An Integral-Form Ensemble Square-Root Filter with Efficient and Precise Model-Space Localization

Robin Armstrong[1], Ian Grooms[2], and Chris Snyder[3]

International Symposium on Data Assimilation (Online), "Data Assimilation Methodology"

March 13th, 2025

1. PhD Candidate, Cornell University, Center for Applied Mathematics
2. Associate Professor, CU Boulder, Department of Applied Mathematics
3. Senior Scientist, NSF-NCAR, Mesoscale and Microscale Meteorology Laboratory

Joint Work With:

Ian Grooms (CU Boulder)

Chris Snyder (NSF-NCAR MMM)

Background and Notation

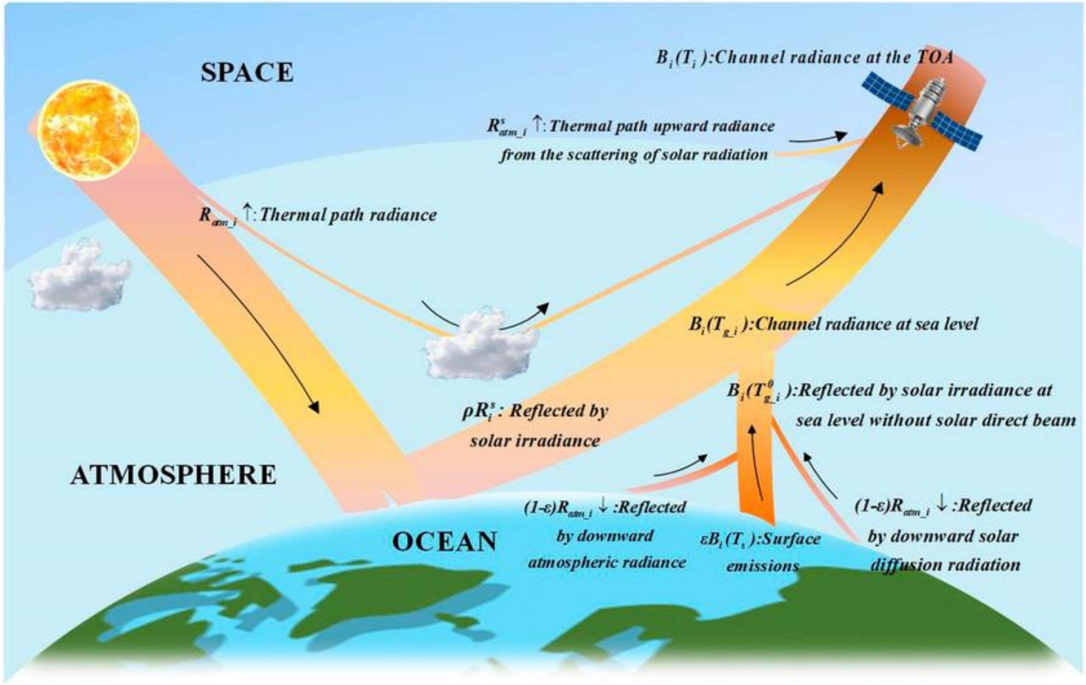
- <u>Forecast</u>: $X_f \sim \mathcal{N}(\mu_x, \mathbf{\Sigma}_{xx})$.
- <u>Data</u>: $Y = h(X_f) + \xi \sim \mathcal{N}(\mu_h, \mathbf{\Sigma}_{hh} + \mathbf{R})$.
- <u>Analysis</u>: $X_a = (X_f \mid Y = y)$.
- <u>Kalman's equations</u>: $X_a \sim \mathcal{N}(\mu_a, \mathbf{\Sigma}_a)$, where

$$\begin{aligned} \mu_a &= \mu_x + \mathbf{K}(y - \mu_h), \\ \mathbf{\Sigma}_a &= \mathbf{\Sigma}_{xx} - \mathbf{K}\mathbf{\Sigma}_{xh}^{\mathrm{T}}, \end{aligned}$$

  and $\mathbf{K} = \mathbf{\Sigma}_{xh}(\mathbf{R} + \mathbf{\Sigma}_{hh})^{-1}$ = "Kalman gain matrix."

- <u>Ensemble Kalman filters (EnKFs)</u>: transform samples of X_f into samples of X_a:

$$f(X_f, y) \sim X_a$$



Source: https://www.researchgate.net/figure/Radiance-received-by-the-satellite-sensor_fig4_375018928

Ensemble Kalman Filters

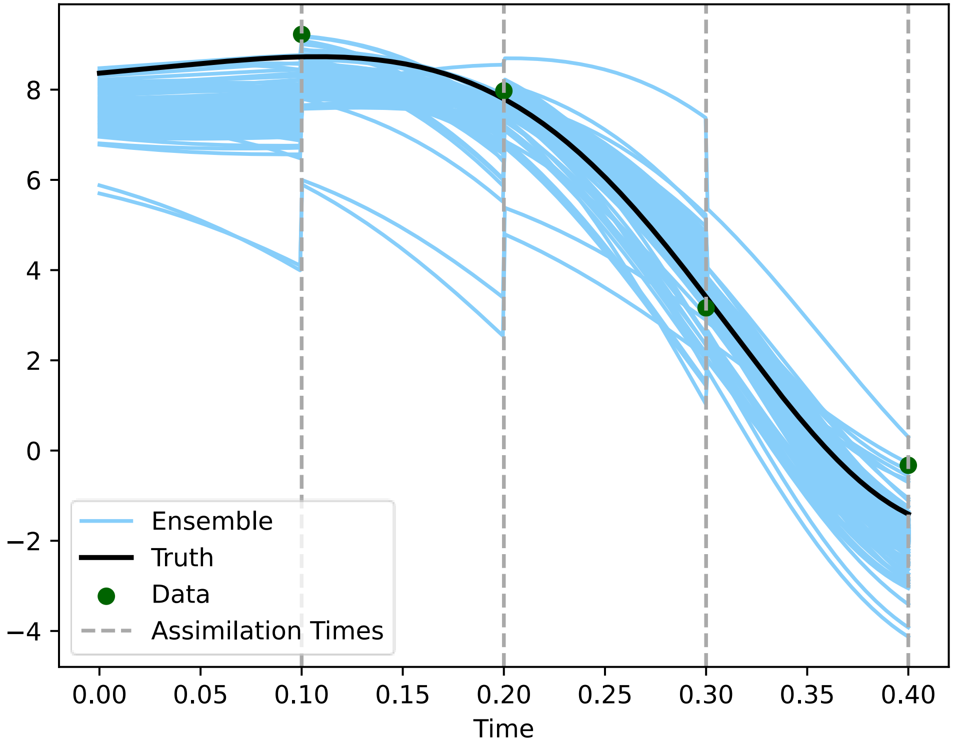
- Original EnKF (Evensen, 1994): updates each sample of X_f as if it were the mean.

