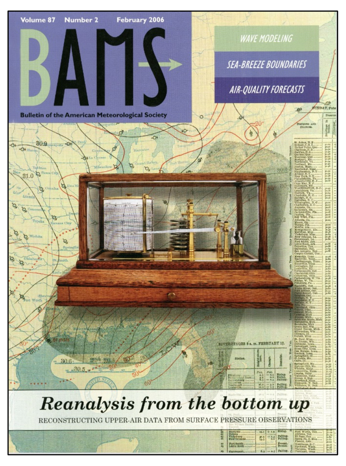
Assimilating real Earth system observations with machine learning models

Laura C. Slivinski (NOAA/OAR/PSL) and Jeffrey S. Whitaker (UCAR/CPAESS & NOAA/NWS/OMD)

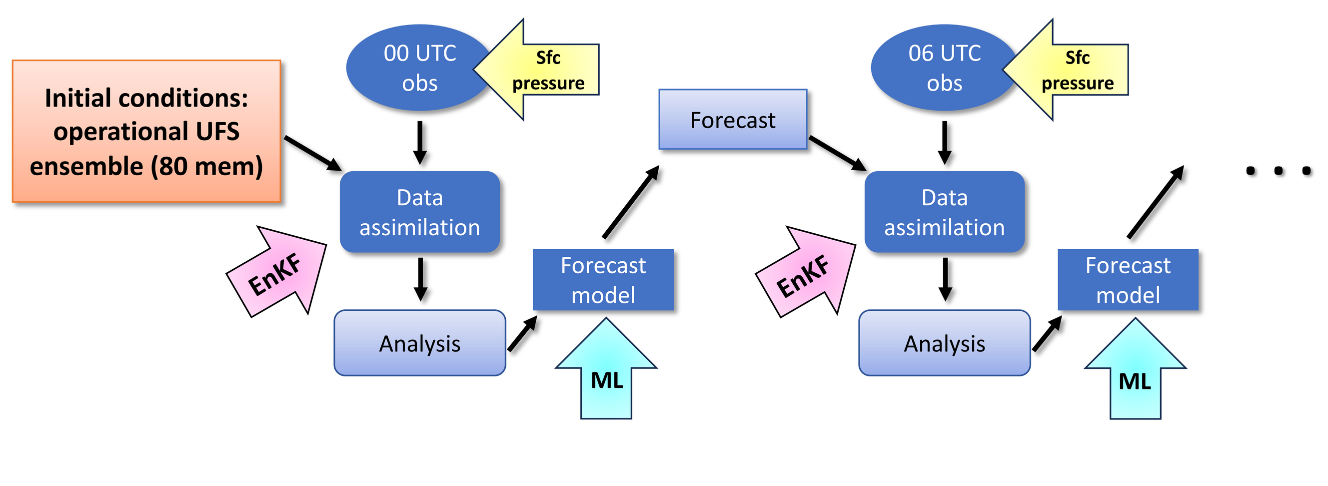
Introduction

- In 2025, we wanted to confront promising new ML models with observations in a cycling data assimilation system.
- A stringent test involves **assimilating only surface pressure in a cycling ensemble Kalman filter** (EnKF), since performance relies on accurate estimates of background-error covariances.
- This setup has been proven in the 20th Century Reanalysis (20CR): EnKF works well with sparse observations because of dynamic covariances, and surface pressure can get "most" of the answer.
- Sfc pres forecast skill is often not optimized in training.



EnKF cycling: experiment setup

6-hourly cycling: traditional cycling EnKF, but short forecasts are generated with ML model

