



# Data Assimilation in a Time of Artificial Intelligence

"Overview of problems and recent progress"

Stephen G. Penny - Head of Weather, Sofar Ocean

\* Research affiliate - CIRES at the University of Colorado Boulder

\* Visiting research professor - University of Maryland College Park

Sofar Weather Team:

Isabel Houghton, Miguel Solano, Wolfgang Langhans, Colin Grudzien

Moriah Cesaretti, Ciara Dorsay

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)



2025 US CLIVAR Summit

22 July 2025

# Disclaimers

- The focus of this presentation is on the potential for Machine Learning Weather Prediction (MLWP) to **fully replace** operational Numerical Weather Prediction (NWP).
- This presentation discusses AI/ML models that are **available *right now***. It does not predict the future, but will describe some fundamental challenges that must be addressed.
- There are many different ways that AI/ML can be incorporated into the NWP forecast/analysis cycle that serve to replace key components of this process. I discuss this in more detail in my talk at the ECMWF:  
Workshop report: “2022 ECMWF-ESA workshop report: current status, progress and opportunities in machine learning for Earth System observation and prediction”  
<https://www.nature.com/articles/s41612-023-00387-2>  
ECMWF-ESA Machine Learning Workshop keynote presentation:  
<https://vimeo.com/770758490/bac45588aa>



# Applying DA with AI/ML

Olivier Talagrand categorized methods based on statistical estimation theory:

“In the late sixties, the development of satellite observing systems, and the perspective that synoptic observations, performed more or less continuously in time, would become more and more numerous in the future, led to the notion that **the dynamical evolution of the flow should be explicitly taken into account in the very definition of the initial conditions of the forecast**. The word *assimilation* was coined at that time for denoting a process in which observations distributed in time are merged together with a dynamical numerical model of the flow in order to determine as accurately as possible the state of the atmosphere” (*Talagrand, 1997*)



# From Observations to Trajectory Estimates

Mapping directly from observations to the trajectory of a dynamical system is in practice an “**ill-posed problem**”

i.e. it is a mathematical problem where one or more of the conditions for well-posed problems are not met:

- existence of a solution,
- uniqueness of the solution, and
- stability (continuous dependence of the solution on the input data).

An ill-posed problem might not have a solution, might have multiple solutions, or its solution might be highly sensitive to small changes in the input. (Hadamard 1923)



# From Observations to Trajectory Estimates

Mapping directly from observations to the trajectory of a dynamical system is in practice an “**ill-posed problem**”

i.e. it is a mathematical problem where one or more of the conditions for well-posed problems are not met:

- existence of a solution,
- **uniqueness of the solution, and**
- **stability** (continuous dependence of the solution on the input data).

An ill-posed problem might not have a solution, **might have multiple solutions, or its solution might be highly sensitive** to small changes in the input. (Hadamard 1923)

Therefore in DA, “all the available information is used in order to estimate as accurately as possible the state” *Talagrand (1997)*



# Key properties of Data Assimilation

- **It is cycled** - DA itself is a dynamical system, and any forecast system must be synchronized with nature via information provided from the observations in order to trust the resulting forecast
- Due to the sparsity\* of observations, it is necessary to have **a complete “first guess” or background/prior** of the atmospheric state that carries information from the past to the present (*Gilchrist and Cressman 1954; Bergthorsson and Doos 1955*), from observed to unobserved variables, and from upstream observed areas to downstream unobserved areas
- Carefully **estimates and accounts for errors** in observations **and** the trajectory of the dynamics to determine the most likely state and its uncertainty - typically relying on Bayesian inference, which means the result is a probability distribution