$$\hat{X}_a = X_f + \mathbf{K}\Big(y - h\big(X_f\big)\Big).$$

Obs. error is undercounted, $\mathrm{Cov}[\hat{X}_a] \prec \mathbf{\Sigma}_a$.

- Stochastic EnKF (Houtekamer & Mitchell, 1998) adds random noise to the observation.

$$X_a = X_f + \mathbf{K}\big(y - h\big(X_f\big) - \hat{\xi}\big), \qquad \hat{\xi} \sim \mathcal{N}(0, \mathbf{R}).$$



- Square-root filter (Whitaker & Hamill, 2002) update the means and perturbations separately.

$$\begin{aligned} \mu_a &= \mu_x + \mathbf{K}(y - \mu_h), & \mathbf{K} &= \mathbf{\Sigma}_{xh}(\mathbf{R} + \mathbf{\Sigma}_{hh})^{-1} \\ X_a - \mu_a &= X_f - \mu_x - \mathbf{G}\big(h\big(X_f\big) - \mu_h\big), & \mathbf{G} &= \mathbf{\Sigma}_{xh}\big(\mathbf{R} + \mathbf{\Sigma}_{hh} + \mathbf{R}(\mathbf{I} + \mathbf{R}^{-1}\mathbf{\Sigma}_{hh})^{1/2}\big)^{-1} \end{aligned}$$

Covariance Localization

- The <u>empirical ensemble covariance</u>,

$$\mathbf{\Sigma}_{xx} = \mathbf{Z}_x \mathbf{Z}_x^{\mathrm{T}},$$

  is low-rank (good for efficiency) but noisy (bad for accuracy).

- The <u>localized ensemble covariance</u>,

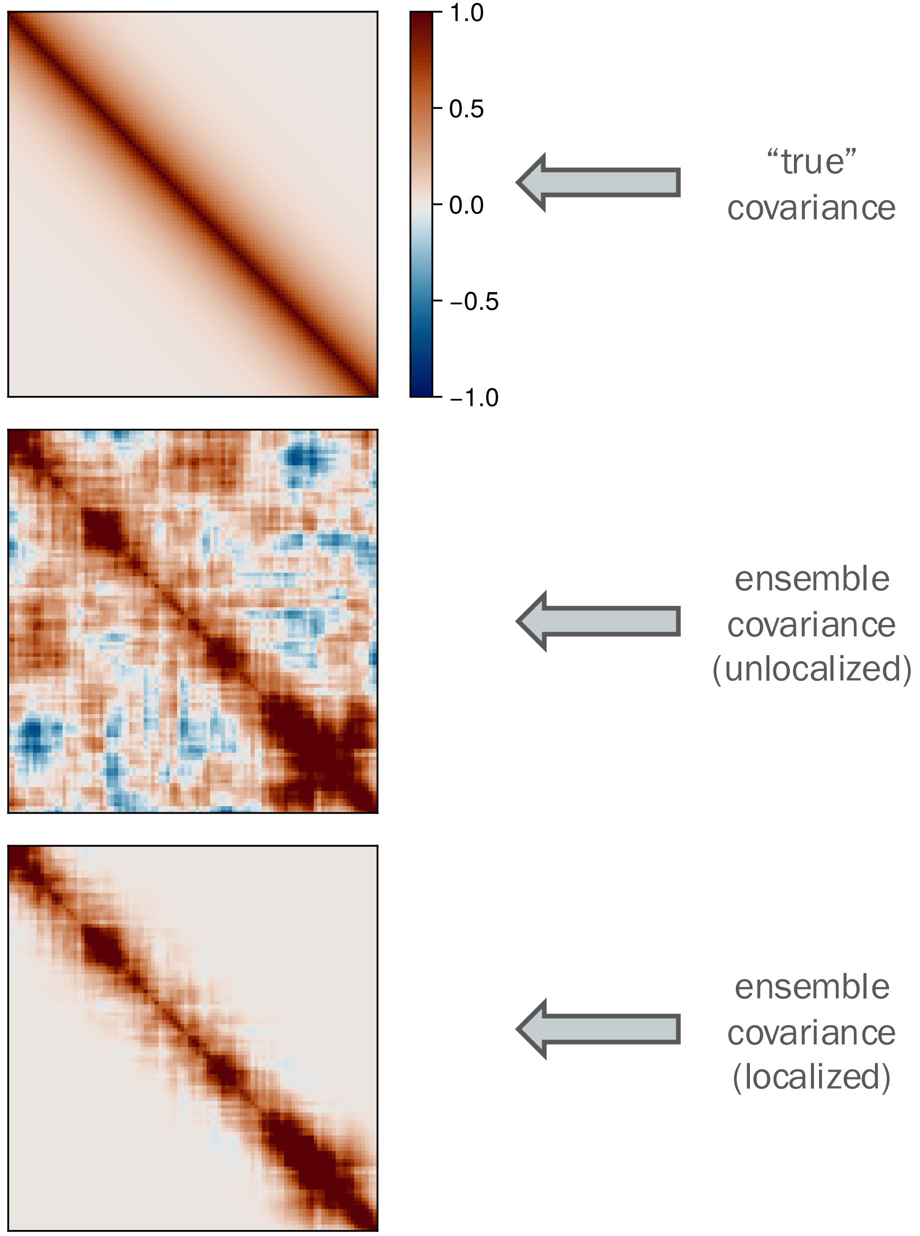
$$\mathbf{\Sigma}_{xx} = \mathbf{L} \circ (\mathbf{Z}_x \mathbf{Z}_x^{T})$$

