Data Assimilation in the era of the AI revolution: where we stand today

Stephen G. Penny - Head of Weather, Sofar Ocean

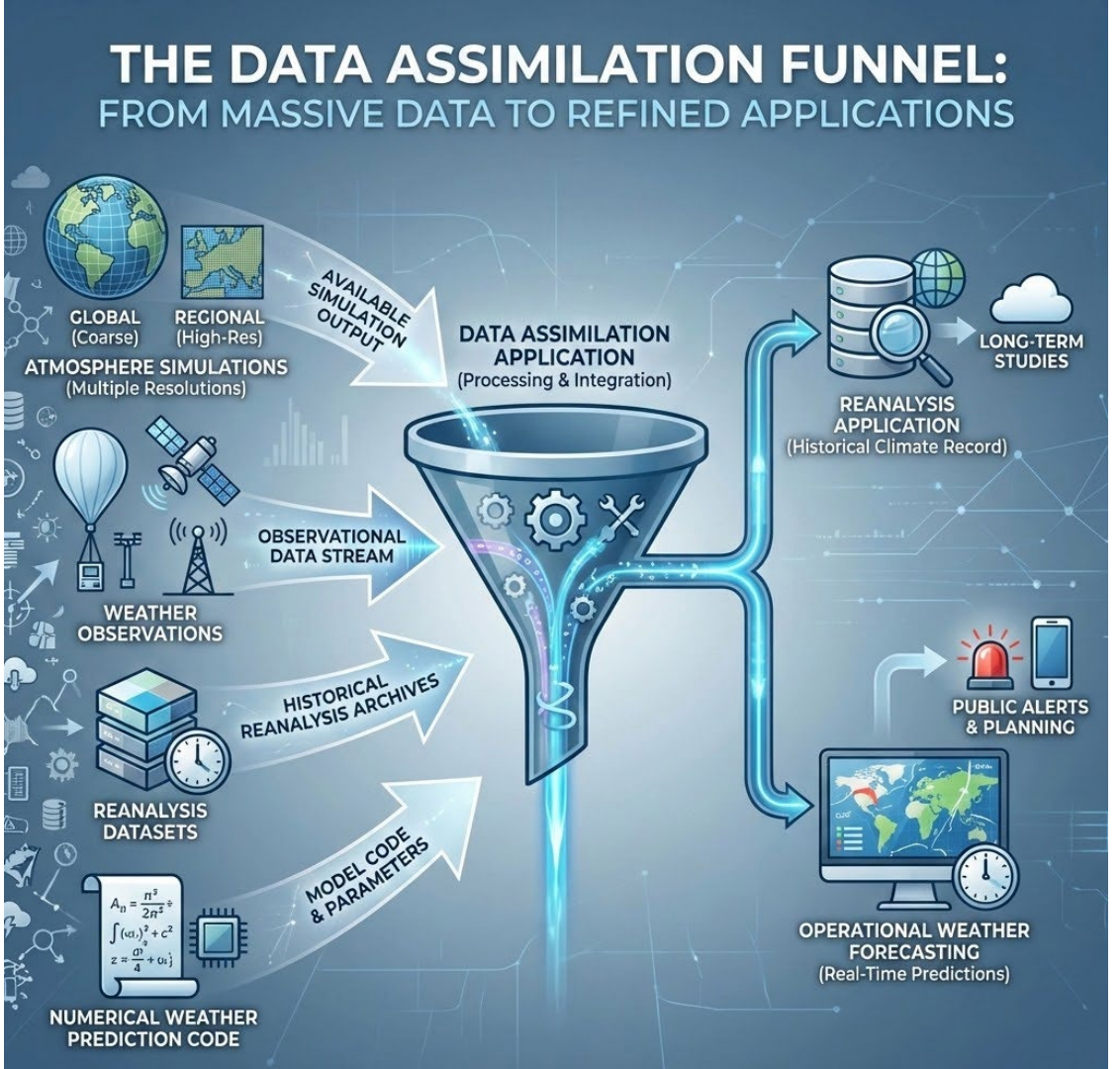
Collaborators:

Tim Smith (NOAA), Tse-Chun Chen (PNNL), Jason Platt (UCSD),
Kylen Solvik (CU Boulder / Columbia U.), Stephan Hoyer (Google Research),
Lucas Harris & FV3-SHiELD team (GFDL), Henry Abarbanel (UCSD)

5th ECMWF-ESA Machine Learning Workshop
15 April 2026

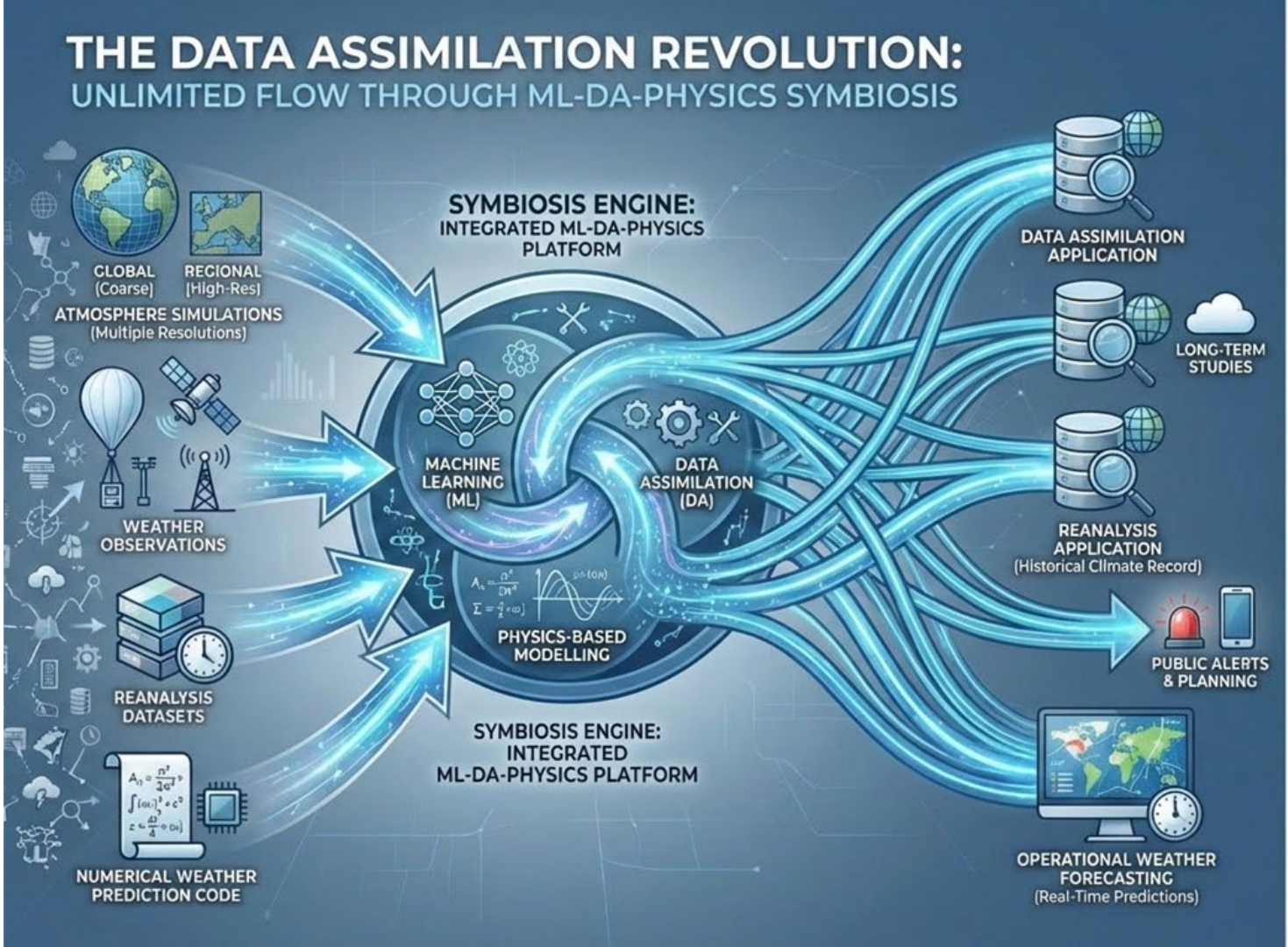
Data assimilation has historically been limited to:

- Primary applications:
  - operational forecasting
  - reanalysis
- Features:
  - Short observation windows (6-12 hours)
  - Primarily for state estimation
  - Single model



AI should bring us a world in which:

- Data assimilation, physics, and ML interact seamlessly
- Vast sources of data and knowledge are fully leveraged



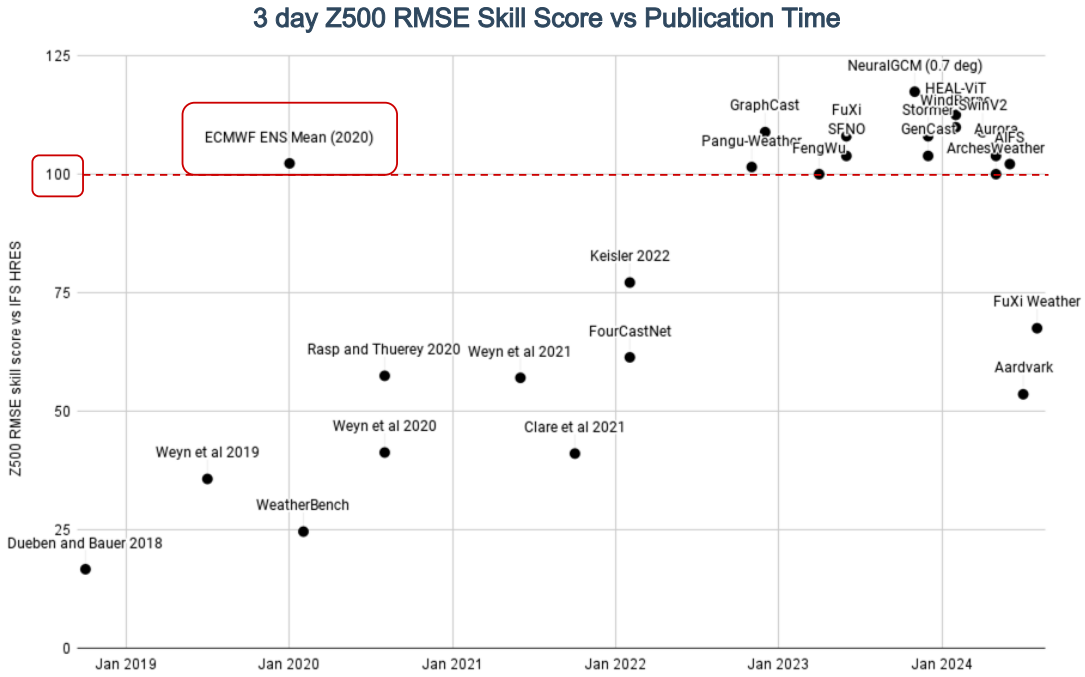
MLWP development over time

→ In the last 5 years, MLWP models have advanced rapidly

→ In the last 2-3 years, they seem to have plateaued

→ These models all depend on NWP inputs

→ Models including attempting an end-to-end solution show the weaknesses in MLWP and a more realistic picture of where they "really stand"



https://raspstephan.github.io/blog/ai-weather-progress/# Last updated November 2024

MLWP development over time

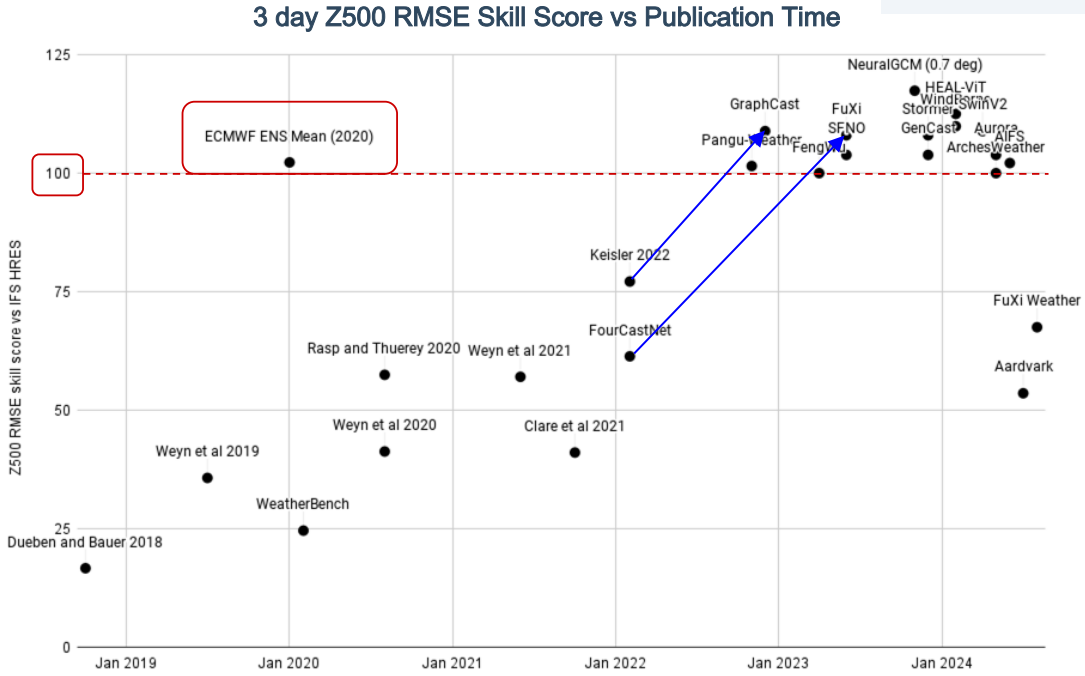
Big-tech driven advances (i.e. scale)

→ In the last 5 years, MLWP models have advanced rapidly

→ In the last 2-3 years, they seem to have plateaued

→ These models all depend on NWP inputs

→ Models including attempting an end-to-end solution show the weaknesses in MLWP and a more realistic picture of where they "really stand"



https://raspstephan.github.io/blog/ai-weather-progress/# Last updated November 2024

