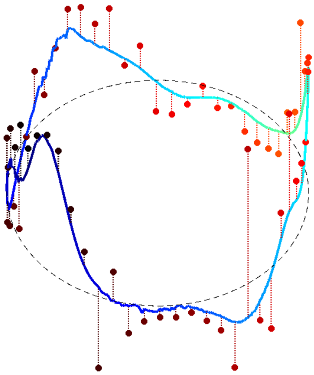
Assimilating obs with non-Gaussian, state-dependent errors.



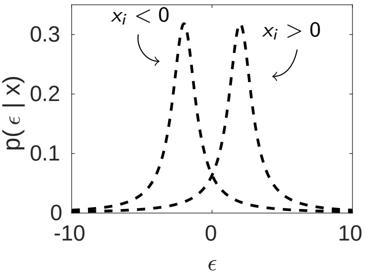
- Model III of Lorenz (2005) on periodic domain
- Model configuration supports chaotic behavior
- Characterized by $N_x = 480$ variables on periodic domain
- Data Assimilation: iterative local particle filter (*Poterjoy 2022, QJRMS; Poterjoy 2022, MWR*)

Revisiting error assumptions for measurements

Assimilating obs with non-Gaussian, state-dependent errors.



- Observations: directly measure every 8^{th} variable at $\Delta t = 0.05$
- $y_i = x_i + \epsilon$ for $i = 1, 2, \ldots, N_y$



Revisiting error assumptions for measurements

Current approach for specifying $p(\mathbf{y}_t|\mathbf{x}_t^n)$:

Assume $\mathbf{y}_t = H(\mathbf{x}_t^{truth}) + \boldsymbol{\epsilon}_t$, and apply assumptions for distribution of $\boldsymbol{\epsilon}_t$.

Revisiting error assumptions for measurements

Current approach for specifying $p(\mathbf{y}_t|\mathbf{x}_t^n)$:

Assume $\mathbf{y}_t = H(\mathbf{x}_t^{truth}) + \boldsymbol{\epsilon}_t$, and apply assumptions for distribution of $\boldsymbol{\epsilon}_t$.

For $\boldsymbol{\epsilon}_t^n = \mathbf{y}_t - H(\mathbf{x}_t^n)$,

$$\begin{aligned} p(\mathbf{y}_t|\mathbf{x}_t^n) &\approx p(\boldsymbol{\epsilon}_t^n), \\ &\approx \mathcal{N}(\boldsymbol{\epsilon}_t^n; \mathbf{0}, \mathbf{R}_t). \end{aligned}$$

Specifying likelihoods

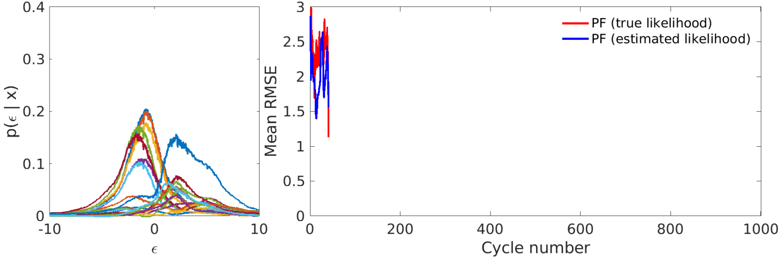
A non-parametric estimate for $p(\mathbf{y}_t|\mathbf{x}_t^n)$:

1. Adopt a low-dimensional representation of $\mathbf{y}_t$ and $\mathbf{x}_t$ from training data using nonlinear manifold learning method (*diffusion maps*; Coifman and Lafon 2006).
2. Compute data-driven estimates of $p(\epsilon_t|\mathbf{x}_t)$ or $p(\mathbf{y}_t|\mathbf{x}_t)$ using *kernel embeddings of conditional distributions* (Song et al. 2013; Berry and Harlim 2017).