  is less noisy but <u>full-rank</u> (difficult to compute with).

- <u>Operator access</u> lets us run Krylov methods and linear solves:

$$\mathbf{\Sigma}_{xx} u = \sum_i z^{(i)} \circ \mathbf{L}(z^{(i)} \circ u)\,,$$

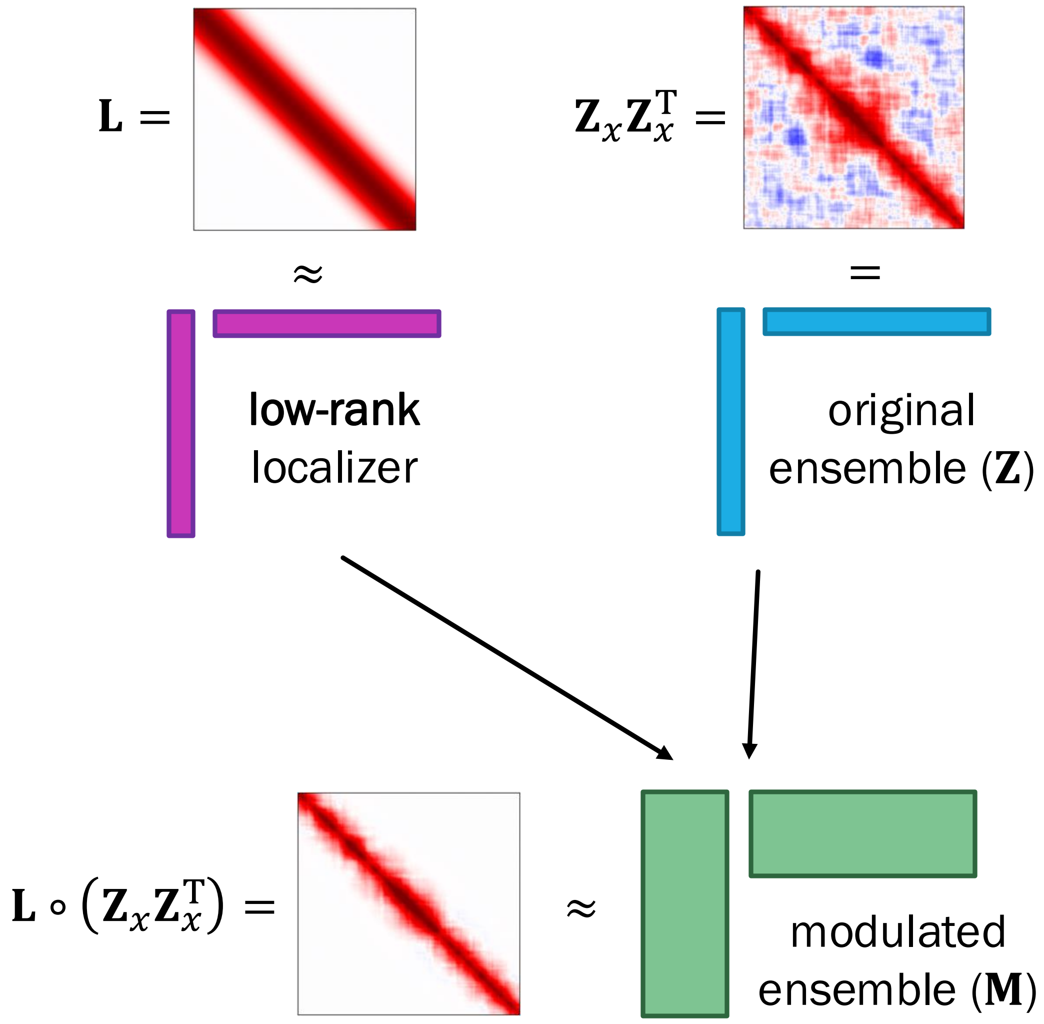
- **How to handle the square-root?**



Ensemble Modulation

- Constructs a larger modulated ensemble whose empirical covariance retains some localization.
- Essentially a low-rank factorization of $\mathbf{L} \circ (\mathbf{Z}_x\mathbf{Z}_x^{\mathrm{T}})$.
- Constructed using a spectral decomposition of $\mathbf{L}$ (Bishop et al., 2017) or using a randomized SVD (Farchi & Bocquet, 2019).
- Only works well if $\mathbf{L} \circ (\mathbf{Z}_x\mathbf{Z}_x^{\mathrm{T}})$ has fast singular value decay!
- Measure of complexity:

$$k = \frac{\text{modulated ensemble size}}{\text{original ensemble size}}$$



A Modulation-Free Perturbation Update

- Stochastic EnKF:

$$\begin{aligned} \mu_a &= \mu_x + \mathbf{K}(y - \mu_h) \\ X_a - \mu_a &= X_f - \mu_x - \mathbf{K}\big(h(X_f) + \xi - \mu_h\big), \qquad \xi \sim \mathcal{N}(0, \mathbf{R}). \end{aligned}$$

- **Key idea 1**: <u>deterministically</u> inflate $\mathbf{R}$.