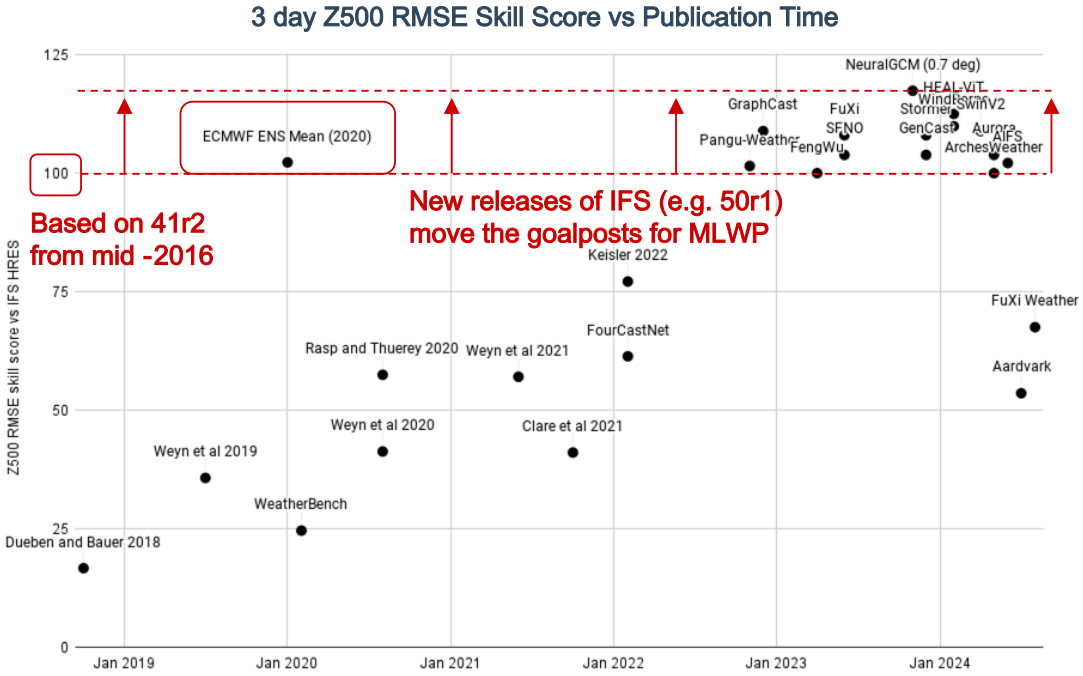
NWP development over time

→ In the last 5 years, MLWP models have advanced rapidly

→ In the last 2-3 years, they seem to have plateaued

→ These models all depend on NWP inputs

→ Models including attempting an end-to-end solution show the weaknesses in MLWP and a more realistic picture of where they "really stand"



https://raspstephan.github.io/blog/ai-weather-progress/# Last updated November 2024

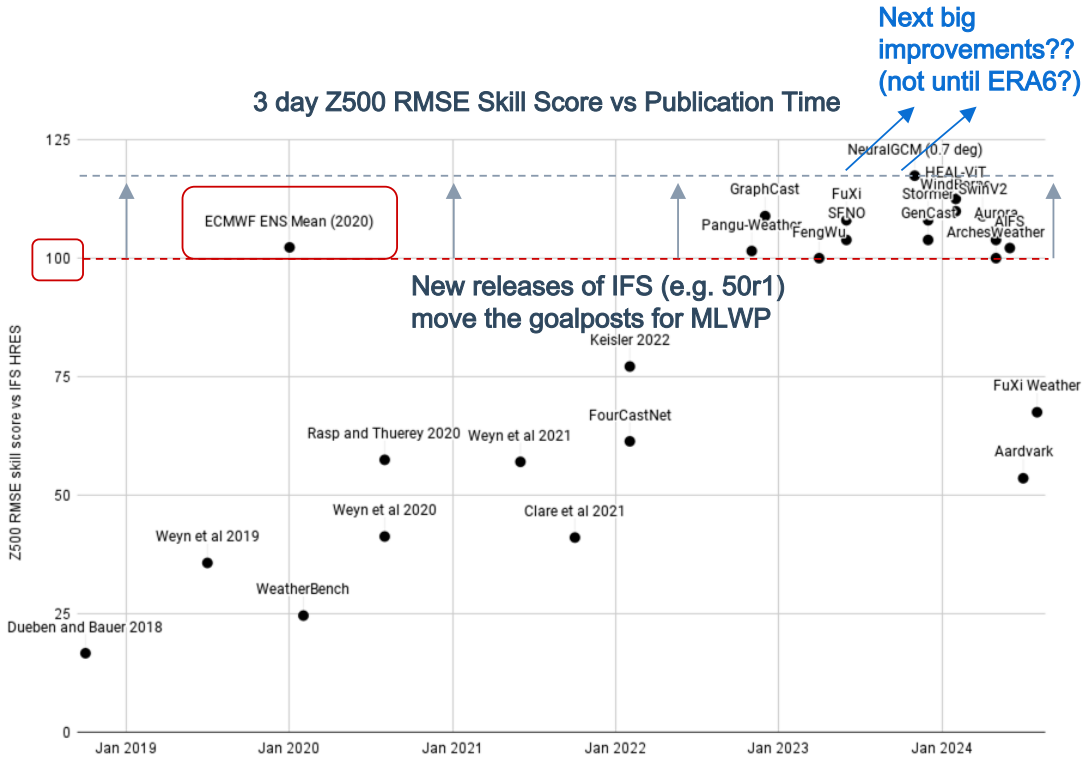
MLWP development *in the future*

→ In the last 5 years, MLWP models have advanced rapidly

→ In the last 2-3 years, they seem to have plateaued

→ These models all depend on NWP inputs

→ Models including attempting an end-to-end solution show the weaknesses in MLWP and a more realistic picture of where they "really stand"



https://raspstephan.github.io/blog/ai_-weather-progress/# Last updated November 2024

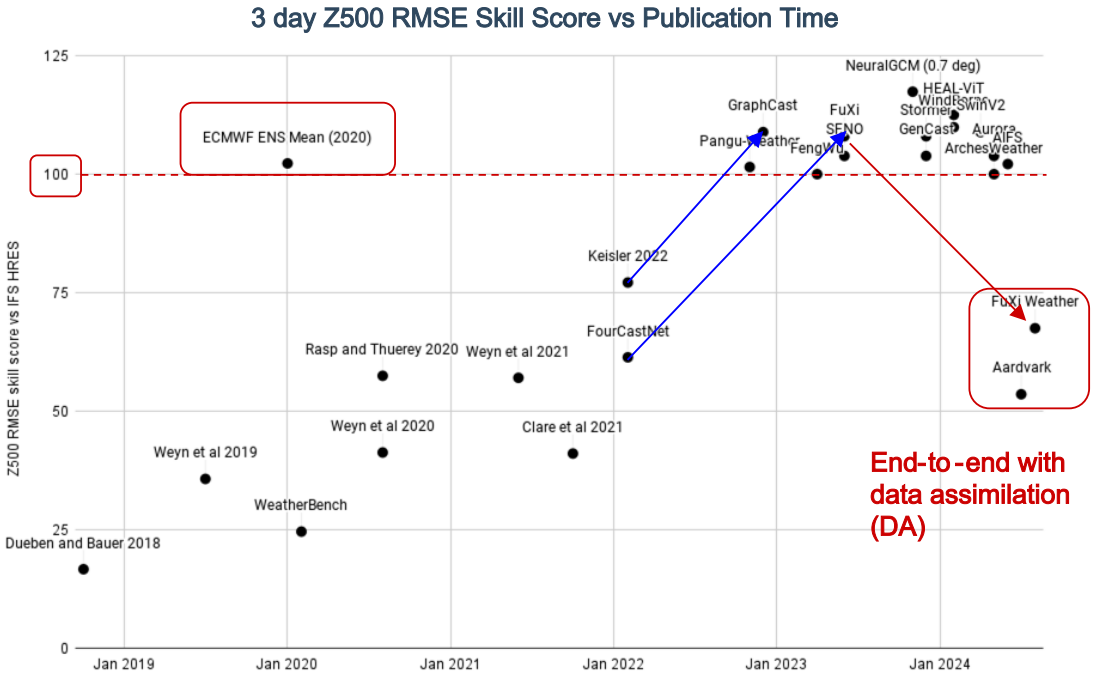
MLWP development over time

→ In the last 5 years, MLWP models have advanced rapidly

→ In the last 2-3 years, they seem to have plateaued

→ **These models all depend on NWP inputs**

→ **Models including attempting an end-to-end solution show the weaknesses in MLWP and a more realistic picture of where they "really stand"**



https://raspstephan.github.io/blog/ai-weather-progress/# Last updated November 2024

What are people trying for ML-DA?

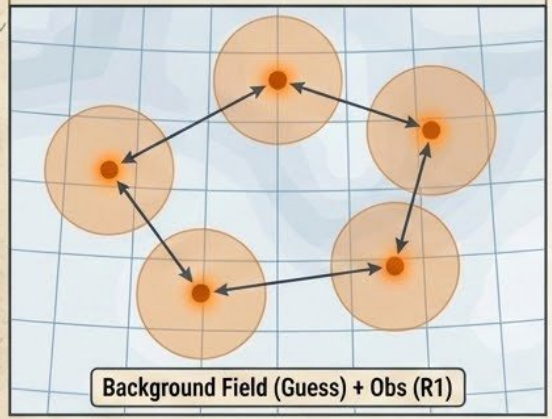
Approach	Modelling Philosophy	The "Traps"	Critique
Direct Mappers	"DA is just In-painting" (Spatial interpolation).	**Hallucination Trap:** High-resolution "fill" with zero dynamical balance.	**"Deep Cressman" Analysis:** Operates as a sophisticated interpolator that maps observations to grid points without a physical prognostic loop.
Ensemble Hybrids	"DA needs cheap Ensembles" (ML as generator).	**The Analog Trap:** The **B** matrix represents historical patterns, not "Errors of the Day."	**Conditional Climatology:** The ensemble spread is a "Library of Past Errors" rather than a prognostic simulation of current chaotic instabilities.
Learned Solvers	"DA is an Optimization Problem" (Differentiable solvers).	**The Adjoint Crisis:** ML gradients are often too noisy/localized for precision DA.	**The Sensitivity Problem:** The adjoints of neural networks often produce "spiky," unphysical sensitivities that fail to preserve dynamical balance during minimization.

CRESSMAN ANALYSIS (1959): AN EARLY OBJECTIVE ANALYSIS METHOD

Ingesting Observations into Numerical Weather Prediction Models

PASS 1: LARGE-SCALE CORRECTION

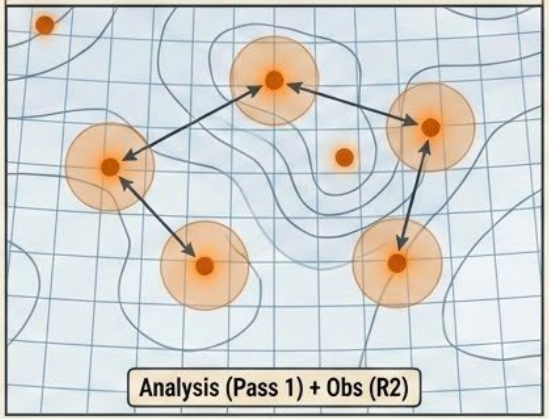
The process begins with a background field (e.g., short-range forecast) on a grid. Observations are used to correct this field using a large radius of influence (R1), capturing broad atmospheric features.



Background Field (Guess) + Obs (R1)

PASS 2: REFINING THE FIELD

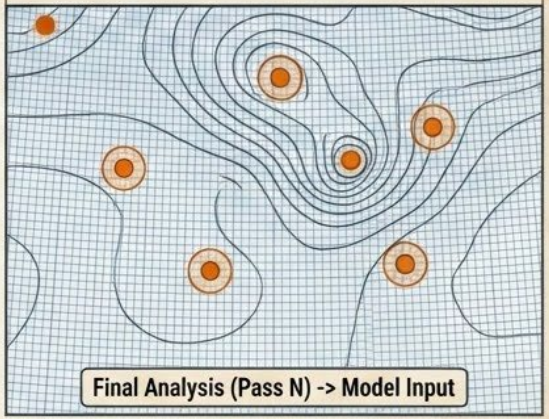
The analysis from Pass 1 becomes the new background. The process is repeated with a smaller radius of influence (R2), allowing for more localized corrections and resolving smaller-scale details.



Analysis (Pass 1) + Obs (R2)

FINAL ANALYSIS: DETAILED FIELD

After multiple passes with progressively smaller radii, a final, detailed analysis is produced. This gridded field is then used as the initial condition for the numerical forecast model.



Final Analysis (Pass N) -> Model Input

THE CRESSMAN WEIGHTING FUNCTION

$$W = \frac{R^2 - d^2}{R^2 + d^2} \quad \text{for } d < R$$

Weights decrease with distance (d) from the grid point, becoming zero at the radius of influence (R).

A pioneering iterative method that paved the way for modern data assimilation, despite its limitations in handling error statistics and anisotropic features.

#1: The Hallucination Trap (The "Deep Cressman" Danger)

- **The Concept:** Direct ML-DA treats assimilation as a "latent in-painting" problem. By mapping sparse observations directly to a dense grid via spatial pattern recognition, it effectively operates as a 21st-century Successive Correction Method (like Cressman Analysis of 1959), acting as a purely spatial interpolator without physical constraints.

- **The Critique:** Where Cressman filled data voids by safely relaxing to a smooth climatology, modern **ML fills voids by "in-painting" sharp, structurally complex features (hallucinations)** based on historical analogs. Because this interpolation is purely statistical—ignoring multivariate mass-wind coupling—the analysis state is physically unbalanced.

The history of Data Assimilation is a history of moving ***away*** from spatial interpolation (Cressman/OI) and ***toward*** dynamical integration (4D-Var).

What is the root of most of the problems here?

What is the root of most of the problems here?

The models.

What is the root of most of the problems here?

The models.

Category 1: Regression-to-the-Mean Emulators

Category 2: Distribution -Mapping Generators

Category 3: Differentiable Hybrid Physics/ML Operators

Category 1: Regression-to-the-Mean Emulators

Models: GraphCast (Google DeepMind), Pangu-Weather (Huawei), FourCastNet (NVIDIA), original FuXi.

These models are trained to minimize MSE/L1 loss against a single trajectory (ERA5). From a DA perspective, their critical characteristics are:

- **The "Dead" Jacobian:** Because they are trained to find the conditional mean $\mathbf{E}[x_{6hr} \mid x_{0hr}]$, their internal gradients (the Tangent Linear Model) are essentially **projection operators onto the low-frequency manifold**. They "kill" the growing Lyapunov vectors because those vectors represent the chaos the model was trained to ignore.
- **Spectral Quenching:** They capture the k^{-3} regime but lack the $k^{-5/3}$ tail. For DA, this means the background state $\mathbf{x}^b$ lacks the "sharpness" needed to resolve observation innovations.
- **Zero Internal Variance:** They provide a deterministic point estimate. To use them in DA, you have to "invent" an error covariance $\mathbf{B}$ from scratch (e.g., using a static climatological $\mathbf{B}$), as the model itself offers no information on uncertainty.