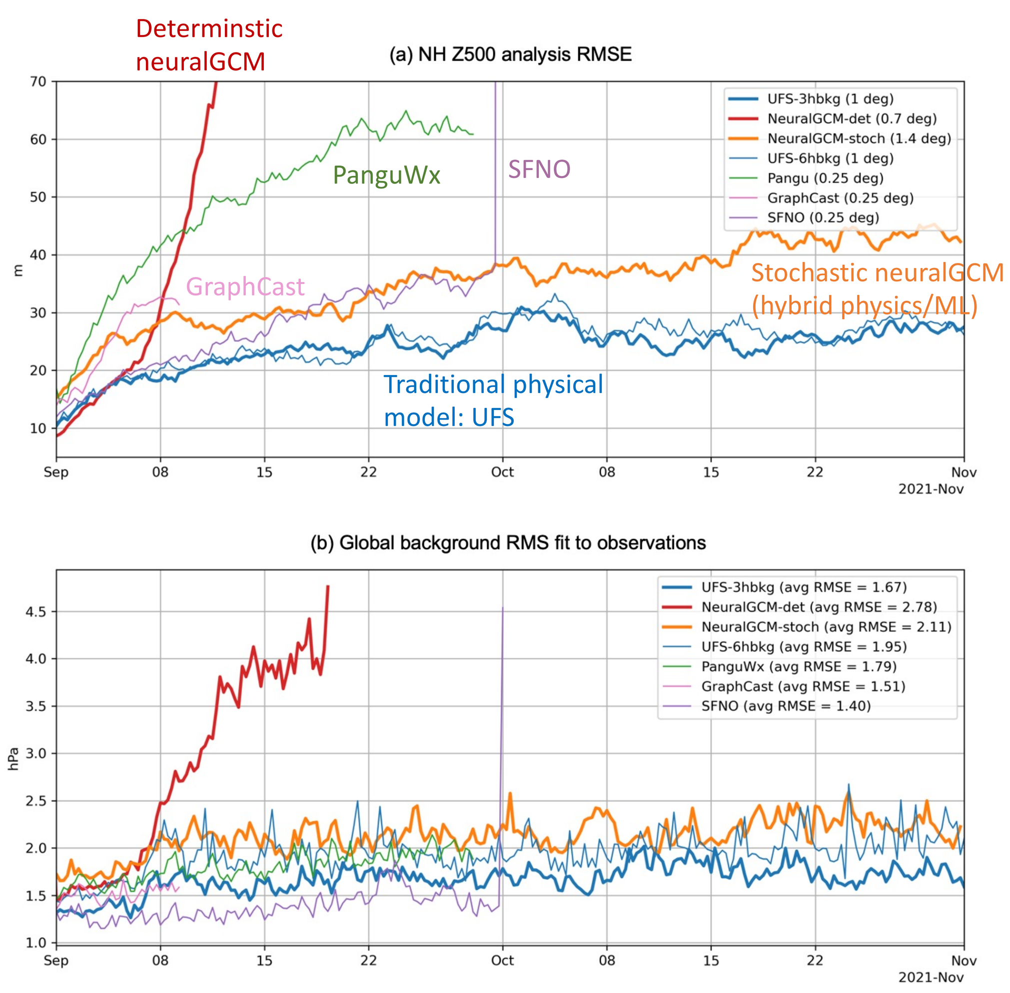


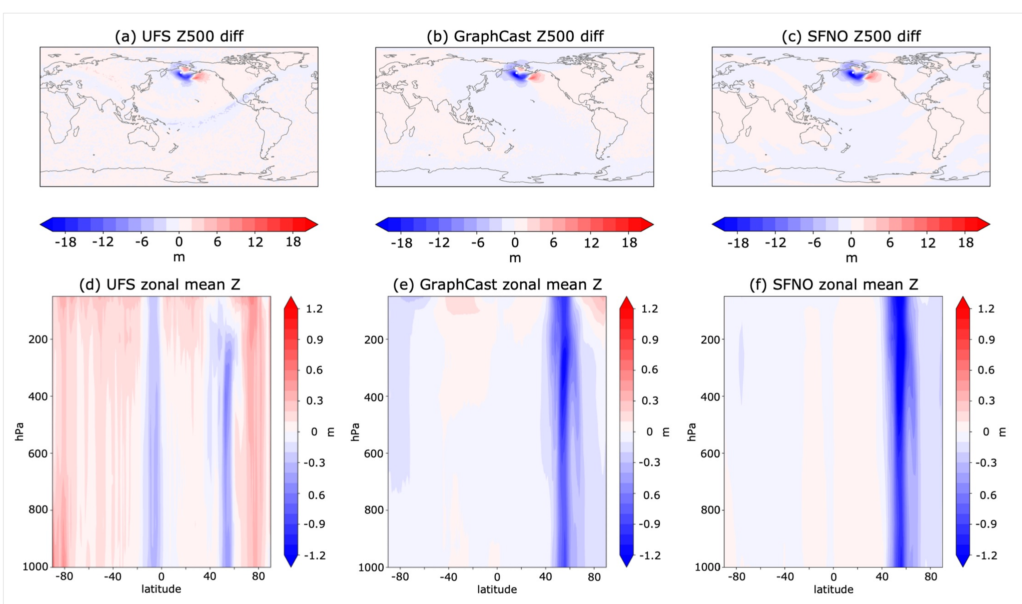
- Serial ensemble square root filter (EnSRF) as in 20CRv3
  - Horiz. & vert. localization, relaxation to prior spread inflation (& relaxation to prior perturbations for non-stochastic emulators)
  - No obs bias-correction or non-linear quality control