<u>Theorem (A. and Grooms, 2025)</u>. Let $\mathbf{K}(s) = \boldsymbol{\Sigma}_{xh}((s+1)\mathbf{R} + \boldsymbol{\Sigma}_{hh})^{-1}$, and let $p(s) = [\pi\sqrt{s}(s+1)]^{-1}$ for $s \in (0, \infty)$. If

$$\begin{aligned} \mu_a &= \mu_x + \mathbf{K}(y - \mu_h), \\ X_a - \mu_a &= X_f - \mu_x - \int_0^\infty \mathbf{K}(s)\big(h(X_f) - \mu_h\big)\, p(s) ds\,, \end{aligned}$$

then $\mathbb{E}[X_a]$ and $\mathrm{Cov}[X_a]$ are as given by Kalman's equations.

- **Key idea 2**: use <u>quadrature</u> to discretize the integral.

Reweighting The Integral

- We want to evaluate

$$\int_0^\infty \mathbf{K}(s)\big(h(X_f) - \mu_h\big)\, p(s) ds\,,$$

where $p(s) = [\pi\sqrt{s}(s+1)]^{-1}$. Very slowly decaying!

<u>Theorem (A. and Grooms, 2025)</u>. Define $\mathcal{D} \subseteq \mathbb{R}_+$ and $\hat{r}, \hat{s} : \mathcal{D} \to \mathbb{R}_+$ such that

$$\frac{1}{\sqrt{c+1}} = \int_{t\in\mathcal{D}} \frac{\hat{r}(t)}{c + \hat{s}(t) + 1} dt$$

for all c in some open set containing $\{0\} \cup \lambda(\mathbf{R}^{-1/2}\boldsymbol{\Sigma}_{hh}\mathbf{R}^{-1/2})$. Then,

$$\int_0^\infty \mathbf{K}(s)\big(h(X_f) - \mu_h\big)\, p(s) ds = \int_{t\in\mathcal{D}} \mathbf{K}(\hat{s}(t))\,\big(h(X_f) - \mu_h\big)\, \hat{p}(t) dt,$$

where $\hat{p}(t) = \hat{r}(t)(\hat{s}(t)+1)^{-1}$ (note that $\int \hat{p}(t)\, dt = 1$).

InFo-ESRF (Integral-Form Ensemble Square-Root Filter)

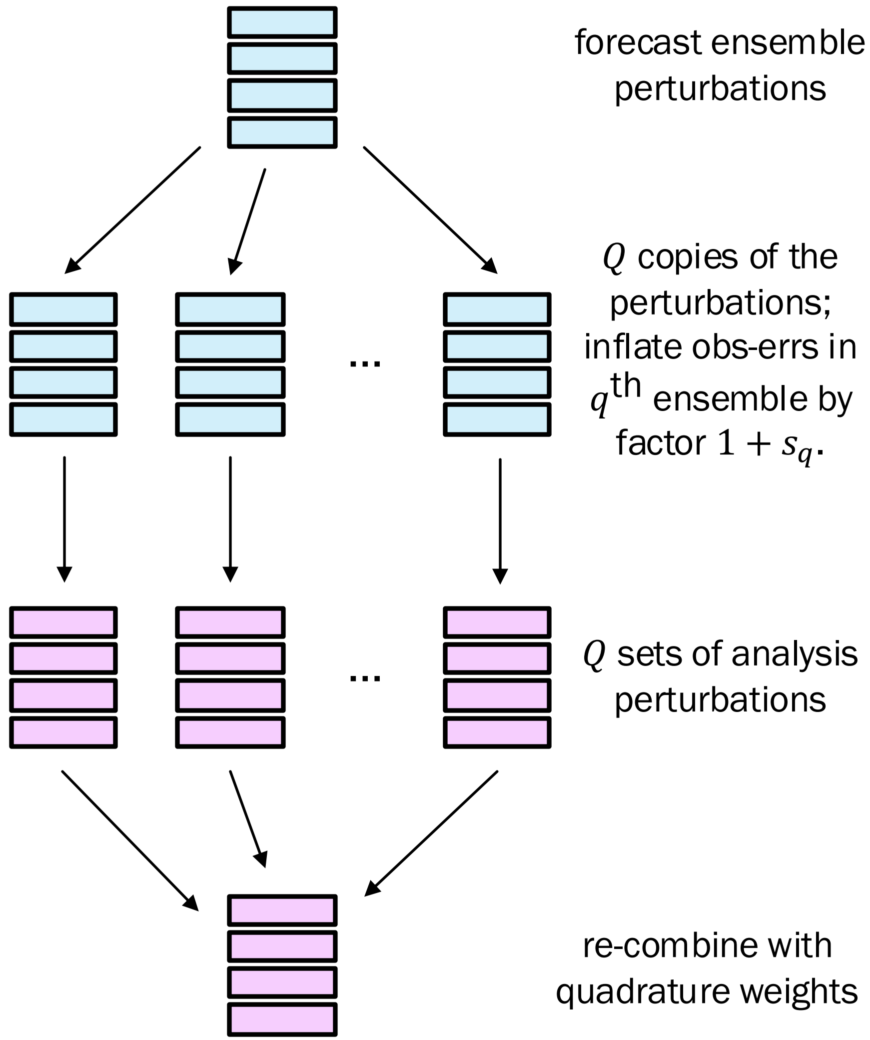
- Quadrature approximation:

$$\int_{t\in\mathcal{D}} \mathbf{K}(\hat{s}(t))\left(h(X_f)-\mu_h\right)\hat{p}(t)dt \approx \sum_{q=1}^{Q} p_q \mathbf{K}(s_q)\left(h(X_f)-\mu_h\right)$$