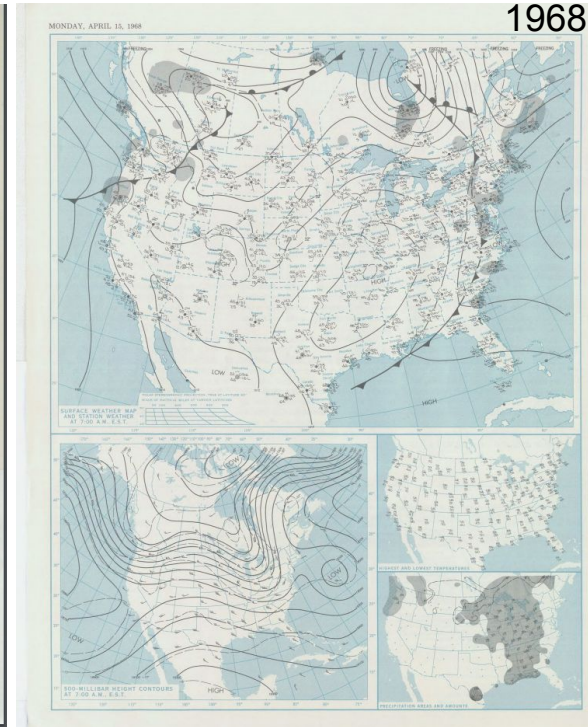
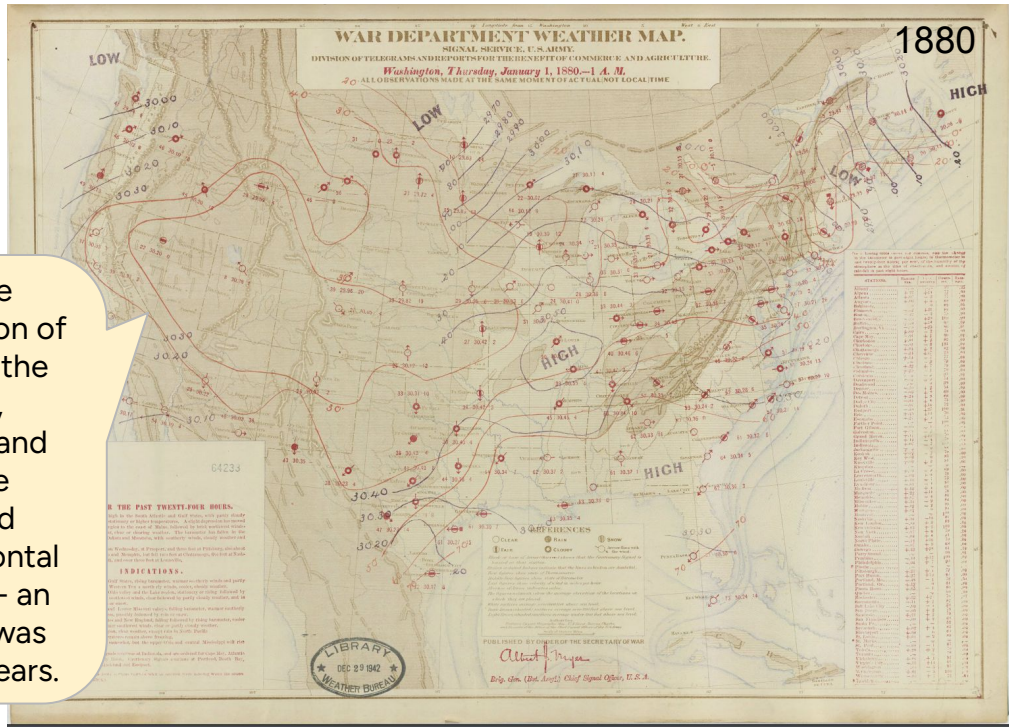
\*relative to the total number of possible observations for all essential physical variables at all points in the 3D volume of the global atmosphere - modern observing systems are still sparse relative to resolved scales in operational forecast models.



# A history of using observations in forecasting

The telegraph, invented in 1837 and first used to send meteorological data in 1844, made the construction of near real-time (NRT) forecasts possible.

Permitted the hand-construction of an "analysis" of the atmosphere, including High and Low pressure systems, wind patterns, and frontal storm systems - an approach that was used for 100+ years.



# A history of using observations in forecasting

## Methods

### **Objective analysis schemes:**

Panofsky (1949),  
Gilchrist and Cressman (1954), 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
- Multivariate statistical DA
  - Static: Objective analysis, 3D-Var
  - Dynamic (i.e. leveraging flow dynamics):
    - 4D-Var (Courtier and Talagrand, 1990),
    - Ensemble Kalman Filter, and hybrids





# A history of using observations in forecasting

## Methods

### **Objective analysis schemes:**

Panofsky (1949),  
Gilchrist and Cressman (1954), Barnes (1964, 1978)

### **Newtonian relaxation / Nudging schemes:**

Hoke and Anthes (1976), Kistler (1974)