EnKF cycling: Results (2025)

NH Z500 errors (rel. to ERA5) grow, and all pure emulators eventually develop NaNs. Hybrid neuralGCM is the only stable ML model.

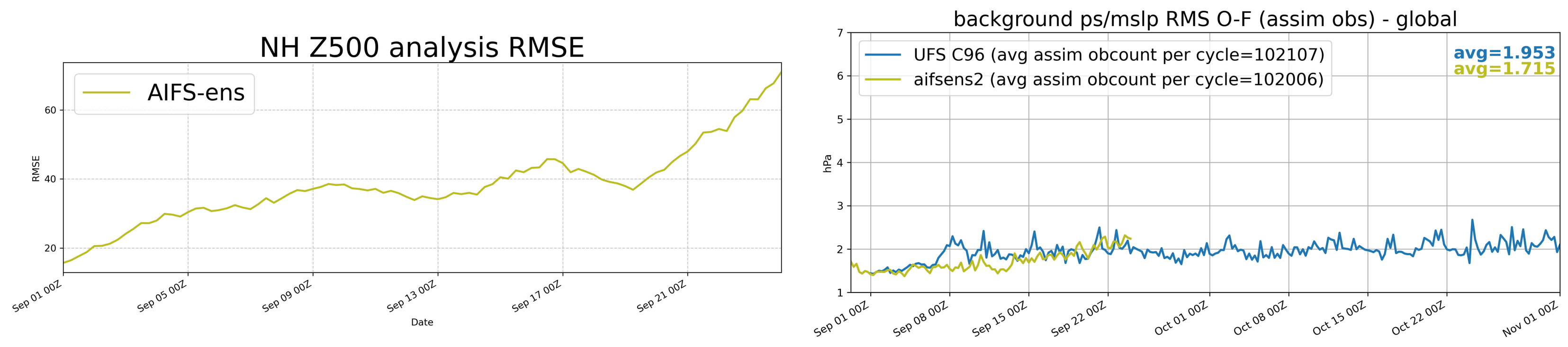
MSLP fields remain realistic at observation locations, while unobserved fields become unphysical.