Category 2: Distribution - Mapping Generators

Models: GenCast (Google DeepMind), AIFS (ECMWF- probabilistic version), FuXi-S, FCNv3.

These models move from regression to **probabilistic sampling** (Diffusion, GANs, or CRPS-optimized stochastic nets).

- **Spectral Recovery (The "Hallucination" Feature):** Unlike Category 1, these models *recover the $k^{-5/3}$ energy* tail - i.e. they produce "sharp" weather. However, for DA, this sharpness is **stochastically independent** of the current observations. They provide a *statistically* plausible state, but not necessarily a *dynamically* or *physically* plausible state.
- **The "Ensemble Prior":** They can natively support Ensemble DA because they can generate a "plausible" backgrounds distribution. However, the result is closer to **Ensemble Optimal Interpolation (EnOI)** than an EnKF
- **The Sensitivity Gap:** While they produce a spread, the spread is conditioned on the *training data's* variance, not the *current* dynamical instability. They essentially provide a **Conditional Climatological B** (hence EnOI), not a truly flow-dependent **B** that knows where the "Errors of the Day" are growing.

Category 3: Differentiable Hybrid Physics/ML Operators

Models: NeuralGCM (Google/ECMWF), ACE (Allen Institute for AI).

This is the most "DA-friendly" category because it doesn't try to emulate the whole atmosphere; it only emulates the parts we don't understand or can't model well

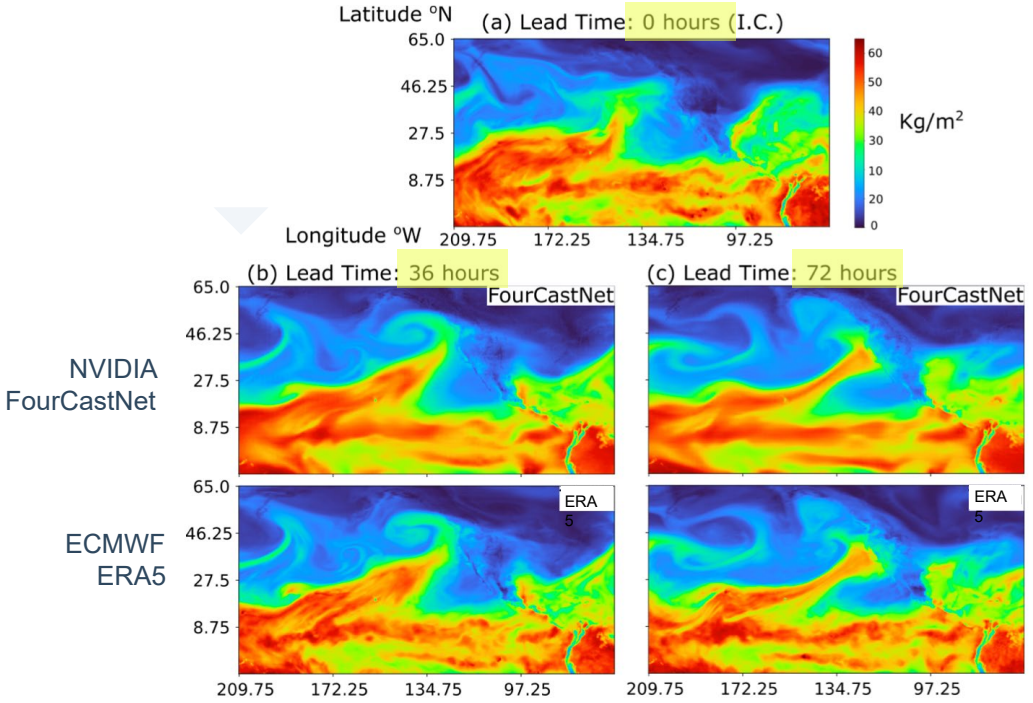
- **Valid (or quasi-valid) Dynamical Jacobian:** For those models that retain a **differentiable dynamical core** (solving the actual Navier-Stokes equations), their Tangent Linear and Adjoint models are physically grounded. They (may) respect conservation laws.
- **Dynamical Balance:** In theory, increments added via DA are less likely to be "rejected" or "shocked" because the model's internal state is governed by fluid dynamics, not just neural weights.
- **Multi-Scale Coupling:** They naturally bridge the k^{-3} and $k^{-5/3}$ regimes by using physics and ML to complement one another.
- **More Costly:** These approaches are not without drawbacks - they are costly even on GPUs and can negate a lot of the benefits of transitioning the technology.

What are other common “traps” that we’re being caught in?

#2: The MSE Trap (Incompatible with Chaos)

Nearly all deterministic MLWP models result in a ‘blurred’ forecast, starting almost immediately (with the first time step)

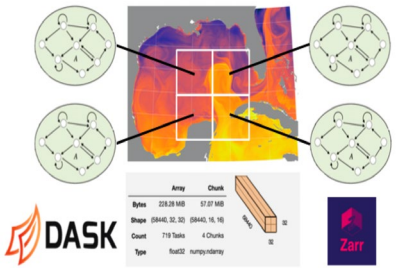
- **The Concept:** Training models to minimize Mean Squared Error (Category 1) targets the conditional mean **E[Y|X]**.
- **The Critique:** This eliminates the chaotic variance in the $k^{-5/3}$ regime—the exact space where leading Lyapunov exponents and singular values live. **You cannot "Fight Chaos"** with a model that has been mathematically trained to filter it out.

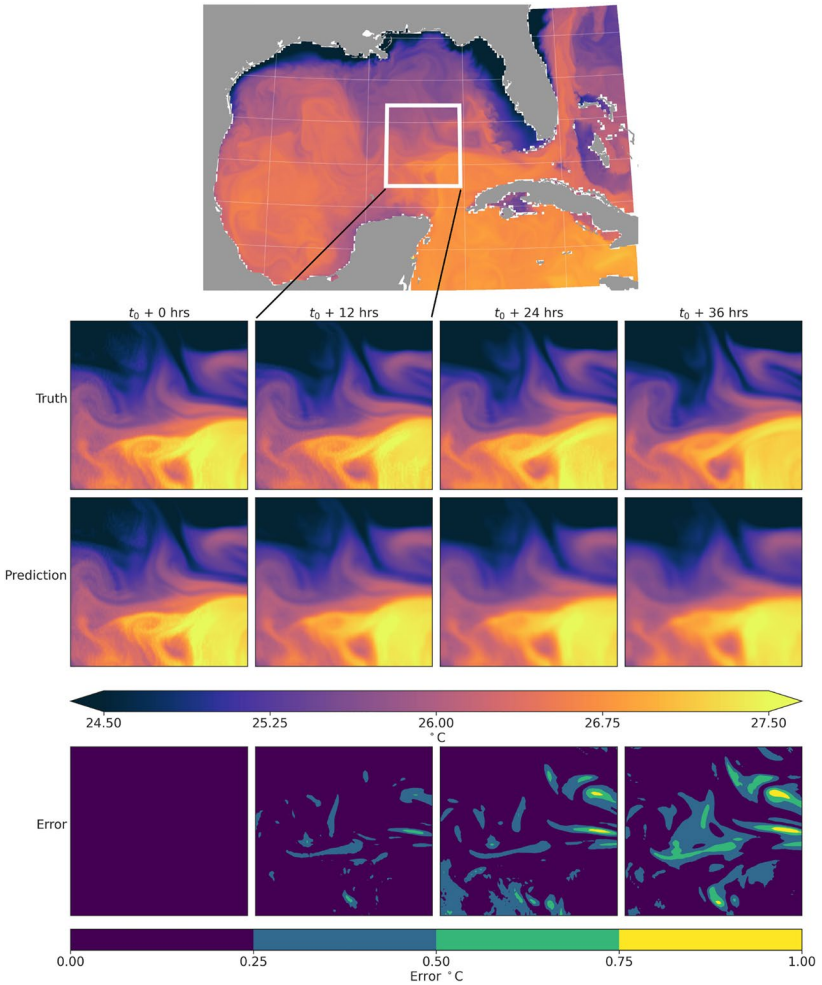


Is the MSE loss function really to blame?

Software: Smith et al., (2024).
xesn: Echo state networks powered by Xarray and Dask.
Journal of Open Source Software, 9(103), 7286,
https://doi.org/10.21105/joss.07286

Designed to address:
1. deploying the code on GPUs,
2. interacting with a parameter optimization algorithm in order to tune the model, and
3. parallelizing the architecture for higher dimensional applications

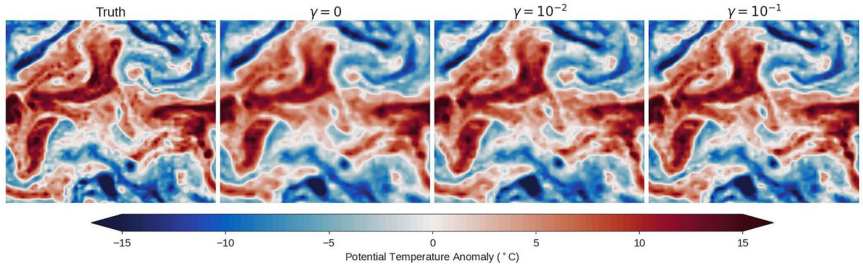


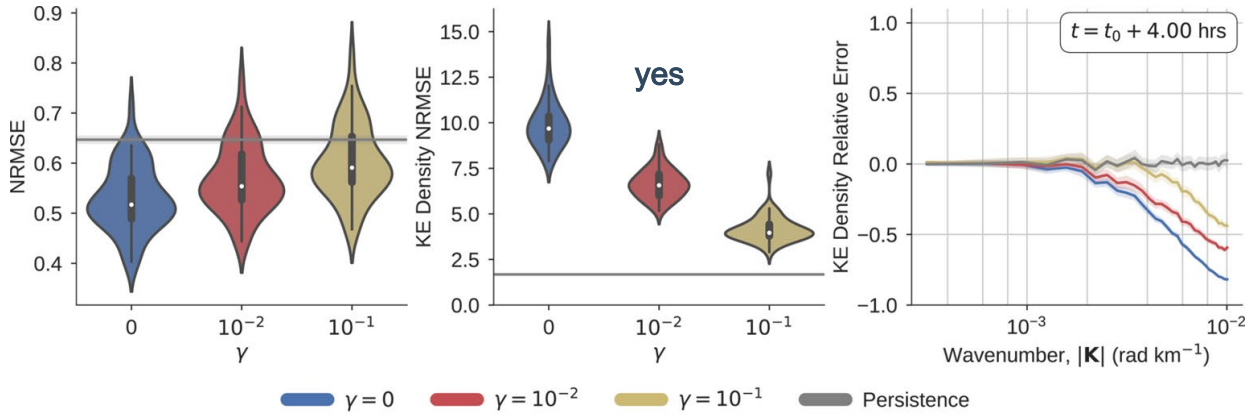


Smith, **Penny**, Platt, Chen (2023) "Temporal Subsampling Diminishes Small Spatial Scales in Recurrent Neural Network Emulators of Geophysical Turbulence"
https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2023MS003792

Is the MSE loss function really to blame?

Tuning the weighting on an added KE term in the loss function improves the fit of the kinetic energy spectrum while degrading the overall RMSE

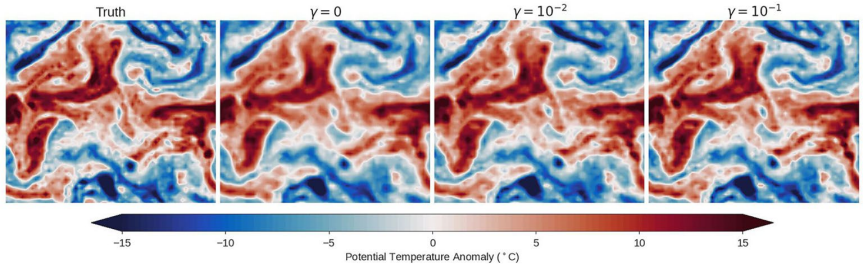


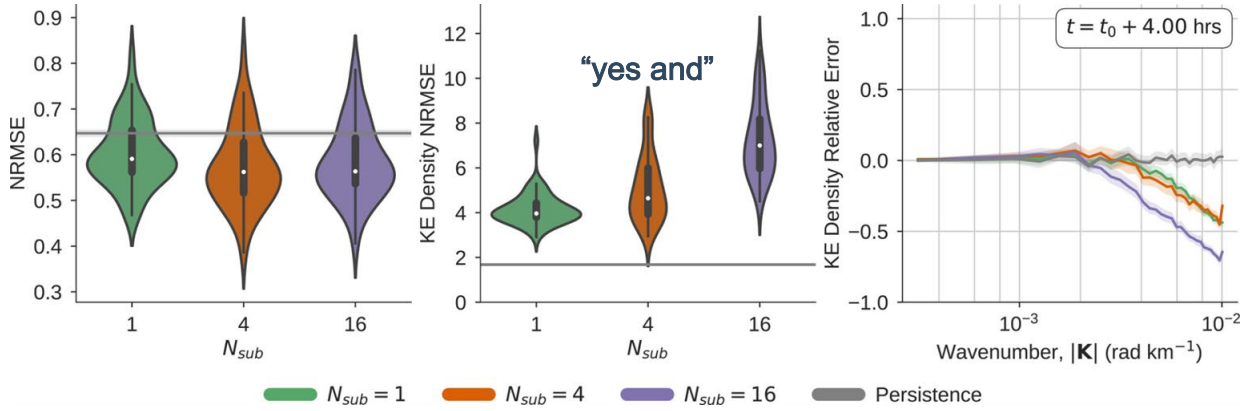


Smith, **Penny**, Platt, Chen (2023) "Temporal Subsampling Diminishes Small Spatial Scales in Recurrent Neural Network Emulators of Geophysical Turbulence" https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2023MS003792

Is the MSE loss function really to blame?