- 1. for $i = 1, \dots, m$ # both loops parallelize
- 2. $z_a^{(i)} \leftarrow x_f^{(i)} - \mu_x$
- 3. for $q = 1, \dots, Q$
- 4. $z_a^{(i)} \leftarrow z_a^{(i)} - p_q \mathbf{K}(s_q)\left(h\left(x_f^{(i)}\right) - \mu_h\right)$ # PGC
- 5. end
- 6. end

- Measure of complexity:

$$k = \text{number of quadrature points}$$

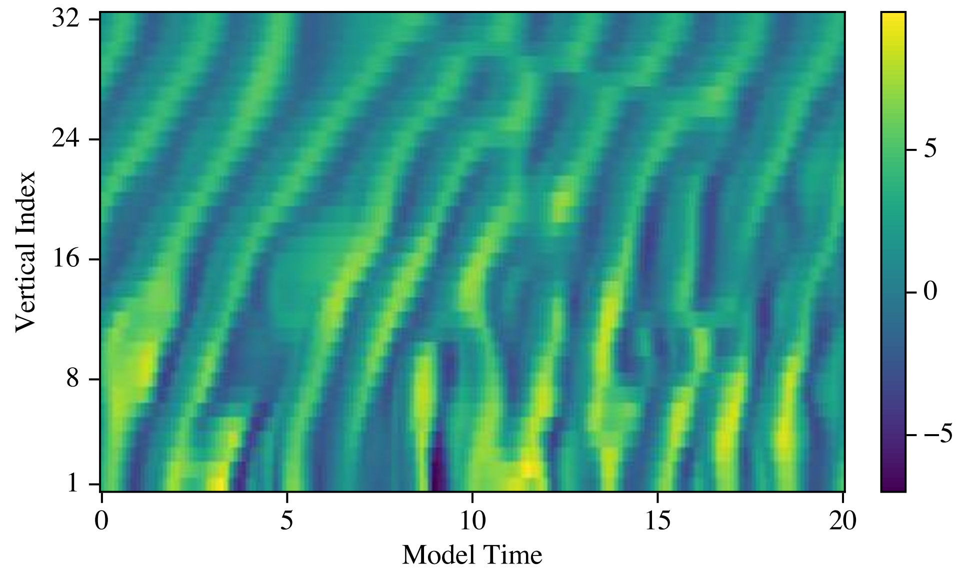


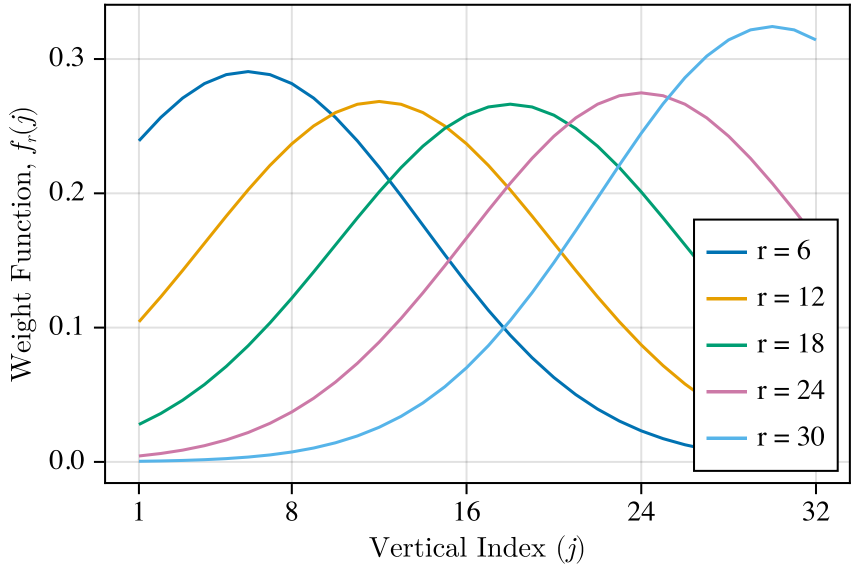
Test Case 1: Multi-Layer Lorenz System

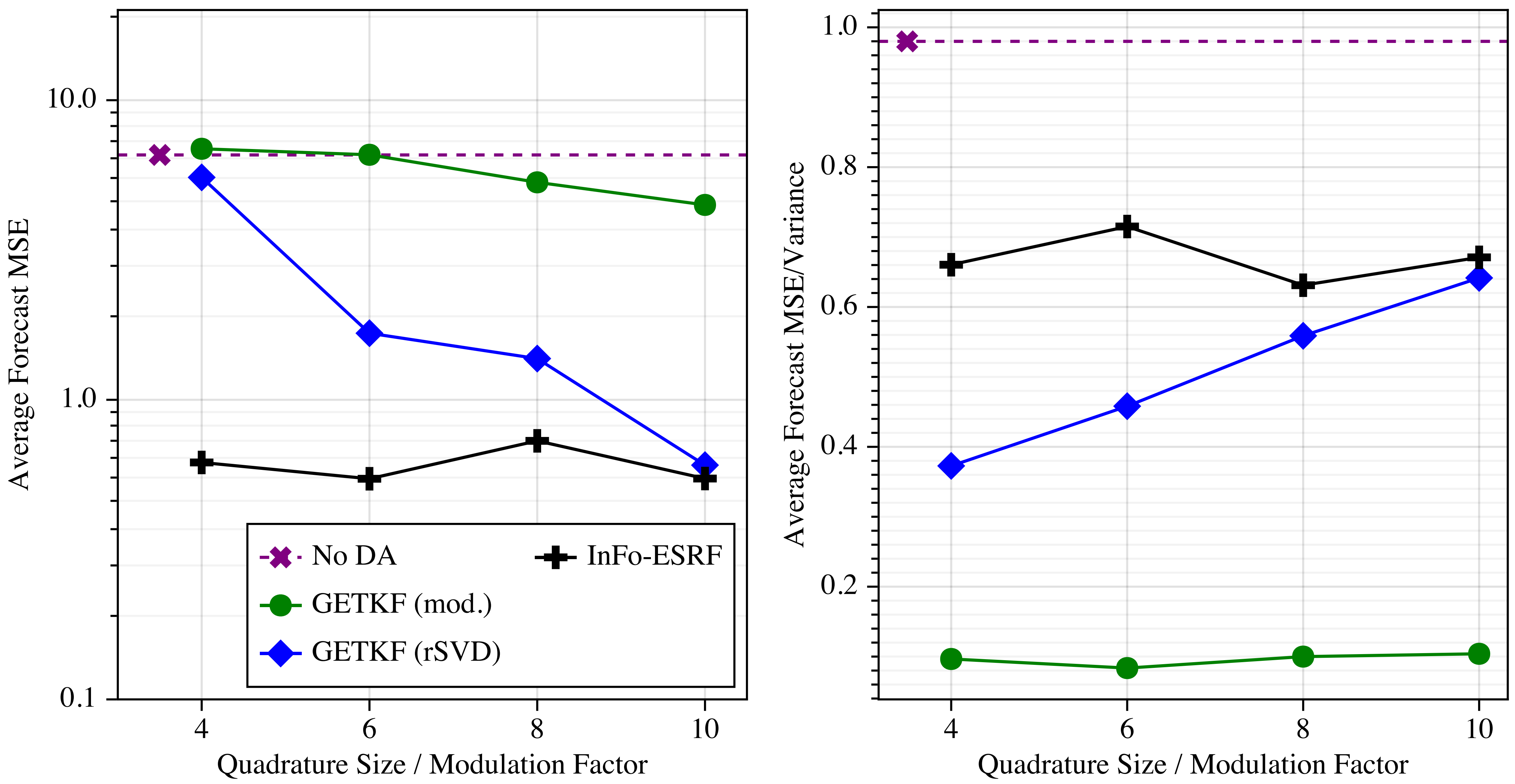
- **Model and Observing System**:
  - 32 coupled layers of 40-variable L'96 systems.
  - Forcing decreases linearly from 8 (bottom) to 4 (top).
  - "Satellite-like" measurements of 5 weighted vertical sums for every 5th column.
  - Gaussian i.i.d. noise with variance ≈ 1% of the climatological variance.
  - From Farchi and Bocquet, FAMS (2019).

- **Localization**: Gaspari-Cohn model-space localization in both the horizontal and vertical.

- **Experiment**: 5000 forecast-assimilation cycles ($\Delta t = 0.05$), average MSE and spread measured over last 4000.