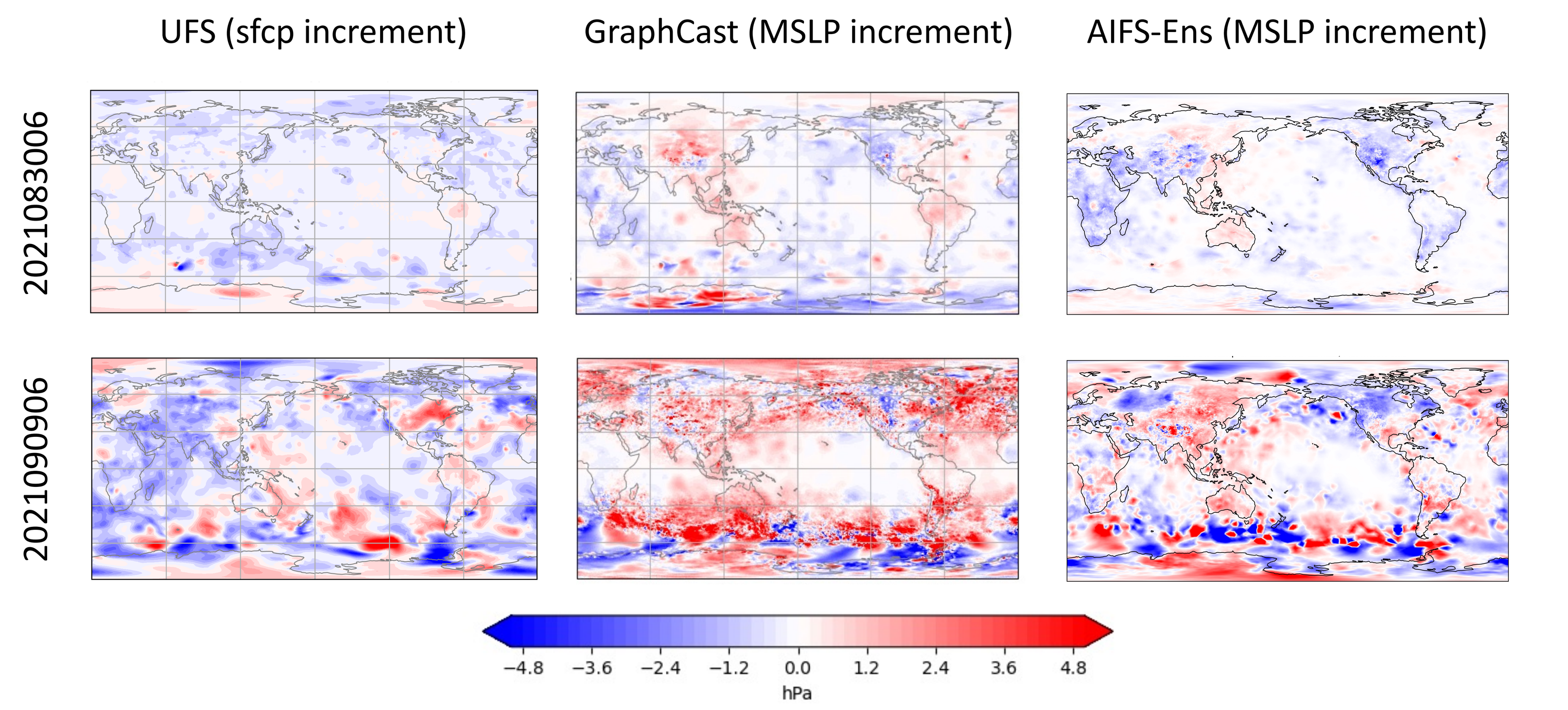
Single-ob assimilation experiments suggest that the pure emulators do not accurately represent growth of small-scale perturbations.

EnKF cycling: Results with AIFS-Ens



AIFS-Ens does not degenerate to NaNs, but Z500 performance is severely degraded (left) while observation departures remain reasonable (right).

AIFS-Ens increments do not develop small-scale noisy increments (which cause GraphCast to fail), but the increments are stronger and noisier than the physical UFS.