Increasing the time step of the data minimally degrades the RMSE but **significantly increases the error of the kinetic energy spectrum**





Consider the difference between a 6-hour time step used by most MLWP models and the NWP model timestep between 2.5 to 12 minutes

Smith, **Penny**, Platt, Chen (2023) "Temporal Subsampling Diminishes Small Spatial Scales in Recurrent Neural Network Emulators of Geophysical Turbulence"
https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2023MS003792

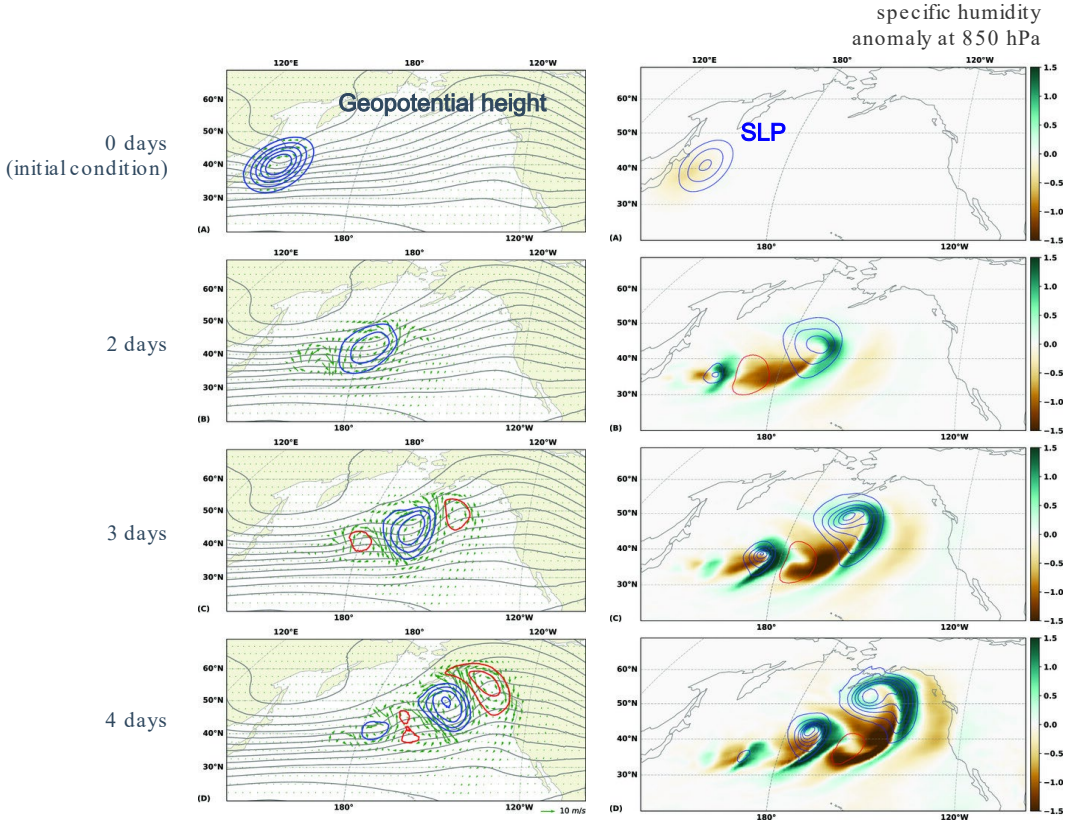
Category 1 MLWP models resolve k^3 scales

Solution at 500 hPa for a localized disturbance on the DJF atmosphere. The “time evolution of a localized 500-hPa trough at the western end of the North Pacific storm track, which is the canonical initial condition preceding surface cyclogenesis.”

At large scales, MLWP models produce “signal propagation and structural evolution qualitatively in accord with previous research in meteorology”

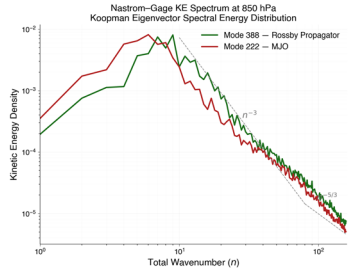
Left - Solution at 500 hPa for a localized disturbance on the DJF-averaged atmosphere state using PanguWeather. Geopotential height is shown by gray lines, every 60 m.

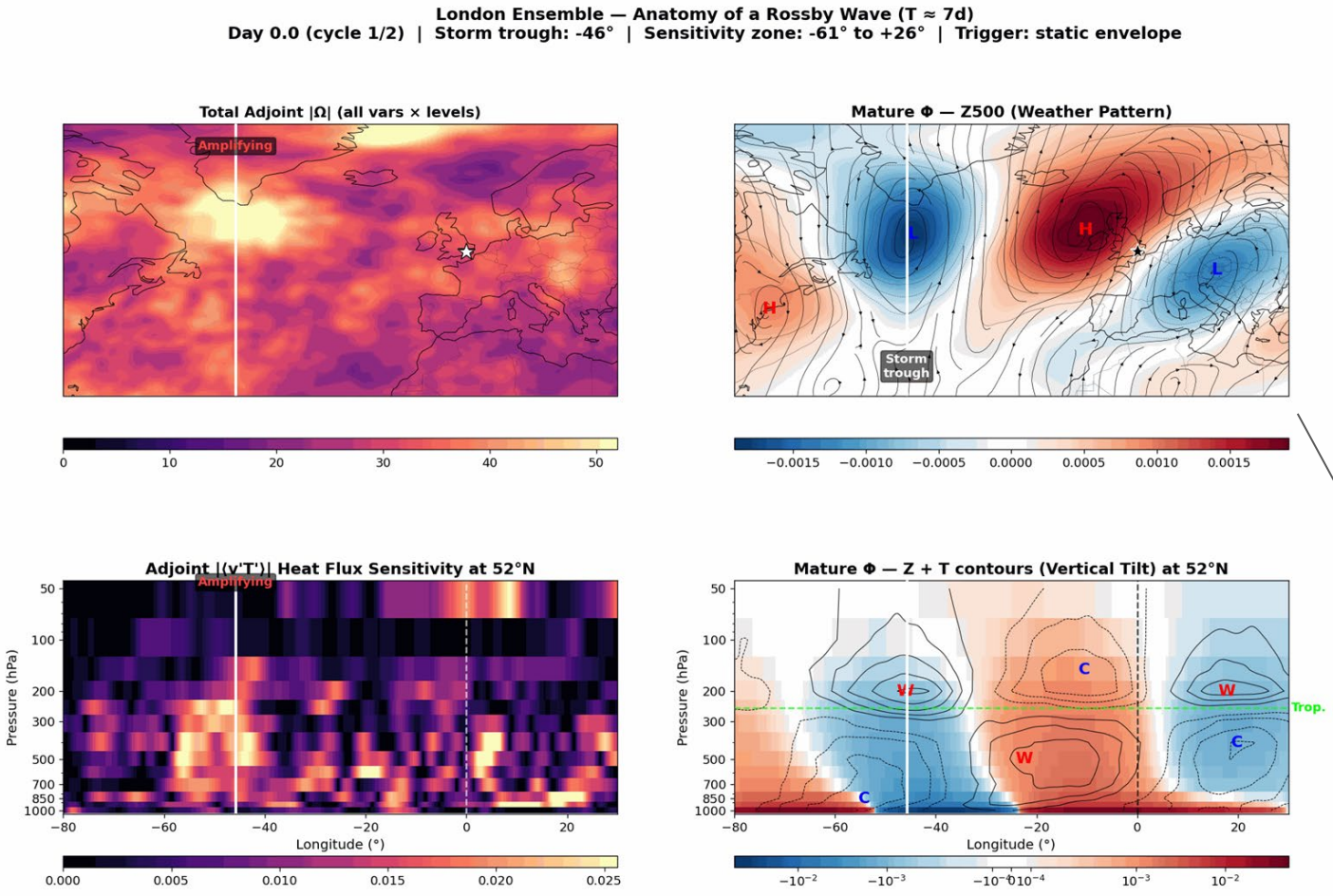
Right - Contour: Anomalies in mean sea level pressure. Shaded: Water vapor specific humidity anomalies (g kg−1) at 850 hPa.



Hakim and Masanam (2024) “Dynamical Tests of a Deep Learning Weather Prediction Model”https://journals.ametsoc.org/view/journals/aies/3/3/AIES-D-23-0090.1.xml

A linear model can do this too…





Here a global 3D LIM/Koopman estimator evolves Rossby waves in a wavetrain crossing the N. Atlantic.

Solution at 500 hPa for a localized disturbance on the DJF atmosphere. The “time evolution of a localized 500-hPa trough at the western end of the North Atlantic storm track

Penny (2026, in prep)

#3: The k^{-3} Trap (The Baseline Illusion)

- **The Concept:** Most MLWP "breakthroughs" are measured in the synoptic, large-scale regime where the atmosphere is relatively deterministic.
- **The Critique:** High skill at these scales is not a revolutionary feat; even simple linear models or low-resolution NWP can perform well here. It is the "Predictability Baseline," not the ultimate goal.

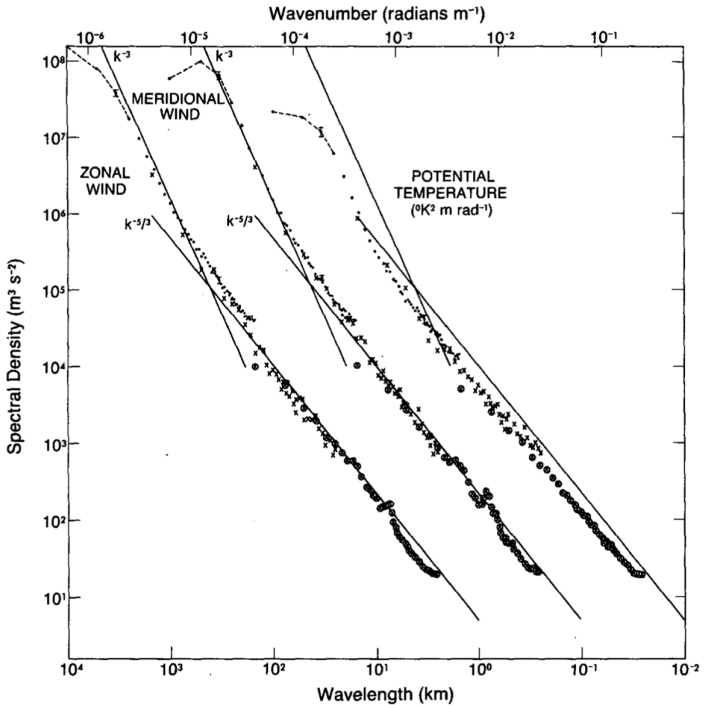
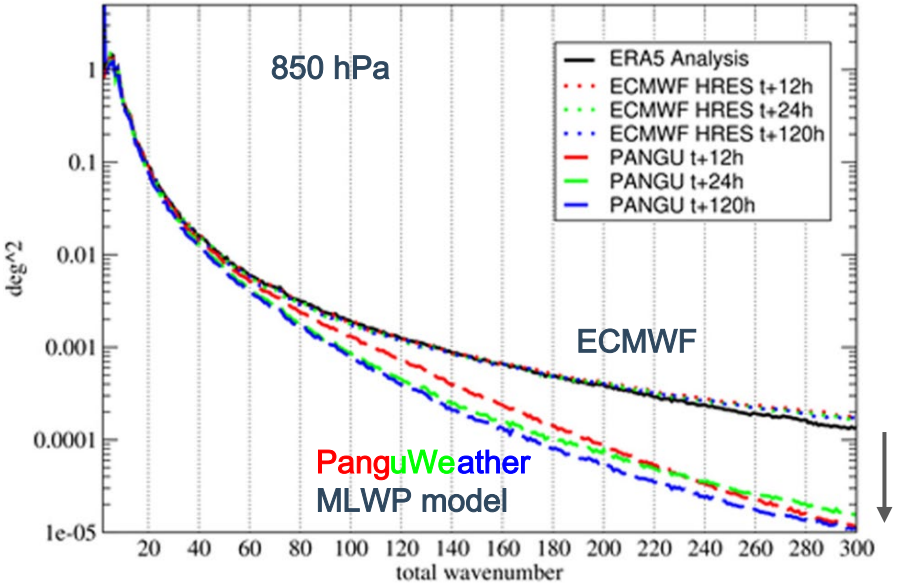


FIG. 3. Variance power spectra of wind and potential temperature near the tropopause from GASP aircraft data. The spectra for meridional wind and temperature are shifted one and two decades to the right, respectively; lines with slopes −3 and $-5/3$ are entered at the same relative coordinates for each variable for comparison.

Flagging the “Blurriness” issue

Bonavita (2024) identified significant problems with the MLWP models that have not yet been resolved in deterministic MLWP models.

Power spectral density as a function of total wavenumber at 850 hPa for temperature.



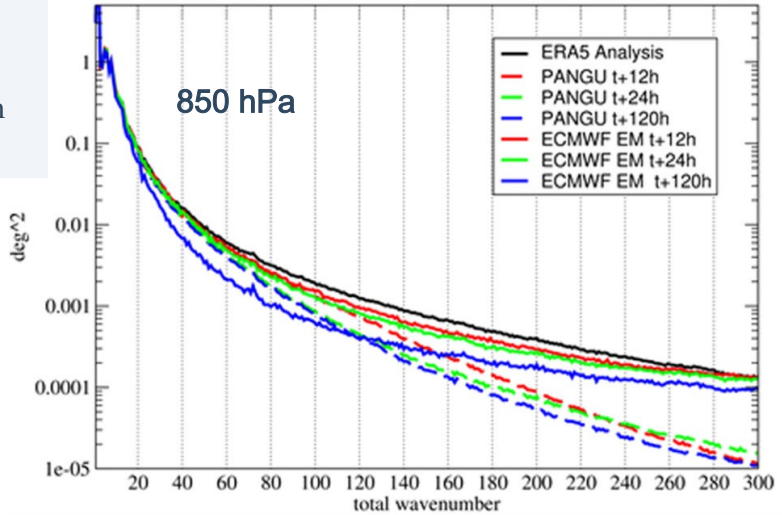
The results “**quantitatively confirm that the ML models produce less spectrally resolved forecasts than the analysis fields used in their training** and those produced by the ECMWF IFS forecasts. The effective resolution of the ML models' forecasts **is closer to 500 –700 km than to the nominal 0.25°** and is gradually decreasing with forecast lead time.” *(Bonavita 2024)*

Bonavita (2024) identified significant problems with the MLWP models that have not yet been resolved in deterministic MLWP models.

Flagging the “Blurriness” issue vs. ENS forecasts

“Blurry” forecasts indicate a **problem** - the MLWP forecasts **do not agree** with the ensemble mean

Power spectral density as a function of total wavenumber at 850 hPa for temperature (vs. ECMWF ENS Ensemble Mean)



What is the “disconnect”?
NWP and ML mean different things by “the mean”

The Ensemble Mean (NWP): This is a **prognostic** average. You take today's specific state, add perturbations that represent today's instabilities, and see where they go. The mean is a representative point of those *specific* futures.