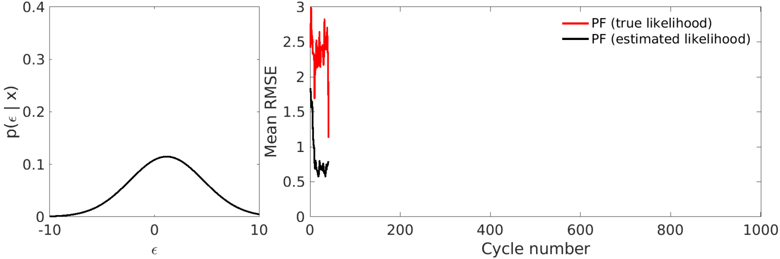
Results in a matrix representation of $p(\mathbf{y}_t|\mathbf{x}_t)$: To specify likelihood for a given member, find element of matrix that is closest to current $\mathbf{y}_t$ and $\mathbf{x}_t^n$.

Lorenz example (training time = 40 cycles)

Posterior RMSEs with non-parametric $p(\epsilon_t|\mathbf{x}_t)$



Best Gaussian estimate of $p(\epsilon_t|\mathbf{x}_t)$ (with QC)



Long-term research implications

The flexibility of data-driven likelihoods opens new research directions.

Another example application:

- We observe the "square" of model variables without knowing this function; i.e., H only <u>selects</u> state variables near obs.
- The distribution for ϵ_t is still unknown.
- All <u>5 parameters</u> (θ) are unknown (control frequency, amplitude, and coupling between large and small-scale waves).

Long-term research implications

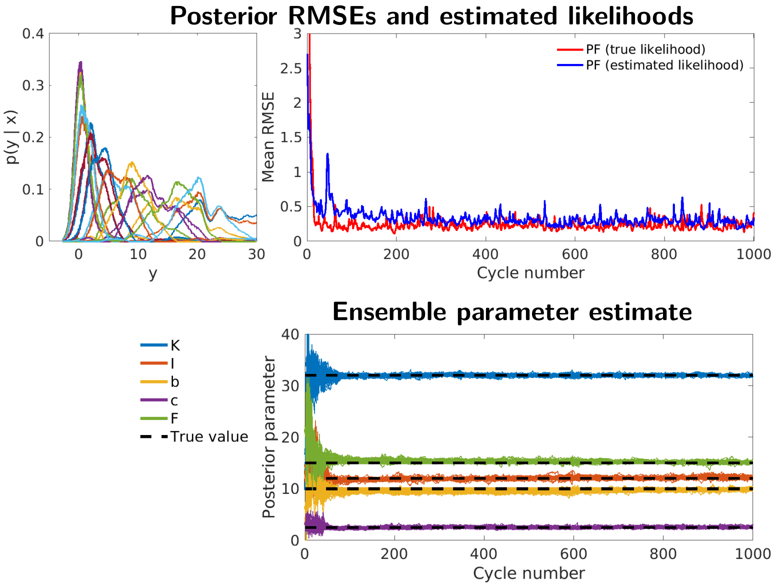
The flexibility of data-driven likelihoods opens new research directions.

Another example application:

- We observe the "square" of model variables without knowing this function; i.e., H only selects state variables near obs.
- The distribution for $\boldsymbol{\epsilon}_t$ is still unknown.
- All 5 parameters ($\boldsymbol{\theta}$) are unknown (control frequency, amplitude, and coupling between large and small-scale waves).

$$p(\mathbf{x}_t, \boldsymbol{\theta} | \mathbf{y}_{0:t}) \quad \propto \quad p(\mathbf{x}_t, \boldsymbol{\theta} | \mathbf{y}_{0:t-1}) p(\mathbf{y}_t | \mathbf{x}_t, \boldsymbol{\theta}).$$

Long-term research implications



Summary

A new non-Gaussian data assimilation strategy is shown to outperform conventional EnVar used for operational weather prediction.

Early results are encouraging, but the full benefits of non-Gaussian data assimilation still need to be explored.