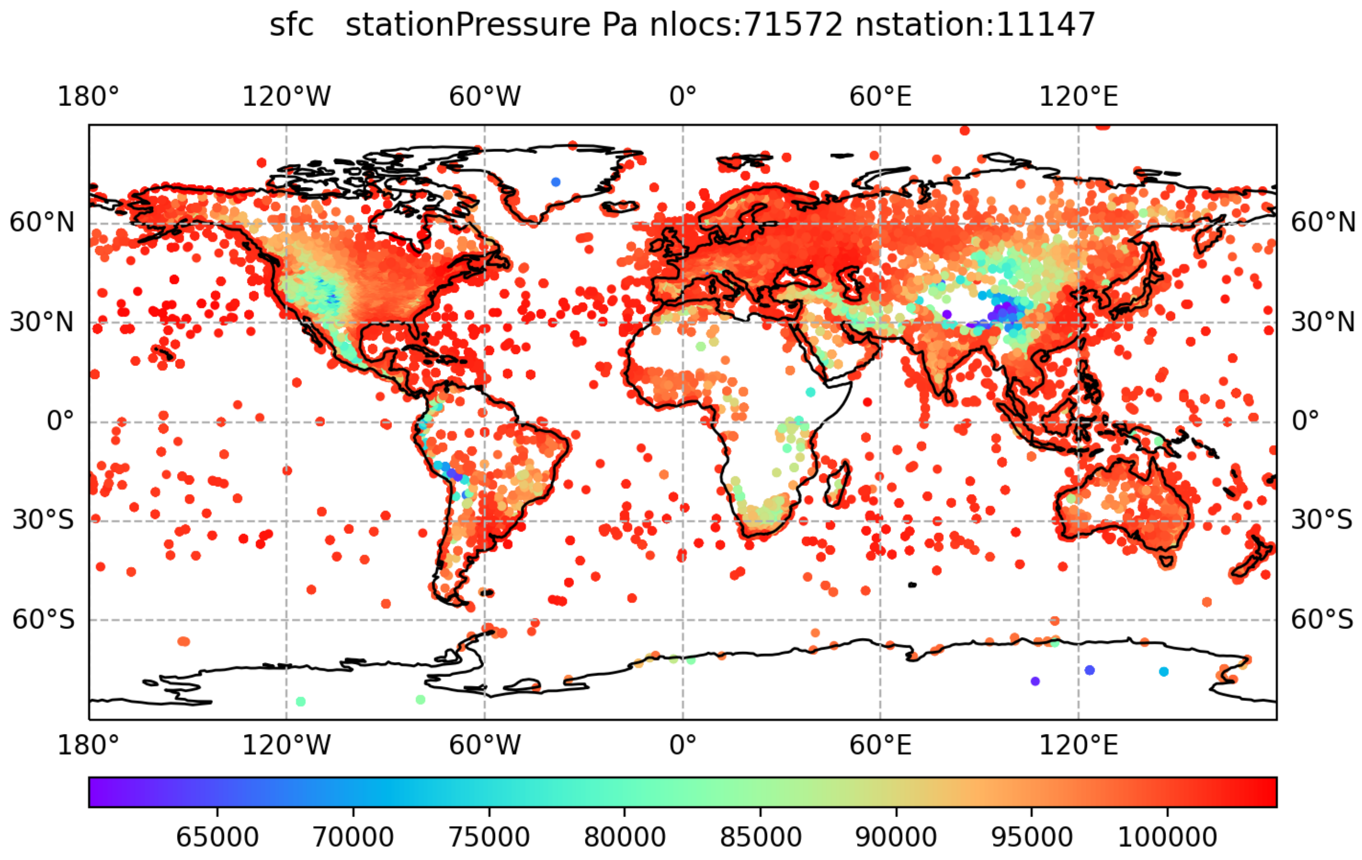




Average Forecast MSE
10.0
1.0
0.1
4
6
8
10
Quadrature Size / Modulation Factor
No DA
InFo-ESRF
GETKF (mod.)
GETKF (rSVD)
Average Forecast MSE/Variance
1.0
0.8
0.6
0.4
0.2
4
6
8
10
Quadrature Size / Modulation Factor

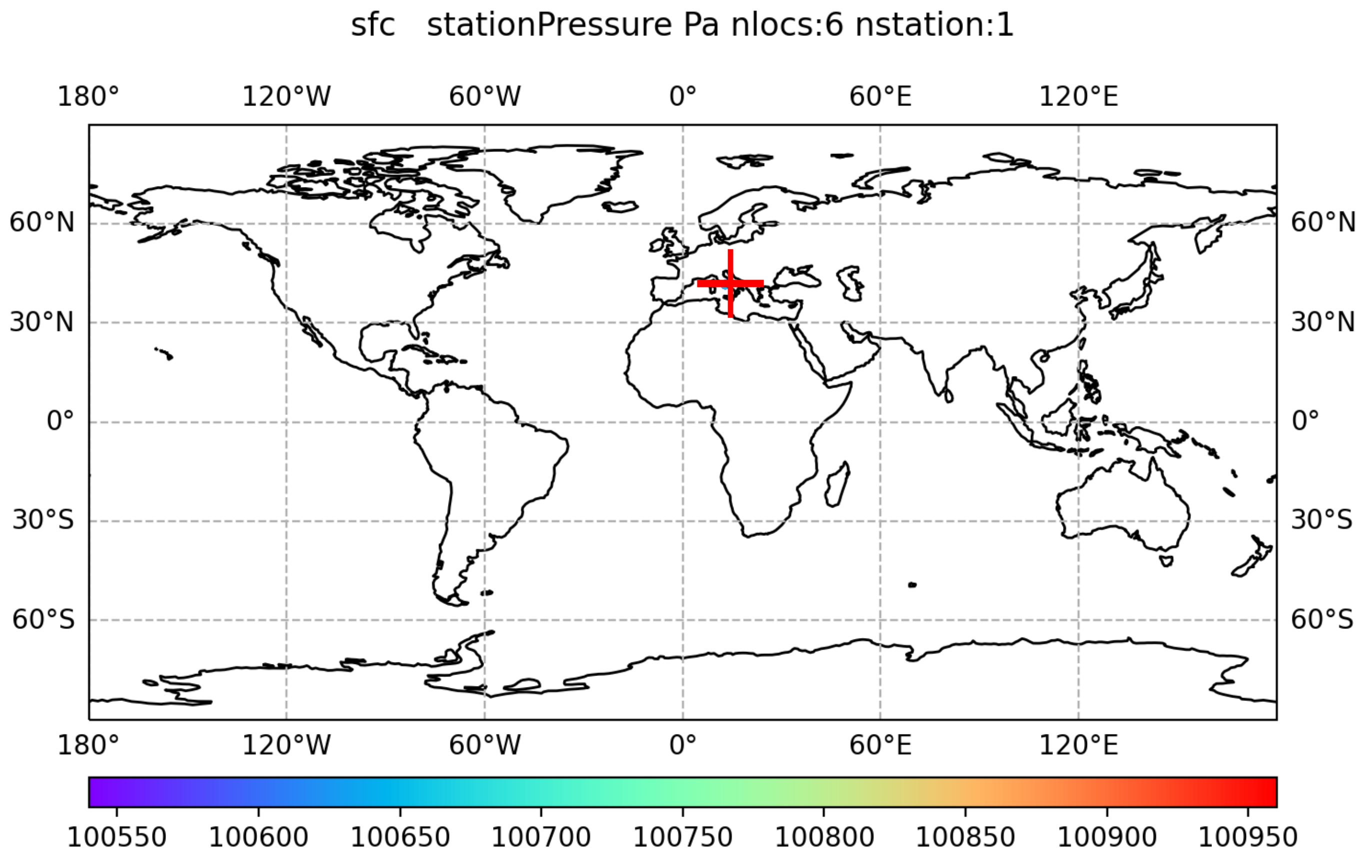
Test Case 2: MPAS-JEDI

- **Background Ensemble:** 20-members from GEFS analysis.
- **Background Covariance:** SABER ensemble covariance model, BUMP-NICAS localization.
- **Observations:** a <u>single</u> surface pressure measurement over Italy.
- **Implementation:** Kalman gain matrices applied using MPAS-JEDI 3D-EnVar executables (<u>one</u> outer loop).