The Statistical Mean (ML): This is a **diagnostic** average. It is effectively an **"Average of Analogs."** The model looks at the current state **x** and, in its high-dimensional weights, essentially asks: *"In the last 40 years, when the weather looked roughly like this, what was the average outcome 6 hours later?"*

Bonavita (2024) identified significant problems with the MLWP models that have not yet been resolved in deterministic MLWP models.

Under an MSE-objective, the ML model learns f(**X**) = **E**[**Y**|**X**]

(k^{-3}) ($k^{-5/3}$)

Law of total variance: Var(**Y**) = Var(**E**[**Y**|**X**]) + **E**[Var(**Y**|**X**)]

Translating into meteorology:

- Var(**Y**): The **Total Variance** of the real atmosphere (the full energy spectrum, including the turbulent $k^{-5/3}$ tail).
- Var(**E**[**Y**|**X**]): The **Variance of the ML Model** (since f(**X**) = **E**[**Y**|**X**]). This is the variance of the "average" states.
- **E**[Var(**Y**|**X**)]: The **Expected Conditional Variance** . This is the chaotic spread—the physical variance of all the possible storms that *could* happen given state **X**.

That means that for the optimal solution to the MSE loss function,
the variance is always underestimated :

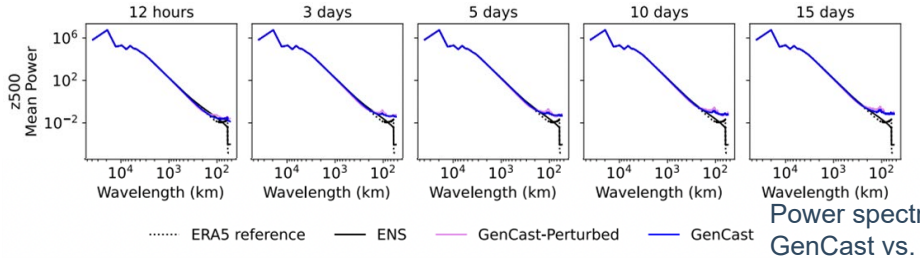
$$\mathrm{Var}(f(\mathbf{X})) = \mathrm{Var}(\mathbf{Y}) - \mathbf{E}[\mathrm{Var}(\mathbf{Y}|\mathbf{X})]$$

Generative AI recovers the $k^{-5/3}$ tail

A generative model (like a diffusion model) looks at the blurry output of an MSE forecast and notices that the $k^{-5/3}$ energy tail is missing. It then samples from its learned latent distribution to "paint" that texture back onto the map (i.e. fill in the missing $\mathbf{E}[\mathrm{Var}(\mathbf{Y}|\mathbf{X})]$).

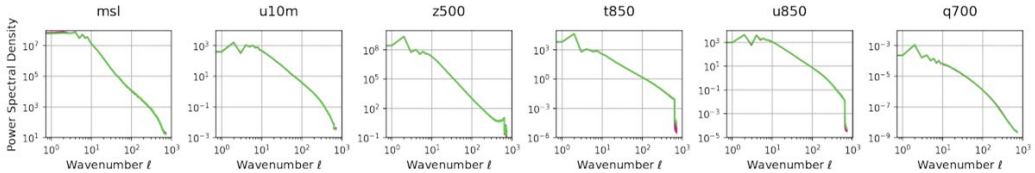
- It matches the **Power Spectrum** . The forecast *looks* like a real 4km atmospheric state.
- But it is entirely guessing the **Phase**.

→ DeepMind's GenCast model *(Price et al. 2025) Supplementary Materials*

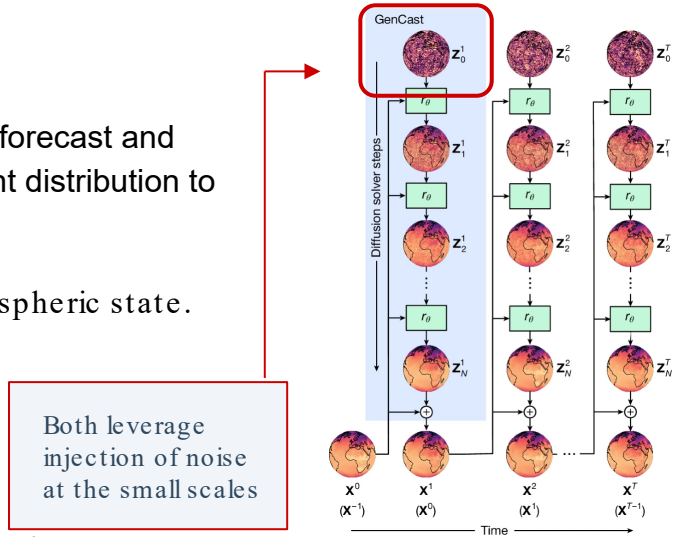


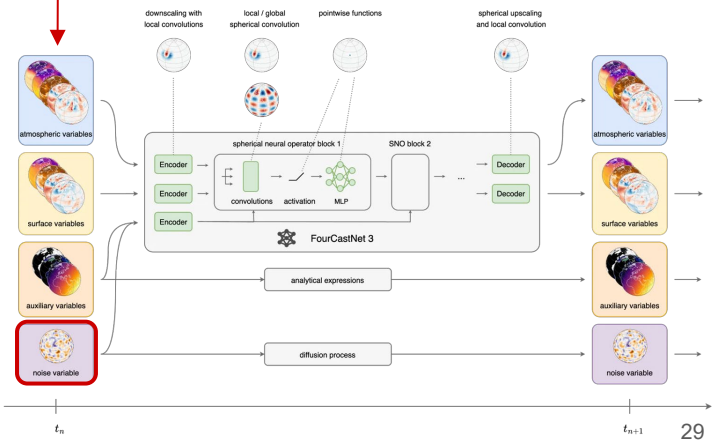
Power spectra for GenCast vs. ERA5 at z500

→ NVIDIA's FourCastNet3 *(Bonev et al. 2025)*



Power spectra for FourCastNet3 vs. ERA5 at 360 hours, averaged over 1 year





#4: The Generative $k^{-5/3}$ Trap (The Stochastic Forgery)

The Concept: Generative models solve the MSE blurring problem by injecting stochastically sampled noise to artificially recover the $k^{-5/3}$ energy spectrum.

The Critique: They match the **Power** (the “what” and “how much”) of the turbulence but hallucinate the **Phase** (the “where”). *At best*: They provide "Conditional Climatology" rather than a prognostic simulation of the "Errors of the Day." *At worst*: the **phase-blind noise creates spurious correlations**, destroys multivariate balance in the **B** matrix, and forces the DA system to fight physical ghosts – this is "Style Transfer," not fluid dynamics.

#5: The Adjoint Crisis (Noisy Loss Landscape Trap)

- **The Concept:** 4D-Var Data Assimilation is an optimization problem that requires a relatively smooth, convex "basin of attraction" to minimize the cost function (J). MLWP models possess internal gradients that create a shattered, highly non-convex, and noisy loss landscape.

- **The Critique:** When we attempt to use the ML model's backpropagation as a surrogate for the physical Adjoint (M^{H}), the gradients point toward numerical noise rather than physical reality. Because deep neural networks suffer from "gradient shattering" and lack physical inductive biases, the DA optimizer gets trapped in meaningless local minima, rendering the analysis state physically useless.

Comparing TLMs

The accuracy of the ensemble statistics is dependent on how the model responds to small perturbations in initial conditions.

Tangent linear model (TLM) response after 6 hours for zonal wind. Comparing (left) DeepMind's Graphcast to (right) NCAR's physics-based Model for Prediction Across Scales (MPAS-A)

Note large errors in the vertical response

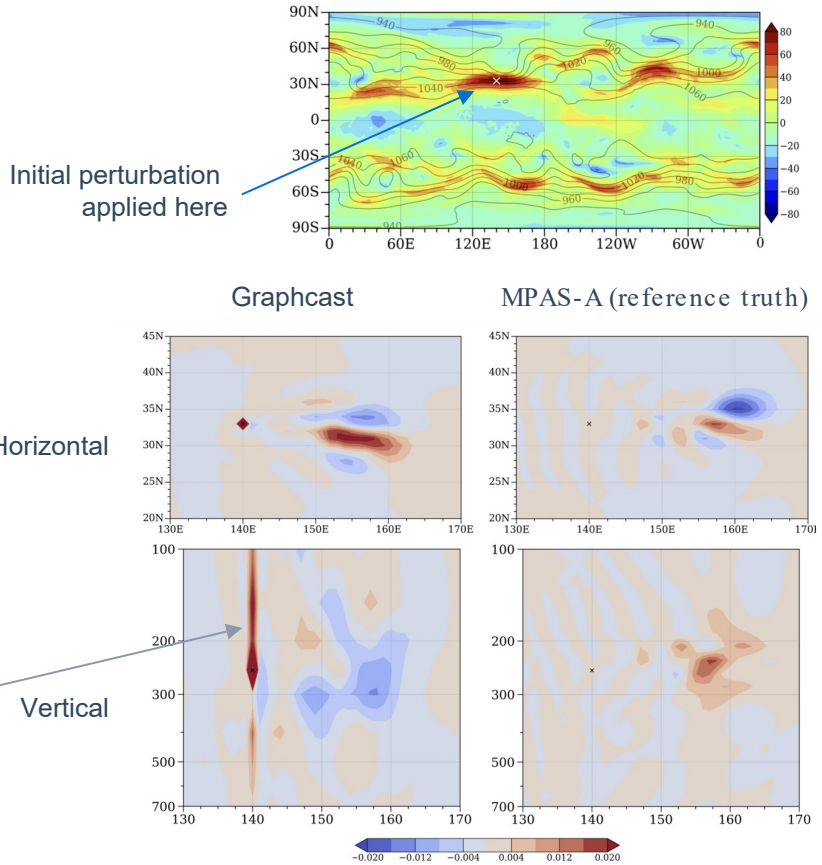
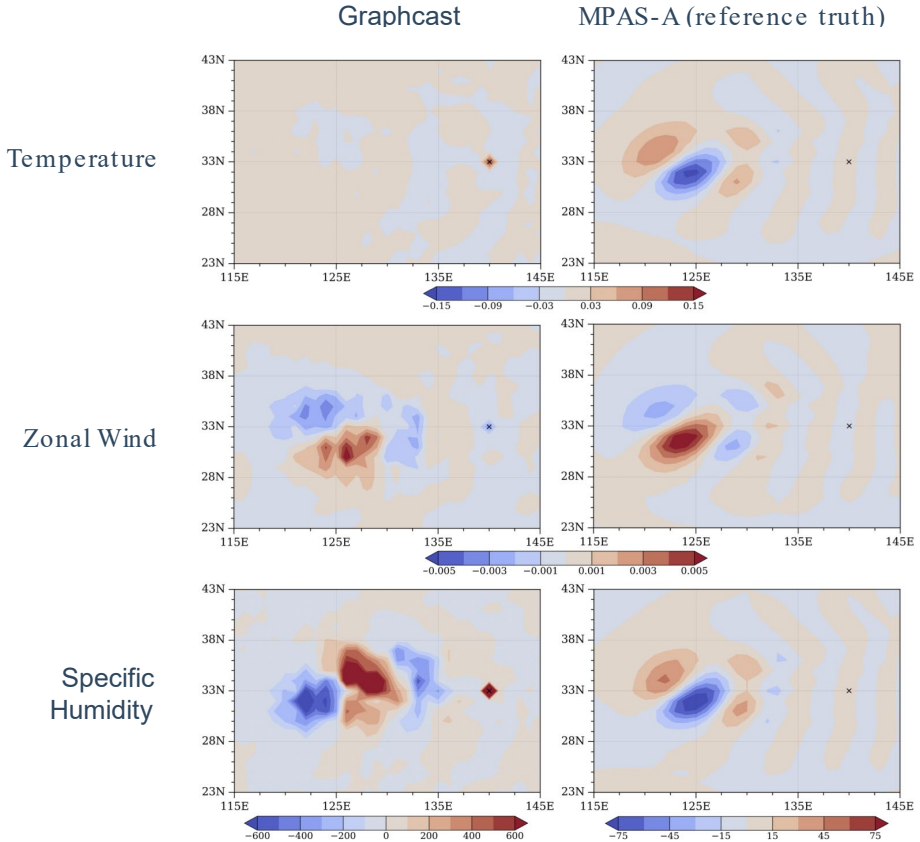


Figure 2: (a) Background geopotential heights (contoured) and zonal wind (shaded) at 00 UTC on January 1, 2022. The cross marks the location of the imposed perturbation. (b)-(c) Horizontal distribution of the TL response in zonal wind 6 hours into the forecast to a zonal wind perturbation at the initial time for GraphCast (left) and MPAS-A (right). (d)-(e) Vertical cross-sections of the TL response in zonal wind along the longitude line at 33°N for GraphCast (left) and MPAS-A (right).

Tian, Holdaway, Kleist (2024) "Exploring the use of Machine Learning Weather models in Data Assimilation" https://arxiv.org/pdf/2411.14677

Comparing Adjoints (horizontal)

The Adjoint sensitivity study determines the upstream impact of the model on a particular point, 6 hours prior.

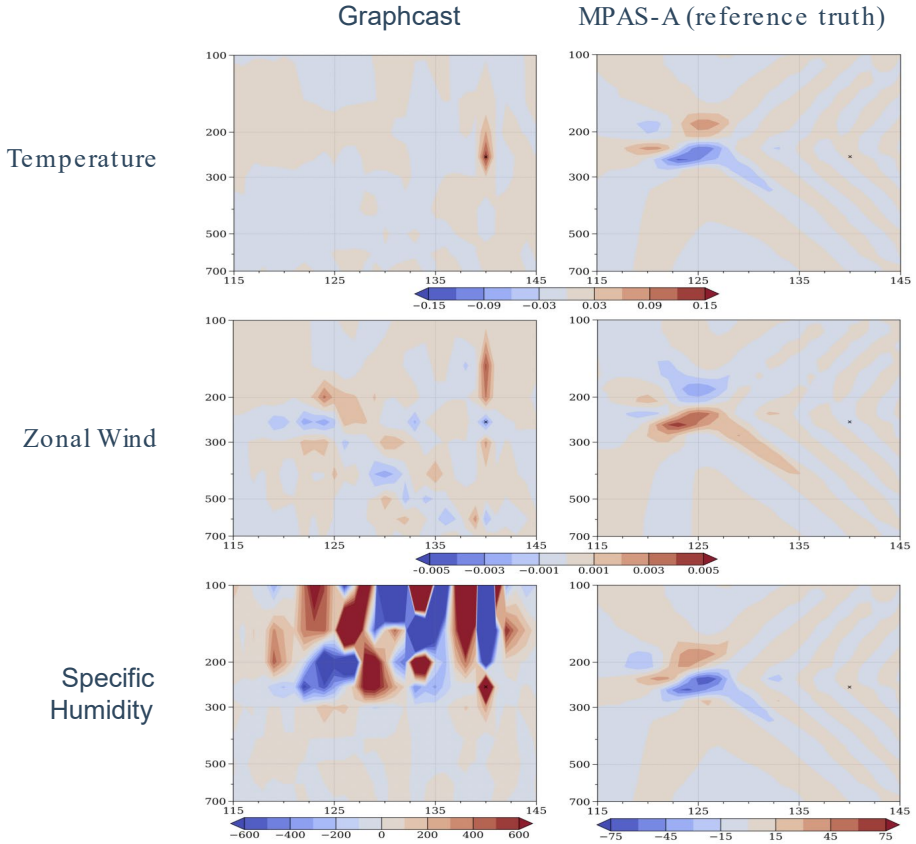


Tian, Holdaway, Kleist (2024) “Exploring the use of Machine Learning Weather models in Data Assimilation” https://arxiv.org/pdf/2411.14677

Comparing Adjoints (vertical)

The impacts in the vertical are particularly poor in the MLWP model.