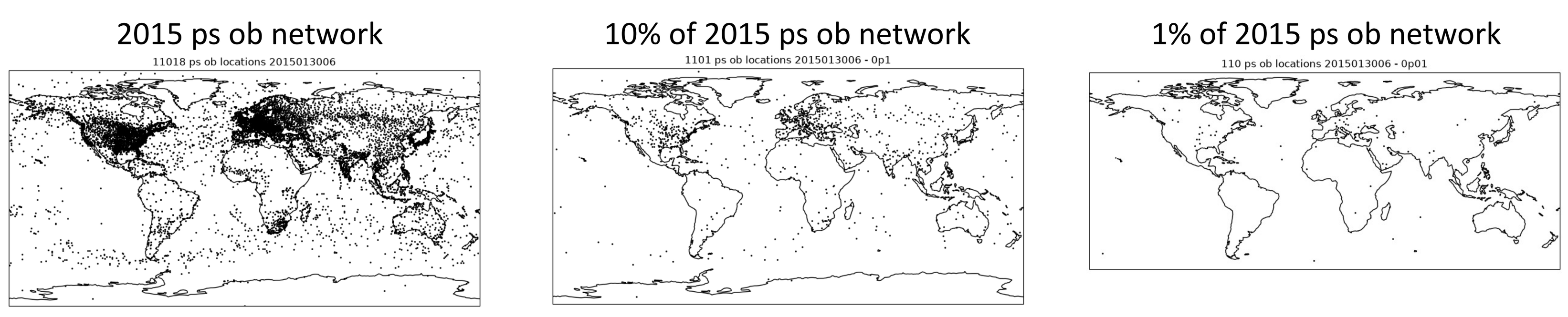


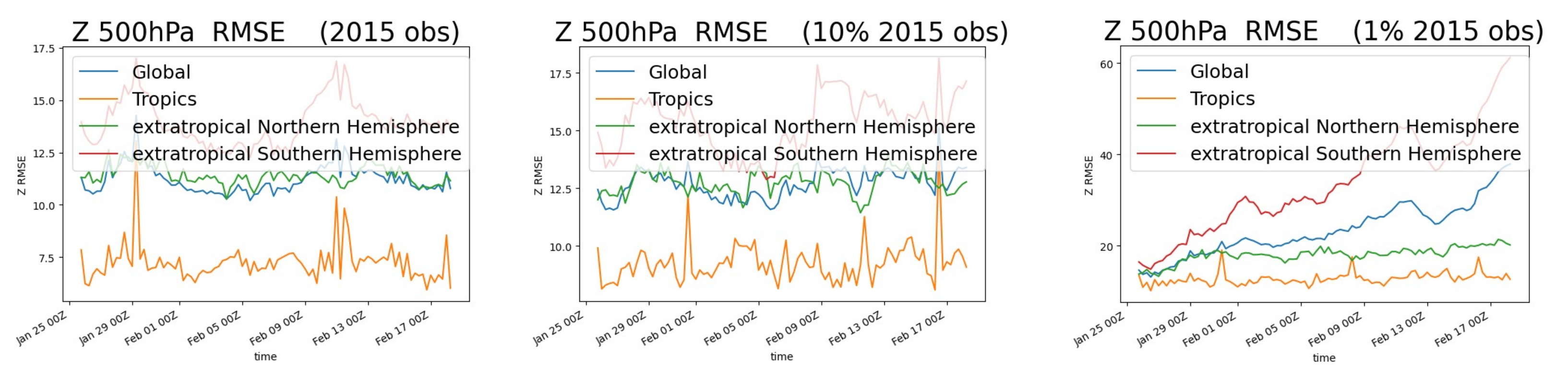
Another way? Long window 4D-Var

- Hakim et. al. test a strong-constraint, long-window 4D-Var formulation using the differentiable NeuralGCM (2.8deg) and only assimilating surface pressure observations.

$$J(\mathbf{x}_0,\ldots,\mathbf{x}_K) = \sum_{k=0}^{K}(\mathbf{y}_k - \mathcal{H}(\mathbf{x}_k))^{\mathrm{T}}\mathbf{R}_k^{-1}(\mathbf{y}_k - \mathcal{H}(\mathbf{x}_k)) + \alpha_t N_{ob}\iiint_V \left(\frac{\partial \mathbf{x}_0}{\partial t}\right)^2 \cos\phi \, d\lambda \, d\phi \, d\sigma.$$

- 4-day window; background term neglected in cost function.
- Analysis errors for NH Z500 are less than half those of 20CRv3 in 2015 (20CRv3 RMSD is ~24m).
- Could we use this to quickly generate a better 20CR?





- ➢ The "default" 4-day configuration works well with 10% observation density (~10% loss in performance).
- ➢ With 1% observation density, errors grow (in SH).
- ➢ Try longer window? Incorporate a prior? Implement EDA?
- ➢ *Note: initial conditions from ERA5; realistic tests should initialize from climatology.*

Compo, G. P., J. S. Whitaker, and P. D. Sardeshmukh, 2006: Feasibility of a 100-Year Reanalysis Using Only Surface Pressure Data. *Bull. Amer. Meteor. Soc.*, **87**, 175–190, https://doi.org/10.1175/BAMS-87-2-175.

Slivinski, L. C., Compo, G. P., Whitaker, J. S., et. al. Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system. *Q J R Meteorol Soc*. 2019; 145: 2876–2908. https://doi.org/10.1002/qj.3598

Slivinski, L. C., Whitaker, J. S., Frolov, S., Smith, T. A., & Agarwal, N. (2025). Assimilating observed surface pressure into ML weather prediction models. *Geophysical Research Letters*, 52, e2024GL114396. https://doi.org/10.1029/2024GL114396

Hakim, G. J., Whitaker, J. S., Huang, B., & Frolov, S., 2026: Long-window 4DVar for reanalysis using a differentiable weather model. https://arxiv.org/pdf/2608.11515

Laura.Slivinski@noaa.gov