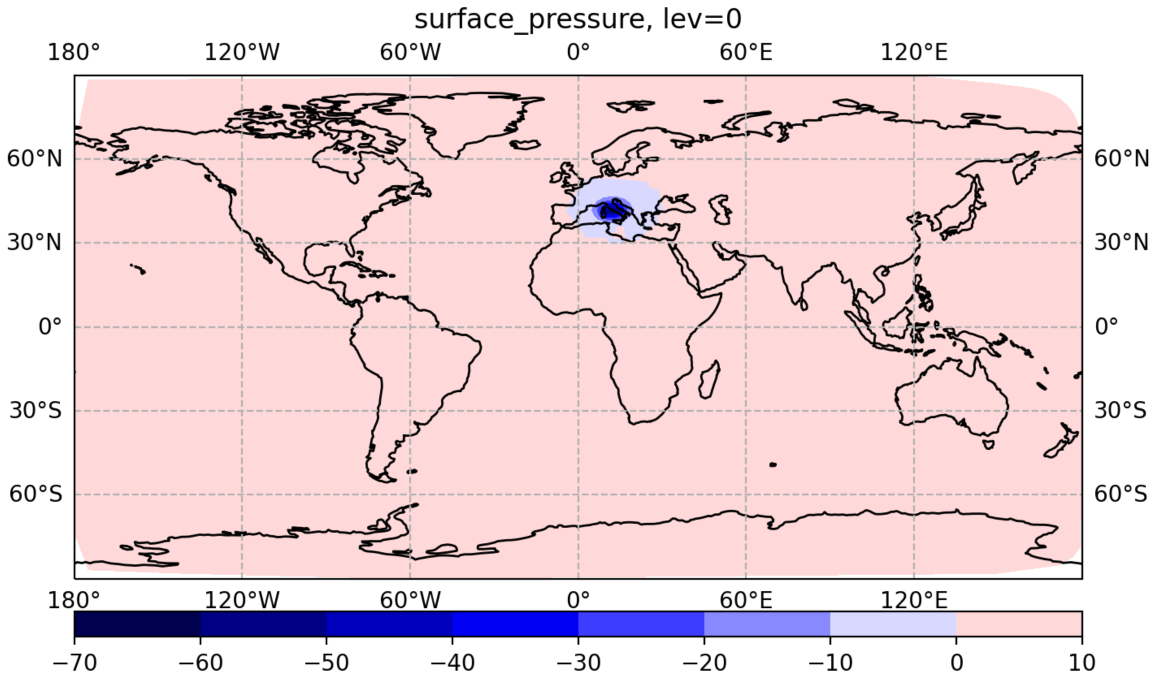


Test Case 2: MPAS-JEDI

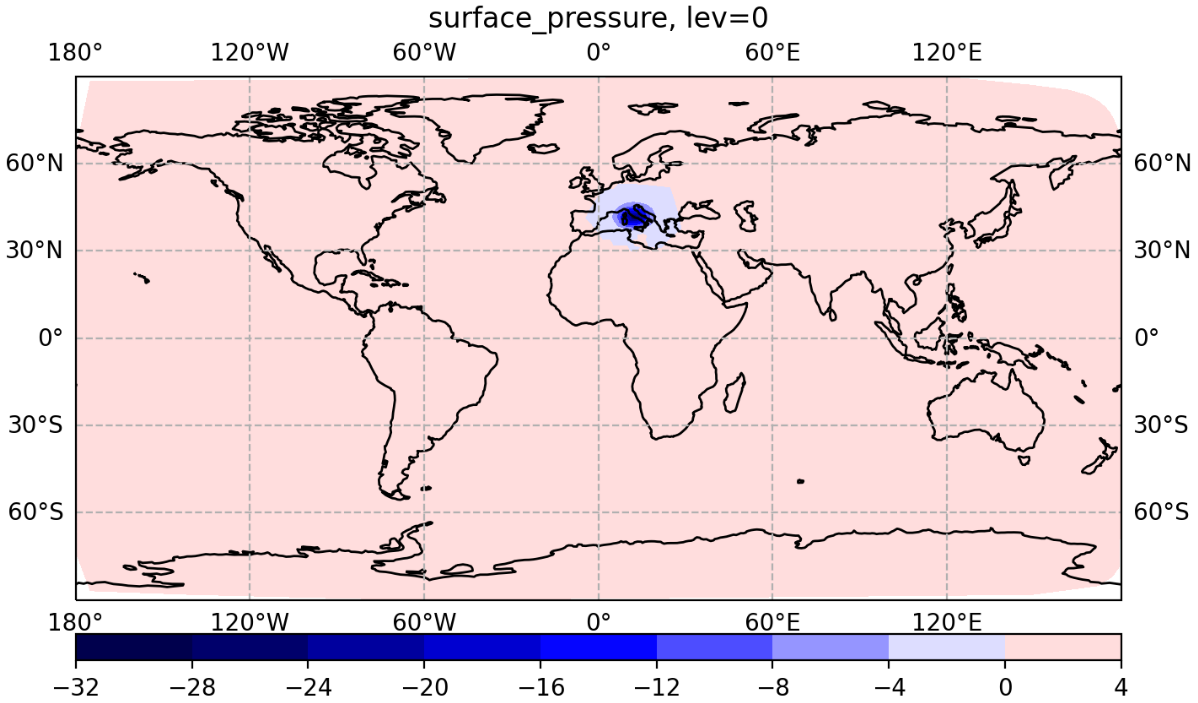
- **Background Ensemble**: 20-members from GEFS analysis.
- **Background Covariance**: SABER ensemble covariance model, BUMP-NICAS localization.
- **Observations**: a <u>single</u> surface pressure measurement over Italy.
- **Implementation**: Kalman gain matrices applied using MPAS-JEDI 3D-EnVar executables (<u>one</u> outer loop).



MPAS-JEDI Results: Changes in Ensemble Variance



EDA/3D-EnVar (perturbed obs)



InFo-ESRF (no obs perturbations)

Conclusion

- **We have demonstrated:** *an ensemble square-root filter which updates perturbations by discretizing an integral form of the Kalman filter update equations. This lets us avoid evaluating a matrix square-root, eliminating the need to approximate the forecast covariance with modulation.*

Preprint available on arXiv!