Minimal impacts for temperature, noisy impacts for zonal winds, and large spurious impact for specific humidity.

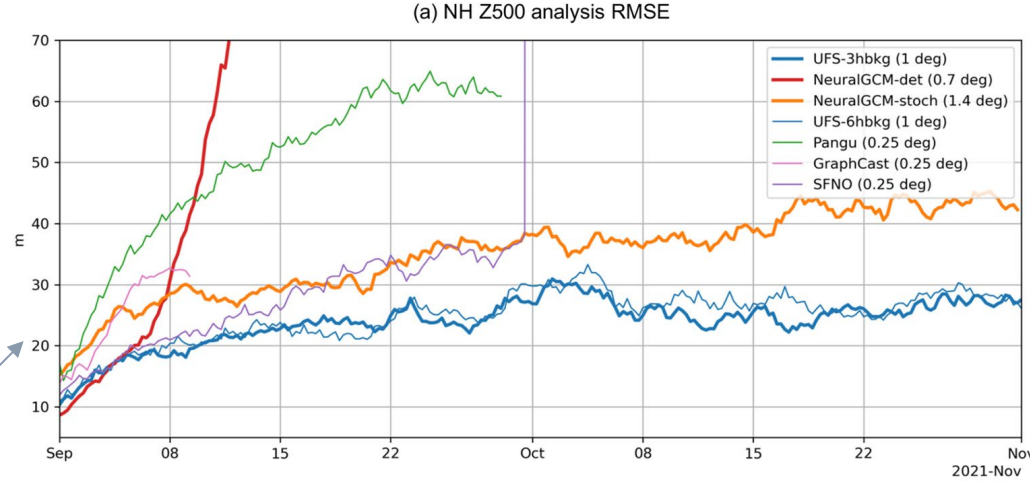


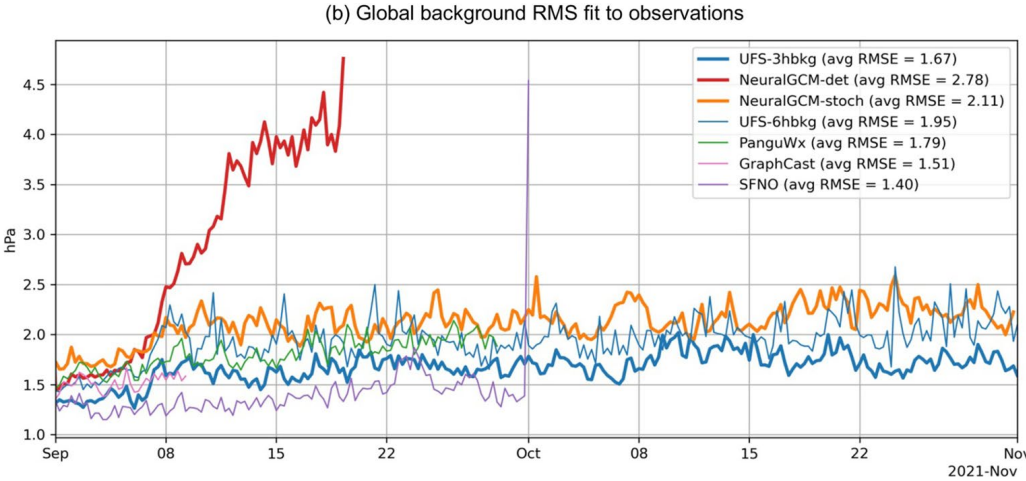
Tian, Holdaway, Kleist (2024) “Exploring the use of Machine Learning Weather models in Data Assimilation” https://arxiv.org/pdf/2411.14677

Comparing in an EnKF

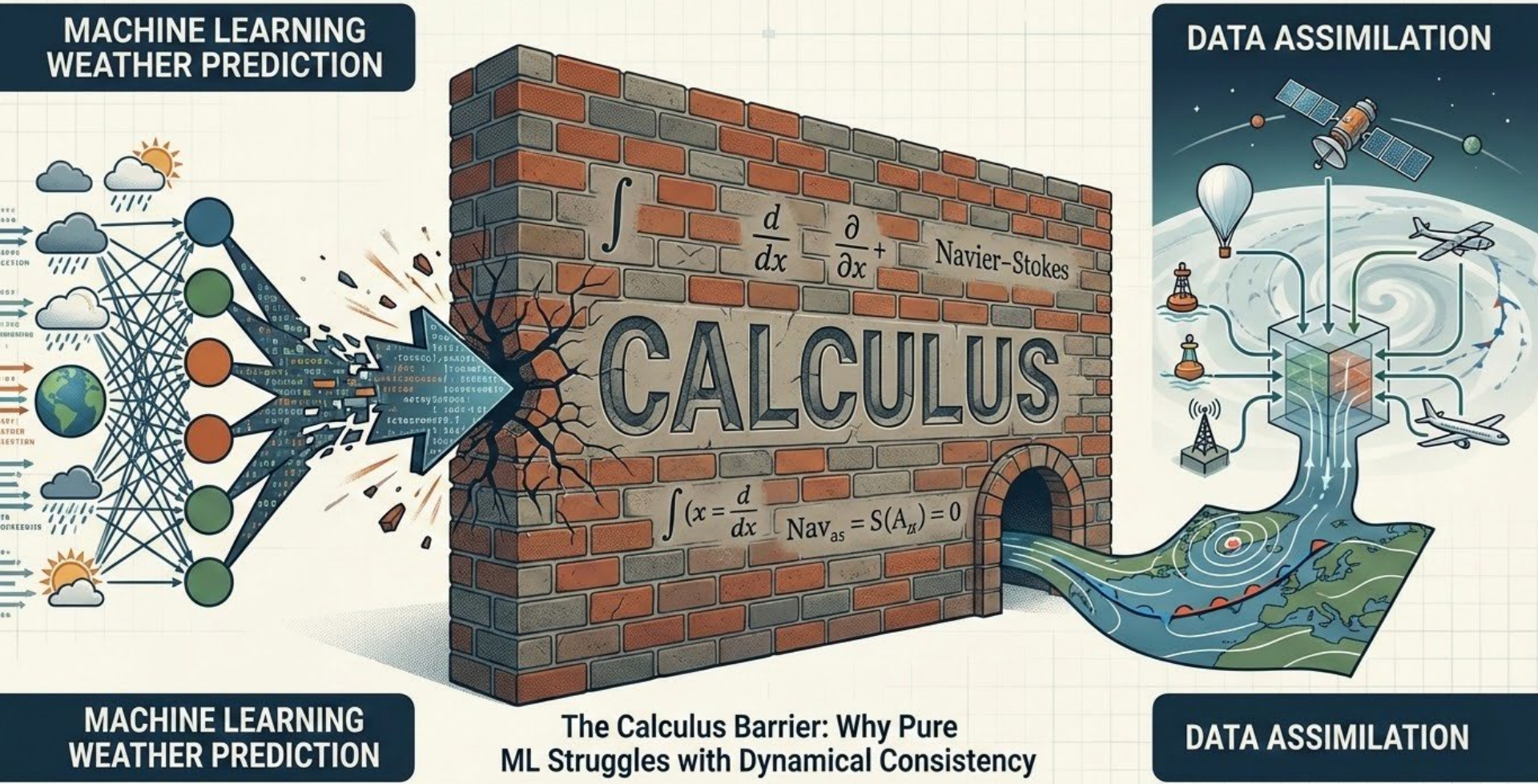
All of the MLWP models underperform the 1-degree NOAA Unified Forecast System (UFS) NWP model when applied within an 80-member ensemble Kalman filter.

Only the **hybrid physics/ML model** makes it beyond a 30-day cycling experiment.





Slivinski et al. (2025) “Assimilating Observed Surface Pressure Into ML Weather Prediction Models”
https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2024GL114396



The Calculus Barrier: Why Pure ML Struggles with Dynamical Consistency

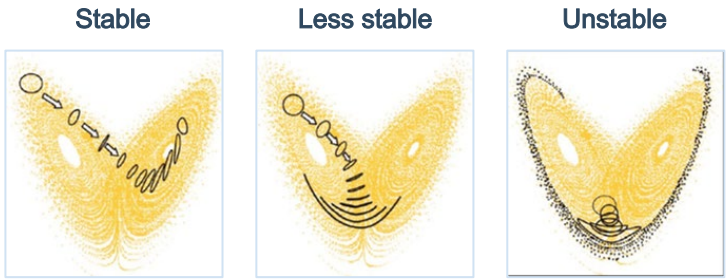
Why does it matter how the model responds to perturbations in initial conditions?

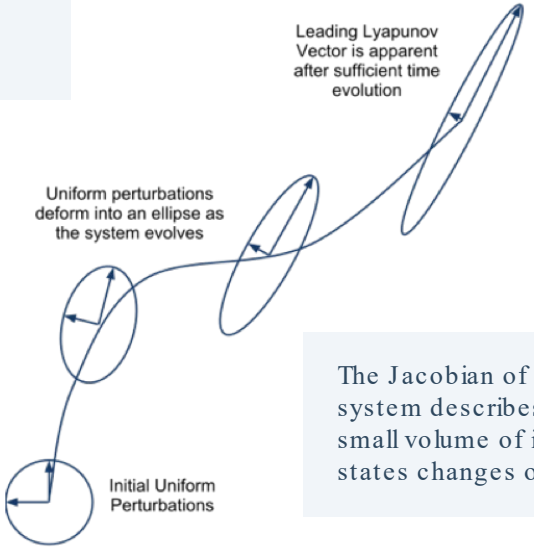
Weather is an archetypal example of a chaotic dynamical system

In 2005, Edward Lorenz visited my advisor Eugenia Kalnay in her office at U. Maryland. At some point during his stay, he penned this on a piece of paper - which later hung on her door for the entire duration of my Ph.D.:

"Chaos: When the present determines the future, but the approximate present does not approximately determine the future."

Predictability depends on the initial conditions (Palmer, 2002):





The Jacobian of the system describes how this small volume of initial states changes over time.

We “fight chaos” by understanding error growth

Small perturbations grow exponentially in some directions, and decay in others.

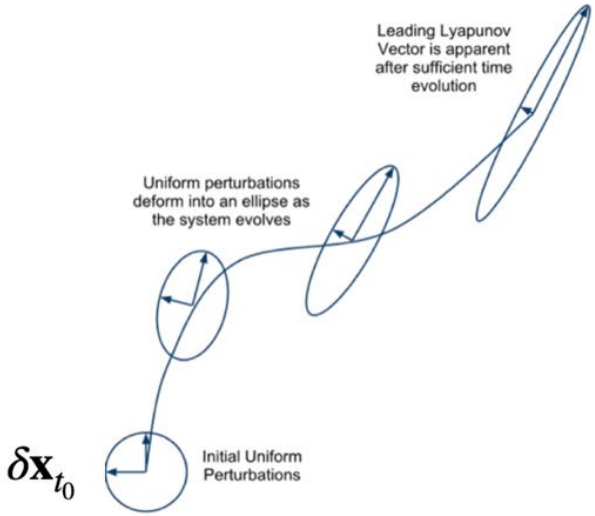
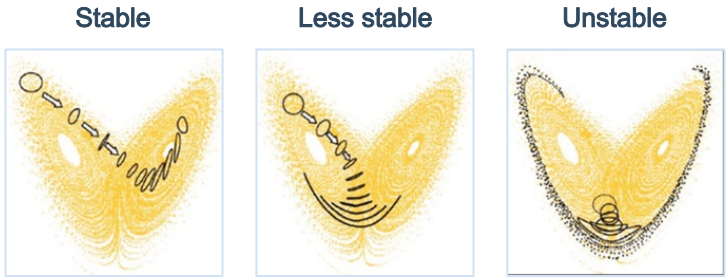


Image Source: Kalnay (2003)

$$\delta \mathbf{x}_{t_n} = \mathbf{M}_{[t_n,t_{n-1}]}\mathbf{M}_{[t_{n-1},t_{n-2}]} \cdots \mathbf{M}_{[t_2,t_1]}\mathbf{M}_{[t_1,t_0]}\delta \mathbf{x}_{t_0}$$

Predictability depends on the initial conditions (Palmer, 2002):



This Lorenz model is 3 Dimensional
Realistic models are upwards of $O(10^9)$

Error growth estimated by the linear propagator

$$\delta \mathbf{x}_{t_n} = \mathbf{M}_{[t_n,t_{n-1}]}\mathbf{M}_{[t_{n-1},t_{n-2}]} \cdots \mathbf{M}_{[t_2,t_1]}\mathbf{M}_{[t_1,t_0]}\delta \mathbf{x}_{t_0}$$

$$\delta \mathbf{x}_{t_n} = \mathbf{M}_{[t_n,t_0]}\delta \mathbf{x}_{t_0}$$

Lyapunov exponents are eigenvalues of:

$$\mathbf{\Lambda} = \lim_{t\to\infty} \frac{1}{t} \ln \left(\mathbf{M}_{[t,t_0]}\mathbf{M}^T_{[t,t_0]} \right)^{\frac{1}{2}}$$

Error growth estimated by the linear propagator

$$\delta \mathbf{x}_{t_n} = \mathbf{M}_{[t_n,t_{n-1}]}\mathbf{M}_{[t_{n-1},t_{n-2}]} \cdots \mathbf{M}_{[t_2,t_1]}\mathbf{M}_{[t_1,t_0]}\delta \mathbf{x}_{t_0}$$

$$\delta \mathbf{x}_{t_n} = \mathbf{M}_{[t_n,t_0]}\delta \mathbf{x}_{t_0}$$

Lyapunov exponents are eigenvalues of:

$$\mathbf{\Lambda} = \lim_{t\to\infty} \frac{1}{t} \ln \left(\mathbf{M}_{[t,t_0]}\mathbf{M}^{T}_{[t,t_0]} \right)^{\frac{1}{2}}$$

Error growth estimated by the linear propagator

$$\delta \mathbf{x}_{t_n} = \mathbf{M}_{[t_n,t_{n-1}]}\mathbf{M}_{[t_{n-1},t_{n-2}]} \cdots \mathbf{M}_{[t_2,t_1]}\mathbf{M}_{[t_1,t_0]}\delta \mathbf{x}_{t_0}$$

$$\delta \mathbf{x}_{t_n} = \mathbf{M}_{[t_n,t_0]}\delta \mathbf{x}_{t_0}$$

Lyapunov exponents are eigenvalues of:

$$\mathbf{\Lambda} = \lim_{t\to\infty} \frac{1}{t} \ln \left(\mathbf{M}_{[t,t_0]}\mathbf{M}^{T}_{[t,t_0]} \right)^{\frac{1}{2}}$$

“Lyapunov exponents are key tools for measuring chaos in dynamical systems. They quantify how fast nearby trajectories diverge or converge, revealing whether a system is stable, periodic, or chaotic.”

