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
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.
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.

← Var analysis alongside EnKF members.
Real-world impact of assuming Gaussian prior

06 UTC Aug. 23 2020

EnKF
(Gaussian DA)

Particle filter
(Non-parametric DA)
Combining particle filters with Var

DA comparisons:

- “EnKF-Var” ← HAFS ensemble updated with EnKF and Var
- “PF-Var” ← 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:

\[ p(x_t | y_{0:t}) \propto p(y_t | x_t)p(x_t | y_{0:t-1}), \]
Future directions

Flexibility provided by non-Gaussian data assimilation:

\[
p(x_t | y_{0:t}) \propto p(y_t | x_t)p(x_t | y_{0:t-1}),
\]

\[
\approx p(y_t | x_t) \frac{1}{N_e} \sum_{n=1}^{N_e} \delta(x - x^n_t),
\]
Future directions

Flexibility provided by non-Gaussian data assimilation:

\[ p(x_t | y_{0:t}) \propto p(y_t | x_t) p(x_t | y_{0:t-1}), \]
\[ \approx p(y_t | x_t) \frac{1}{N_e} \sum_{n=1}^{N_e} \delta(x - x^n_t), \]
\[ \propto \sum_{n=1}^{N_e} p(y_t | x^n_t) \delta(x - x^n_t). \]

Large freedom exists in how we specify \( p(y_t | x^n_t) \).
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 \((\text{Poterjoy 2022, QJRMS}; \text{Poterjoy 2022, MWR})\)
Revisiting error assumptions for measurements

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

- Observations: directly measure every 8th 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 $p(y_t|x_t^n)$:

Assume $y_t = H(x_t^{\text{truth}}) + \epsilon_t$, and apply assumptions for distribution of $\epsilon_t$. 
Revisiting error assumptions for measurements

Current approach for specifying $p(y_t | x^n_t)$:

Assume $y_t = H(x^n_t^{\text{truth}}) + \epsilon_t$, and apply assumptions for distribution of $\epsilon_t$.

For $\epsilon^n_t = y_t - H(x^n_t)$,

$$p(y_t | x^n_t) \approx p(\epsilon^n_t),$$

$$\approx \mathcal{N}(\epsilon^n_t; 0, R_t).$$
Specifying likelihoods

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

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 $\mathbf{x}^n_t$. 
Lorenz example (training time = 40 cycles)

**Posterior RMSEs with non-parametric** $p(\epsilon_t|x_t)$

**Best Gaussian estimate of** $p(\epsilon_t|x_t)$ (with QC)
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 $\epsilon_t$ is still unknown.
- All 5 parameters ($\theta$) 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 $\epsilon_t$ is still unknown.
- All 5 parameters ($\theta$) are unknown (control frequency, amplitude, and coupling between large and small-scale waves).

$$p(x_t, \theta | y_{0:t}) \propto p(x_t, \theta | y_{0:t-1}) p(y_t | x_t, \theta).$$
Long-term research implications

**Posterior RMSEs and estimated likelihoods**

![Graph showing posterior RMSEs and estimated likelihoods with cycle number on the x-axis and mean RMSE on the y-axis.](image)

**Ensemble parameter estimate**

![Graph showing ensemble parameter estimate with cycle number on the x-axis and posterior parameter on the y-axis.](image)
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


We can represent likelihoods using kernel embeddings:

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

See Song et al. (2009, 2013)

\[
\mu_{kj} = \sum_{l=1}^{M} \psi_l(\hat{y}) [C\tilde{C}^{-1}]_{kl},
\]

\[
C_{lk} = \frac{1}{N} \sum_{j=1}^{N} \phi_l(\mathbf{d}_j) \psi_k(\hat{y}_j),
\]

\[
\tilde{C}_{lk} = \frac{1}{N} \sum_{j=1}^{N} \psi_l(\hat{y}_j) \psi_k(\hat{y}_j).
\]

where \(\mu_{kj}\) coefficients determine dependence across \(\mathbf{d}\) and \(\hat{y}\).
Constructing marginals and basis

For $q(d)$, adopt a kernel estimate:

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

$$q(d) = \sum_{k=1}^{N} N(d_k, B_k),$$

where $B_k$ is a covariance.
For $q(\mathbf{d})$, adopt a kernel estimate:

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

$$q(\mathbf{d}) = \sum_{k=1}^{N} N(\mathbf{d}_k, \mathbf{B}_k),$$
where $\mathbf{B}_k$ 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