As a motivating example, we show how likelihoods can be estimated non-parametrically and used for data assimilation with particle filters.

References

Berry, T. and J. Harlim, 2017: Correcting biased observation model error in data assimilation. Mon. Wea. Rev. 145, 2833 – 2853.

Coifman, R., and S. Lafon, 2006: Diffusion maps .Appl. Comput. Harmonic Anal., 21, 5 – 30.

Kurosawa, K., and Poterjoy, J., 2021: Data assimilation challenges posed by nonlinear operators: A comparative study of ensemble and variational filters and smoothers, Mon. Wea. Rev. 149, 2369 – 2389.

Kurosawa, K. and J. Poterjoy, 2023: A statistical hypothesis testing strategy for adaptively blending particle filters and ensemble Kalman filters for data assimilation. Mon. Wea. Rev., 151, 105 – 125.

McCurry, J., J. Poterjoy, K. Knopfmeier, and L. Wicker, 2023: An Evaluation of Non-Gaussian Data Assimilation Methods in Moist Convective Regimes. Mon. Wea. Rev., 151, Provisionally accepted.

Poterjoy, J. 2022: Implications of multivariate non-Gaussian data assimilation for multi-scale weather prediction. Mon. Wea. Rev. 150, 1475 – 1493.

Poterjoy, J., 2022: Regularization and tempering for a moment-matching localized particle filter. Q. J. Roy. Meteor. Soc., Published online 31 May 2022.

Song, K. Fukumizu, and A. Gretton, 2013: Kernel embeddings of conditional distributions: A unified kernel framework for non-parametric inference in graphical models. IEEE Signal Process. Mag., 30, 98 – 111.

Kernel embeddings of conditional distributions

We can represents likelihoods using kernel embeddings:

$$\boxed{p(\mathbf{d}_i|\hat{\mathbf{y}}_j) = \sum_{k=1}^{M} \mu_{kj}\phi_k(\mathbf{d}_i)q(\mathbf{d}_i)}$$

See Song et al. (2009,2013)

$$\begin{aligned}
\mu_{kj} &= \sum_{l=1}^{M} \psi_l(\hat{\mathbf{y}})[\mathbf{C}\tilde{\mathbf{C}}^{-1}]_{kl}, \\
\mathbf{C}_{lk} &= \frac{1}{N}\sum_{j=1}^{N} \phi_l(\mathbf{d}_j)\psi_k(\hat{\mathbf{y}}_j), \\
\tilde{\mathbf{C}}_{lk} &= \frac{1}{N}\sum_{j=1}^{N} \psi_l(\hat{\mathbf{y}}_j)\psi_k(\hat{\mathbf{y}}_j).
\end{aligned}$$

where μ_{kj} coefficients determine dependence across $\mathbf{d}$ and $\hat{\mathbf{y}}$.

Constructing marginals and basis

For $q(\mathbf{d})$, adopt a kernel estimate:

- Variable bandwidth kernel densities provide non-parametric representation of marginal pdfs.

$$q(\mathbf{d}) = \sum_{k=1}^{N} N(\mathbf{d}_k, \mathbf{B}_k), \text{ where } \mathbf{B}_k \text{ is a covariance.}$$

Constructing marginals and basis

For $q(\mathbf{d})$, adopt a kernel estimate:

- Variable bandwidth kernel densities provide non-parametric representation of marginal pdfs.

$$q(\mathbf{d}) = \sum_{k=1}^{N} N(\mathbf{d}_k, \mathbf{B}_k), \text{ where } \mathbf{B}_k \text{ is a covariance.}$$

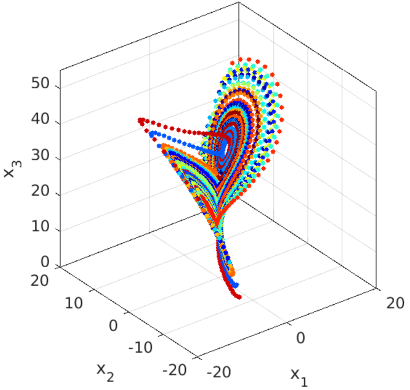
For basis functions, diffusion maps (Coifman and Lafon 2006) is a reasonable choice:

- Manifold learning method for represent data in lower-dimensional space
- Similar strategy applied by Berry and Harlim (2017)

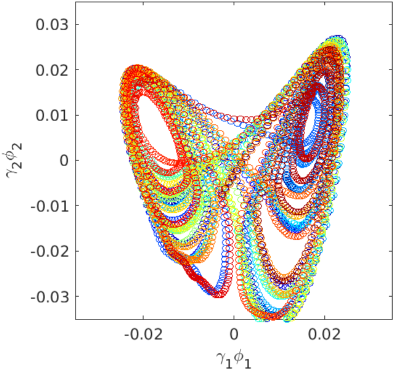
Constructing basis functions

Example: Data produced from Lorenz (1963) model

Model data



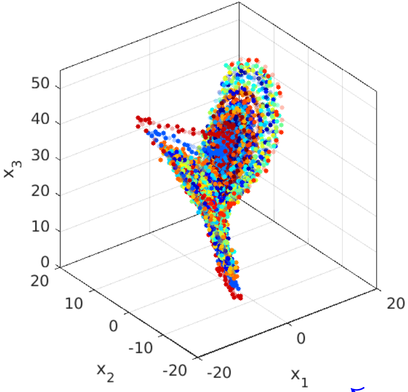
Two-dimensional embeddings



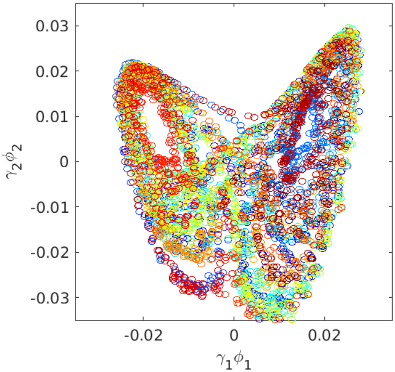
Constructing basis functions

Example: Data produced from Lorenz (1963) model

Observations



Two-dimensional embeddings



Unbiased Gaussian errors

References

Kurosawa, K., and Poterjoy, J., 2021: Data assimilation challenges posed by nonlinear operators: A comparative study of ensemble and variational filters and smoothers, Mon. Wea. Rev. 149, 2369 – 2389. https://doi.org/10.1175/MWR-D-20-0368.1

Kurosawa, K. and J. Poterjoy, 2023: A statistical hypothesis testing strategy for adaptively blending particle filters and ensemble Kalman filters for data assimilation. Mon. Wea. Rev., 151, 105 – 125. https://doi.org/10.1175/MWR-D-22-0108.1

Poterjoy, J., G. J. Alaka, Jr., and H. R. Winterbottom, 2021: The irreplaceable utility of sequential data assimilation for numerical weather prediction system development: Lessons learned from an experimental HWRF system. Wea. Forecasting, 36, 661 – 677. https://doi.org/10.1175/WAF-D-20-0204.1

Poterjoy, J. 2022a: Implications of multivariate non-Gaussian data assimilation for multi-scale weather prediction. Mon. Wea. Rev., 150, 1475 – 1493. https://doi.org/10.1175/MWR-D-21-0228.1

Poterjoy, J., 2022b: Regularization and tempering for a moment-matching localized particle filter. Q. J. Roy. Meteor. Soc., Published online 31 May 2022. https://doi.org/10.1002/qj.4328

Hodyss, D. and P. A. Reinecke, 2013: Skewness of the prior through position errors and its impact on data assimilation. Data Assimilation for Atmospheric, Oceanic, and Hydrologic Applications. S. K. Park and L. Xu, Eds., Vol. II Springer, 843 pp.