Error growth estimated by the linear propagator

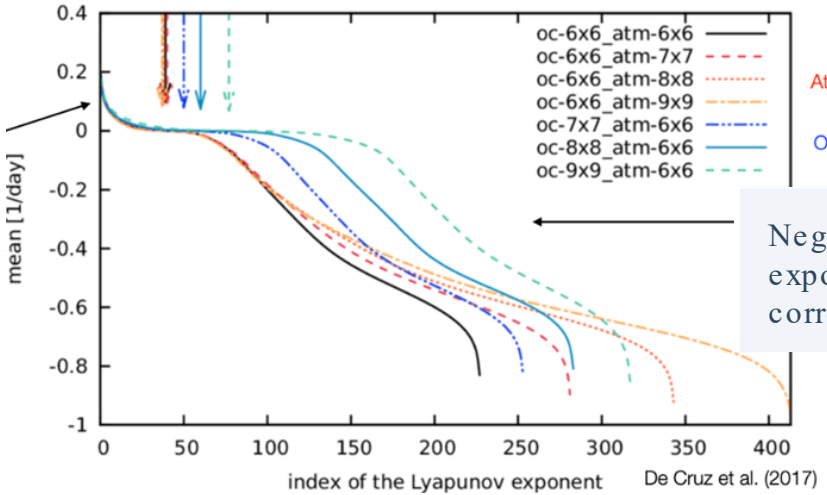
$$\delta \mathbf{x}_{t_n} = \mathbf{M}_{[t_n,t_{n-1}]}\mathbf{M}_{[t_{n-1},t_{n-2}]} \cdots \mathbf{M}_{[t_2,t_1]}\mathbf{M}_{[t_1,t_0]}\delta \mathbf{x}_{t_0}$$

$$\delta \mathbf{x}_{t_n} = \mathbf{M}_{[t_n,t_0]}\delta \mathbf{x}_{t_0}$$

Lyapunov exponents are eigenvalues of:

$$\mathbf{\Lambda} = \lim_{t\to\infty} \frac{1}{t} \ln \left(\mathbf{M}_{[t,t_0]}\mathbf{M}^T_{[t,t_0]} \right)^{\frac{1}{2}}$$

Lyapunov exponents for the dissipation model configurations



Positive exponents indicate exponential error growth in corresponding directions

Negative exponents indicate exponential error decay in corresponding directions

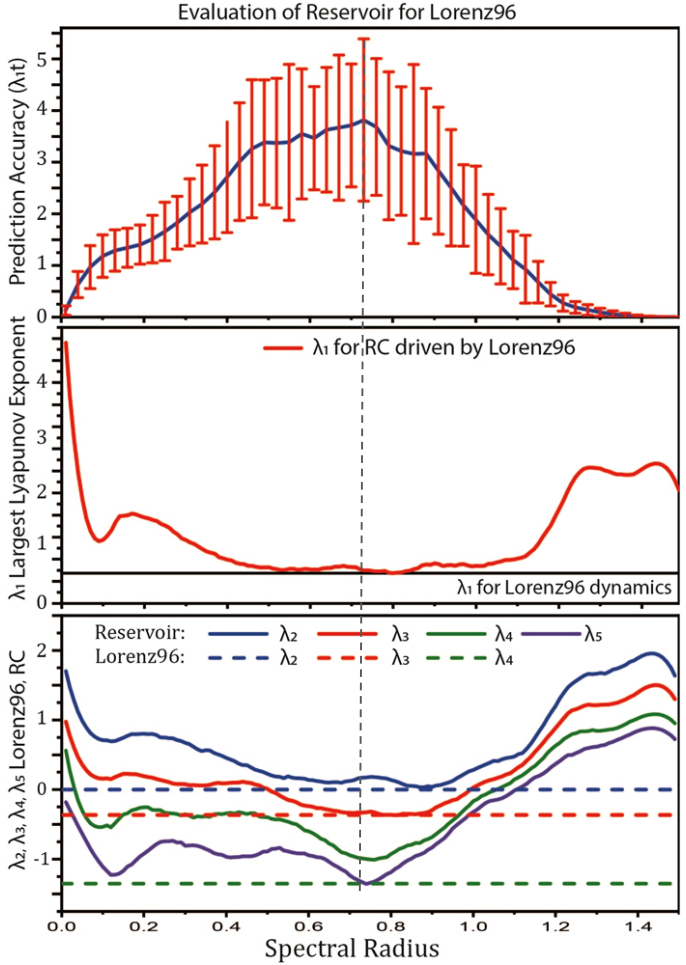
De Cruz et al. "Exploring the Lyapunov instability properties of high-dimensional atmospheric and climate models." https://npg.copernicus.org/articles/25/387/2018/

Getting the Lyapunov exponents/vectors correct leads to better forecasts

The fundamental feature of any model that makes it successful at forecasting chaotic systems **is the recovery of the Lyapunov spectrum** .

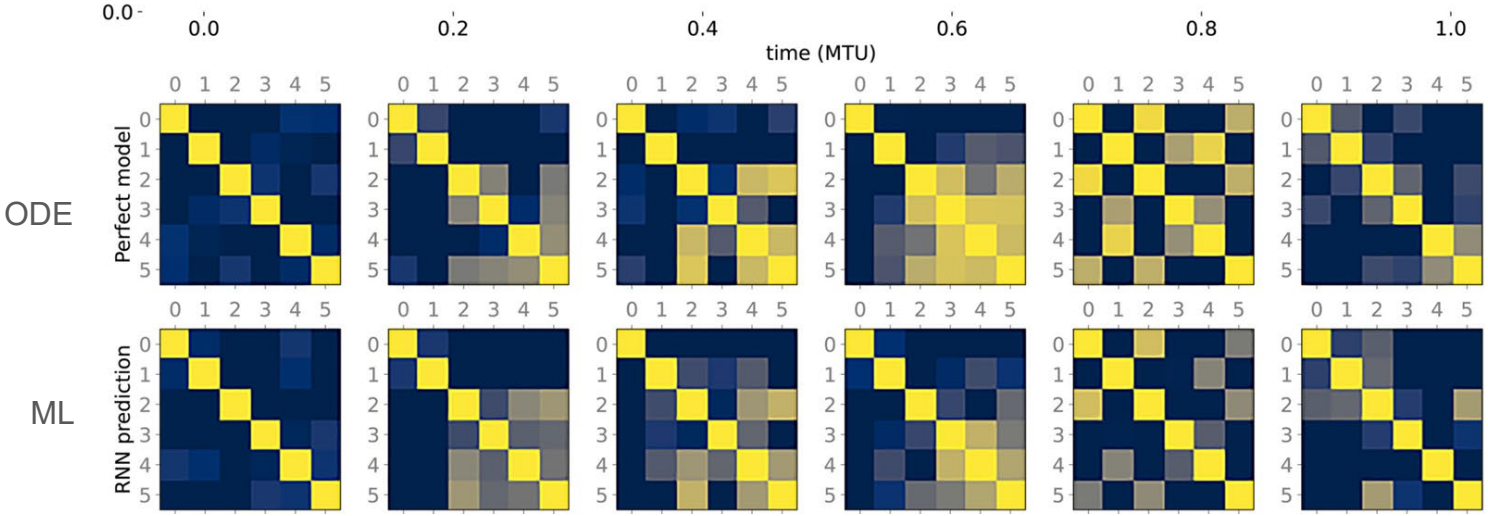
Why?

Otherwise, errors grow exponentially in the dimensions that are not resolved by the model.



Platt, **Penny**, Abarbanel, et al. (2021) "Robust forecasting using predictive generalized synchronization in reservoir computing" https://doi.org/10.1063/5.0066013

And that helps us get correct error covariance statistics

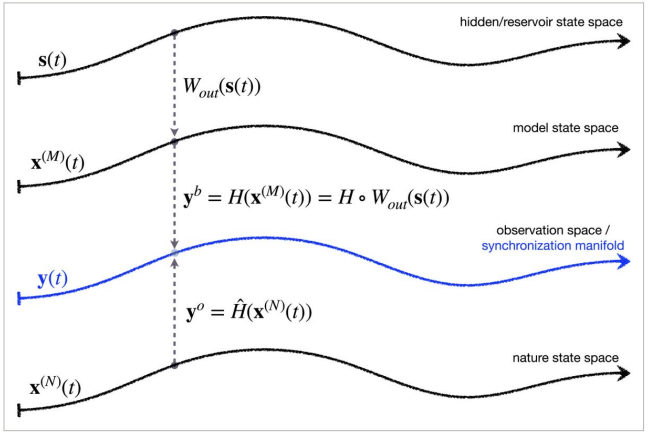


Example error correlation matrices throughout a sample forecast for the L96 numerical versus ML model

Penny et al. (2022) "Integrating Recurrent Neural Networks With Data Assimilation for Scalable Data-Driven State Estimation"
https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2021MS002843

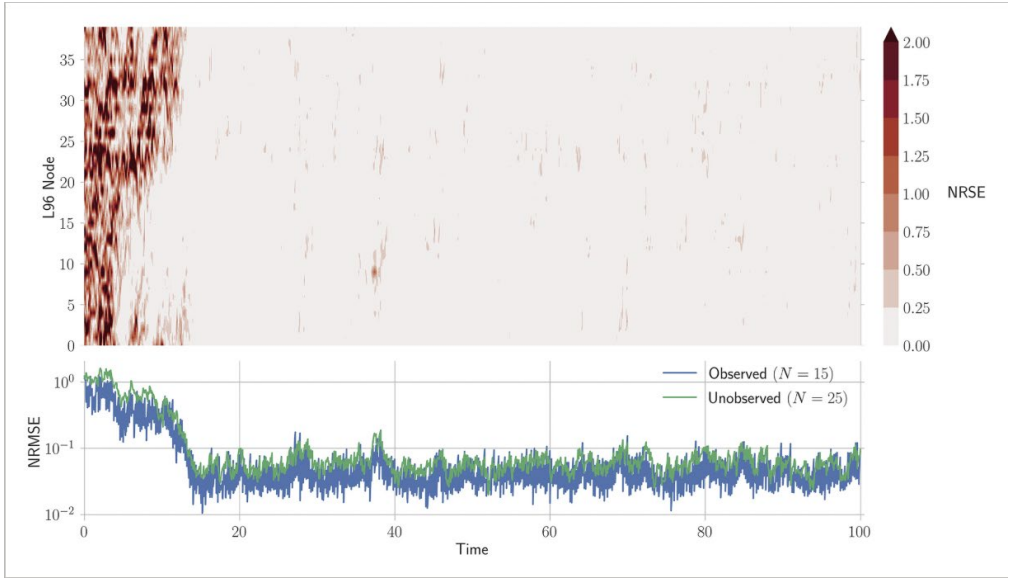
And we can validate that in sparsely-observed DA experiments

Augmented data assimilation:



So it **can** be done, the questions is just,
"can it be done for the atmosphere at scale"?

Important test for DA: assimilate sparse observations and then cycle the system…



Penny et al. (2022) "Integrating Recurrent Neural Networks With Data Assimilation for Scalable Data -Driven State Estimation"
https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2021MS002843

#6: The Analogue Trap (The Historian's Paradox)

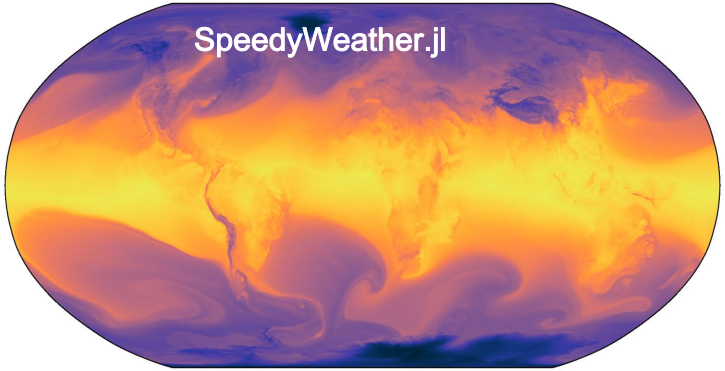
- **The Concept:** MLWP models are **Historians, not Physicists** . They have memorized 40+ years of ERA5 and "blend" past patterns to predict the future.

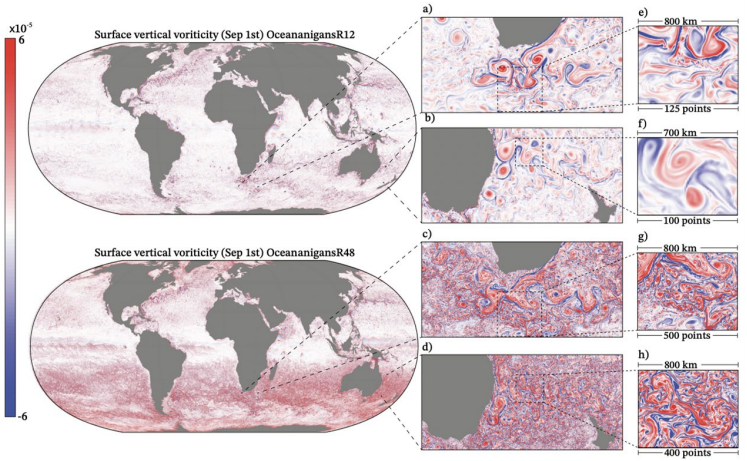
- **The Critique:** By acting as nonlinear analog forecasters, they inherit the same weaknesses as traditional ensembles: rank deficiency, spurious correlations, and states off of the atmosphere's attractor. They tell us what *usually* happens, but they are blind to why a specific instability is growing *now*.

Case Study: SFNO vs. NeuralGCM

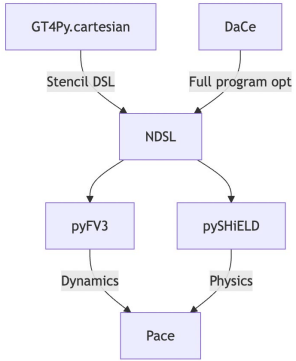
Feature	SFNO (NVIDIA / FCN) (Category 1)	NeuralGCM (Google) (Category 3)
Overall Philosophy	**The "Historian"** (Data-driven Mapper)	**The "Physicist"** (Physics-based Solver)
Shared Architecture	**SHT-MLP Sandwich:** Uses GPUs for dense SHT transforms and MLPs for local grids.	**SHT-MLP Sandwich:** Uses GPUs for dense SHT transforms and MLPs for local grids.
The Spectral Core	**Learned Filter:** Maps coefficients via learned statistical weights (Dense GEMM).	**PDE Solver:** Calculates primitive equations & fluid dynamics in spectral space.
Time-Step (CFL)	**Unconstrained:** Takes massive 6-hour leaps.	**Shackled:** Requires ~30-minute steps for numerical stability.
Inference Speed	**Ultra-Fast:** Single-pass execution.	**Moderate:** Pays the "tax" of iterative PDE bookkeeping (10x–50x slower).
Jacobian Fidelity	**Statistical:** Prone to "local spikes" and thermodynamically blind gradients.	**Physical:** Differentiable solver ensures honest, dynamically coupled gradients.
DA Readiness	**Low:** A fast emulator, but struggles with the "Rossby Adjoint Test" for T and q.	**High:** Provides the Tangent Linear/Adjoint operators of the dynamics, as needed for 4D-Var.

Category 3 models: enabled via Growing collection of modernized physics -based models:

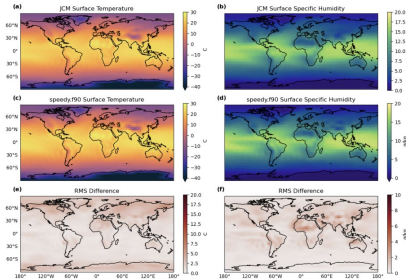




ClimaAtmos (CalTech) / ClimaOcean (MIT)



PACE - NOAA/GFDL



JCN - JAX Circulation Model

With accurate Jacobians: Employ DA using ML tools

Leveraging (A) automatic differentiation and (B) ML software tools to minimize the data assimilation cost function

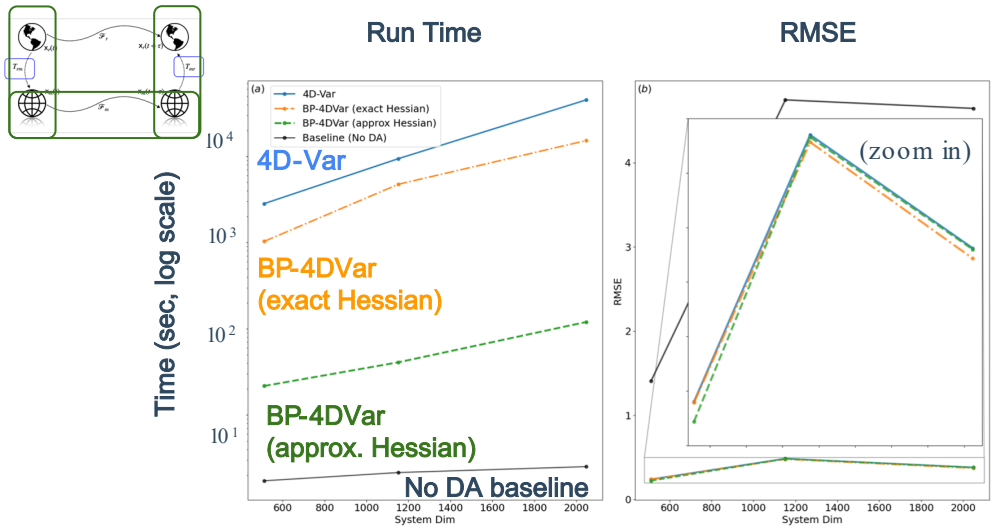
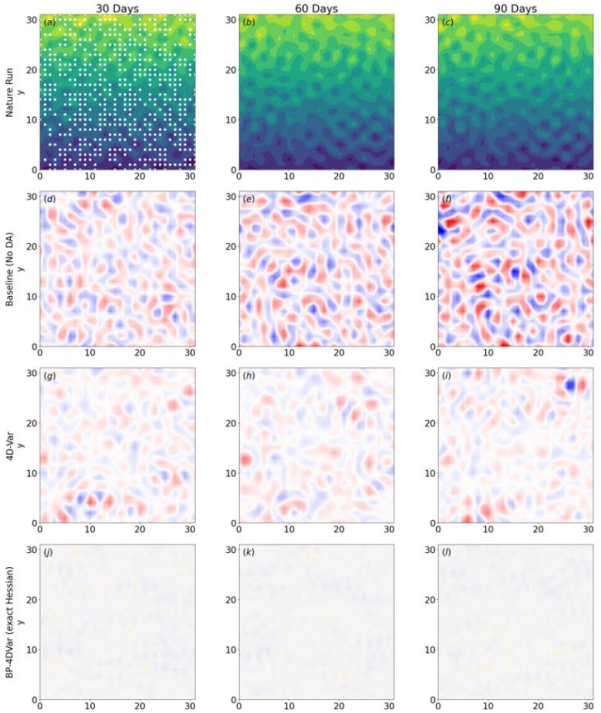


Figure 5. (a) Run times (log scale) and (b) RMSE for the QG dynamics using the PyQG-JAX forecast model, for 4D-Var and Backprop-4DVar. An unconstrained free run without data assimilation is provided as a baseline for comparison. While all three DA methods show similar performance in terms of RMSE, Backprop-4DVar using the approximate Hessian (green) is an order of magnitude faster than the reference methods.

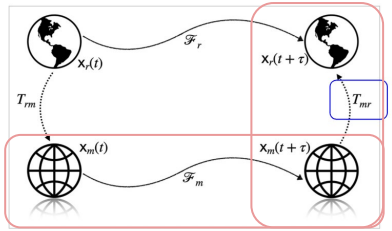


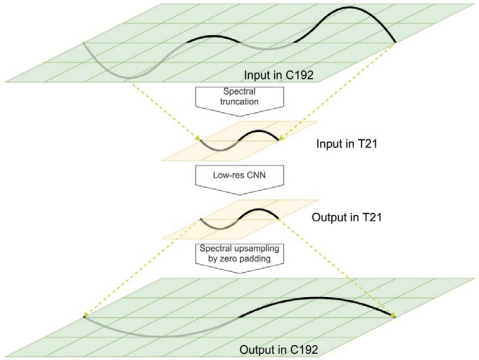
*Solvik, **Penny**, and Hoyer (2025)* "4D-Var Using Hessian Approximation and Backpropagation Applied to Automatically Differentiable Numerical and Machine Learning Models" https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004608

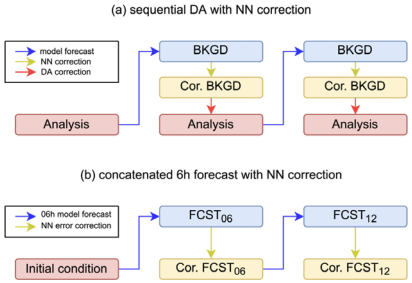
49

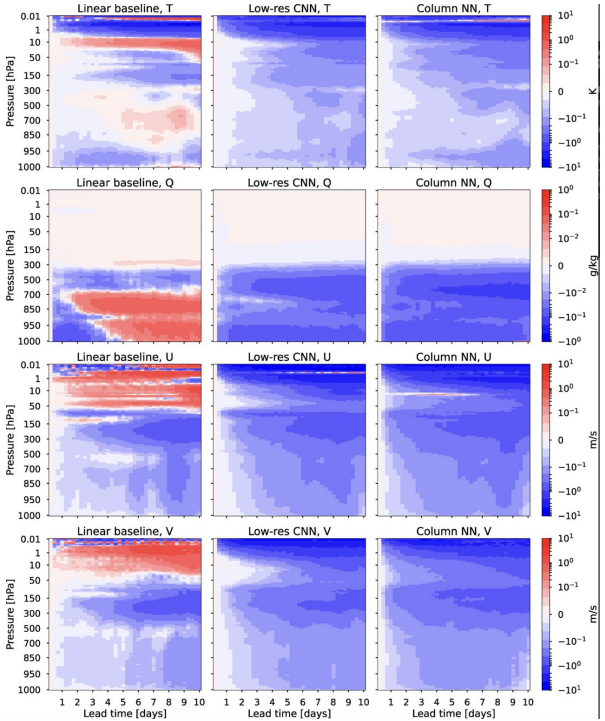
Category 3 models: AI/ML to enhance NWP

NWP with ML post - processing









Chen, **Penny**, Whitaker, et al. (2022) "Correcting Systematic and State-Dependent Errors in the NOAA FV3-GFS Using Neural Networks"
https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2022MS003309

Summary: the community has mastered visual emulation but is still struggling with the core capabilities needed for data assimilation

The k^{-3} Trap: Resolving large-scale deterministic flow - It is not a revolution; it's a baseline.

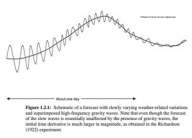
The $k^{-5/3}$ Trap: Recovering the energy is not sufficient - we need to resolve the chaos.

Jacobian Fidelity : The "Real Test" of an MLWP model for DA is its **Adjoint Sensitivity** - this is how we connect sparse observations to physical corrections - it's a mathematical constraint, not a technological one.

The "analogue trap" (The Historian vs. The Physicist) : A historian tells stories based on what happened before. A physicist predicts what *must* happen based on the laws of the universe.

A brief history of numerical weather prediction

1904: Vilhelm Bjerknes derived the primitive equations and proposed the idea of a forecast with a diagnostic step and prognostic step



1922: Lewis Fry Richardson attempted to use these differential equations to evolve the state of the atmosphere, computed by hand - and failed miserably

1950s: Jule Charney filtered out fast gravity waves with the quasi -geostrophic equations and with John Von Neumann pioneered the first numerical integration on the ENIAC computer

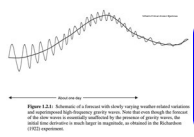
1959: Karl-Heinz Hinklemann produces a primitive equation forecast, ~40 years after Richardson's failed attempt

1960s: Ed Lorenz discovered "chaos" in simple weather models

A brief history of numerical weather prediction

MLWP models are about here

1904: Vilhelm Bjerknes derived the primitive equations and proposed the idea of a forecast with a diagnostic step and prognostic step



1922: Lewis Fry Richardson attempted to use these differential equations to evolve the state of the atmosphere, computed by hand - and failed miserably

1950s: Jule Charney filtered out fast gravity waves with the quasi -geostrophic equations and with John Von Neumann pioneered the first numerical integration on the ENIAC computer

1959: Karl-Heinz Hinklemann produces a primitive equation forecast, ~40 years after Richardson's failed attempt

1960s: Ed Lorenz discovered "chaos" in simple weather models

Conclusion

“The forecast is based on the supposition that what the atmosphere did then, it will do again now. …The past history of the atmosphere is used, so to speak, as a full -scale working model of its present self”

– Lewis Fry Richardson,
lamenting the use of data -driven ‘analog’ forecasting in the early 1900’s

A brief history of using observations in forecasting

Objective Analysis

Objective analysis schemes:
Panofsky (1949), Gilchrist and Cressman (1954), Cresman (1959), Barnes (1964, 1978)

Newtonian relaxation / Nudging schemes :
Hoke and Anthes (1976), Kistler (1974)

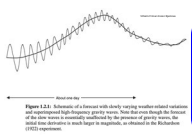
Data Assimilation

- Incorporate a 'first guess' or 'background field
  - Using climatology - Gandin (1963), Bergthorsson and Doos (1955)
  - Using short-range forecasts
- Static multivariate statistical DA
  - Optimal Interpolation, 3D-Var
- Modern flow-resolving DA
  - Dynamic:
    - 4D-Var (Courtier and Talagrand, 1990),
    - Ensemble Kalman Filter, and hybrids

A brief history of numerical weather prediction

1904: Vilhelm Bjerknes derived the primitive equations and proposed the idea of a forecast with a diagnostic step and prognostic step





1922: Lewis Fry Richardson attempted to use these differential equations to evolve the state of the atmosphere, computed by hand - and failed miserably




1950s: Jule Charney filtered out fast gravity waves with the quasi -geostrophic equations and with John Von Neumann pioneered the first numerical integration on the ENIAC computer

1959: Karl-Heinz Hinklemann produces a primitive equation forecast, ~40 years after Richardson's failed attempt

1960s: Ed Lorenz discovered "chaos" in simple weather models

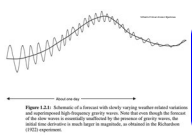
MLWP models are about here

ML-DA is about here

A brief history of numerical weather prediction

1904: Vilhelm Bjerknes derived the primitive equations and proposed the idea of a forecast with a diagnostic step and prognostic step





1922: Lewis Fry Richardson attempted to use these differential equations to evolve the state of the atmosphere, computed by hand - and failed miserably




1950s: Jule Charney filtered out fast gravity waves with the quasi -geostrophic equations and with John Von Neumann pioneered the first numerical integration on the ENIAC computer

1959: Karl-Heinz Hinklemann produces a primitive equation forecast, ~40 years after Richardson's failed attempt

1960s: Ed Lorenz discovered "cha[illegible]" in simple weather models

MLWP models are about here

ML-DA is about here

This hasn't been discovered yet.

Conclusion

“The forecast is based on the supposition that what the atmosphere did then, it will do again now. …The past history of the atmosphere is used, so to speak, as a full -scale working model of its present self”

– Lewis Fry Richardson,
lamenting the use of data -driven ‘analog’ forecasting in the early 1900’s

100 years later – should we come full circle?


steve.penny@sofarocean.com