## **Data Assimilation**

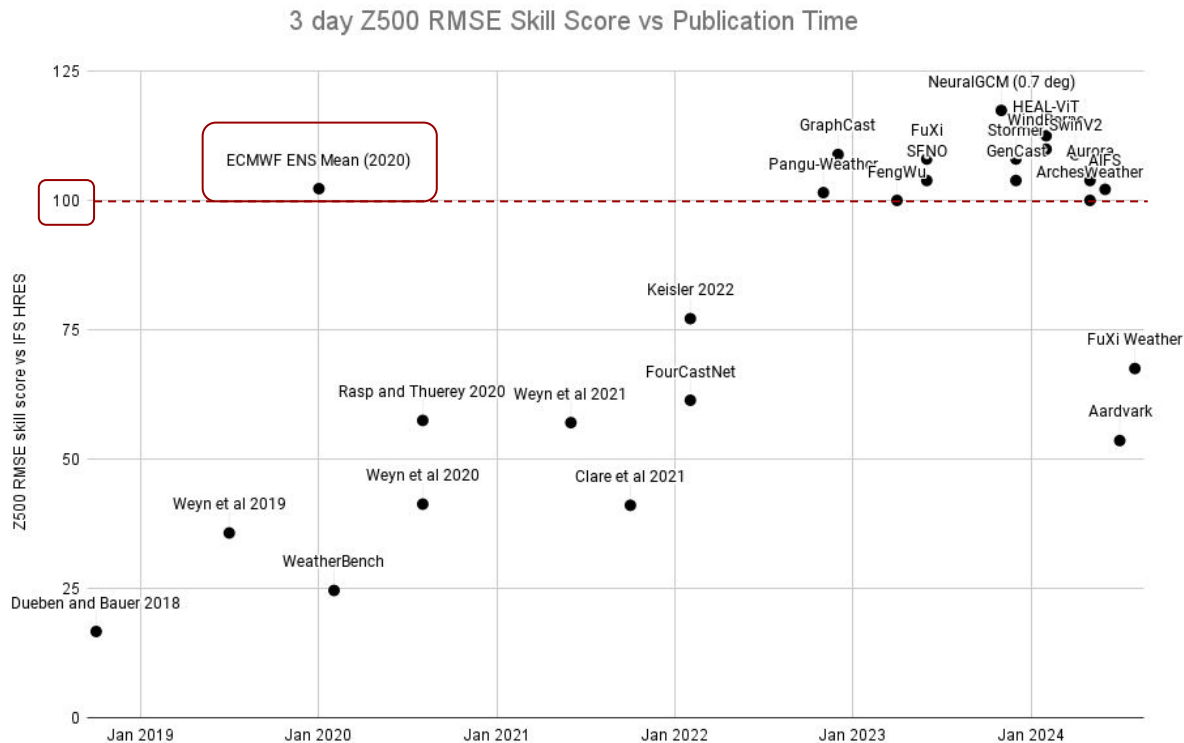
- Incorporate a ‘first guess’ or ‘background field’
  - Using climatology - Gandin (1963), Bergthorsson and Doos (1955)
  - Using short-range forecasts
- Multivariate statistical DA
  - Static: Objective analysis, 3D-Var
  - Dynamic (i.e. leveraging flow dynamics):
    - 4D-Var (Courtier and Talagrand, 1990),
    - Ensemble Kalman Filter, and hybrids

This is where most current AI/ML-based “data assimilation” approaches are categorized.



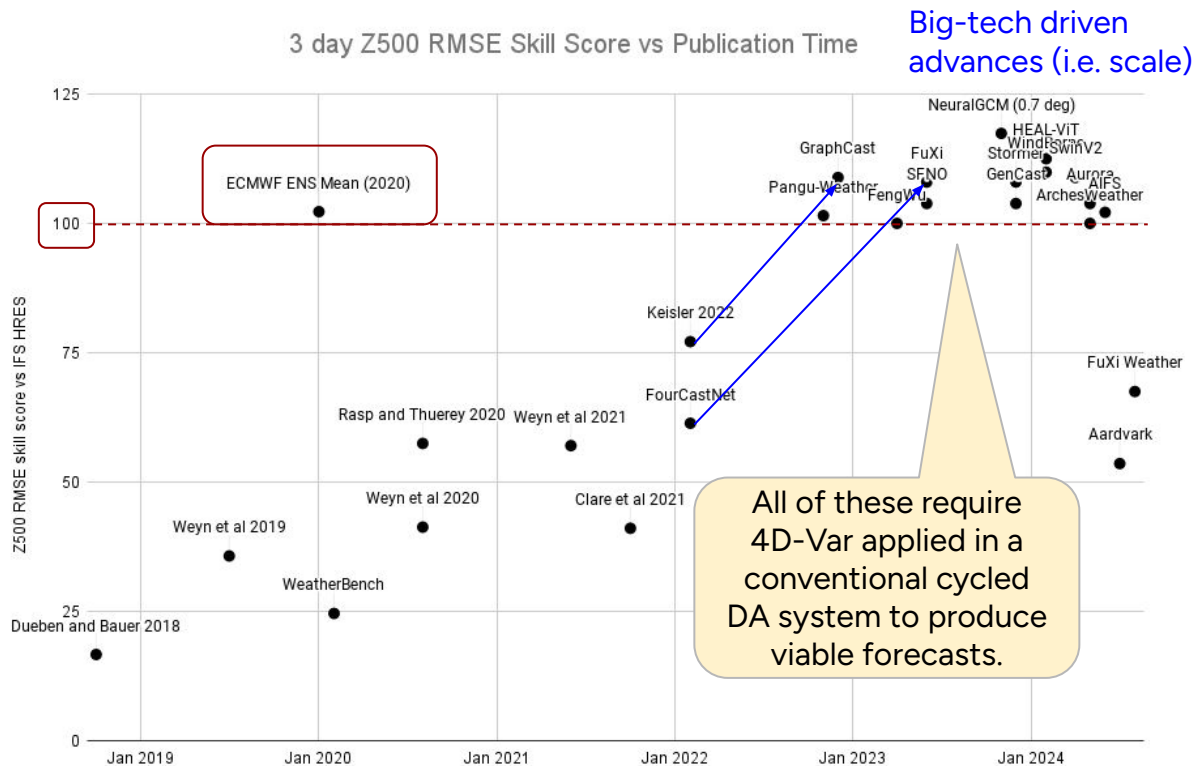
# Machine Learning Weather Prediction (MLWP)

- In the last 5 years, MLWP models have advanced rapidly
- In the last 2 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”



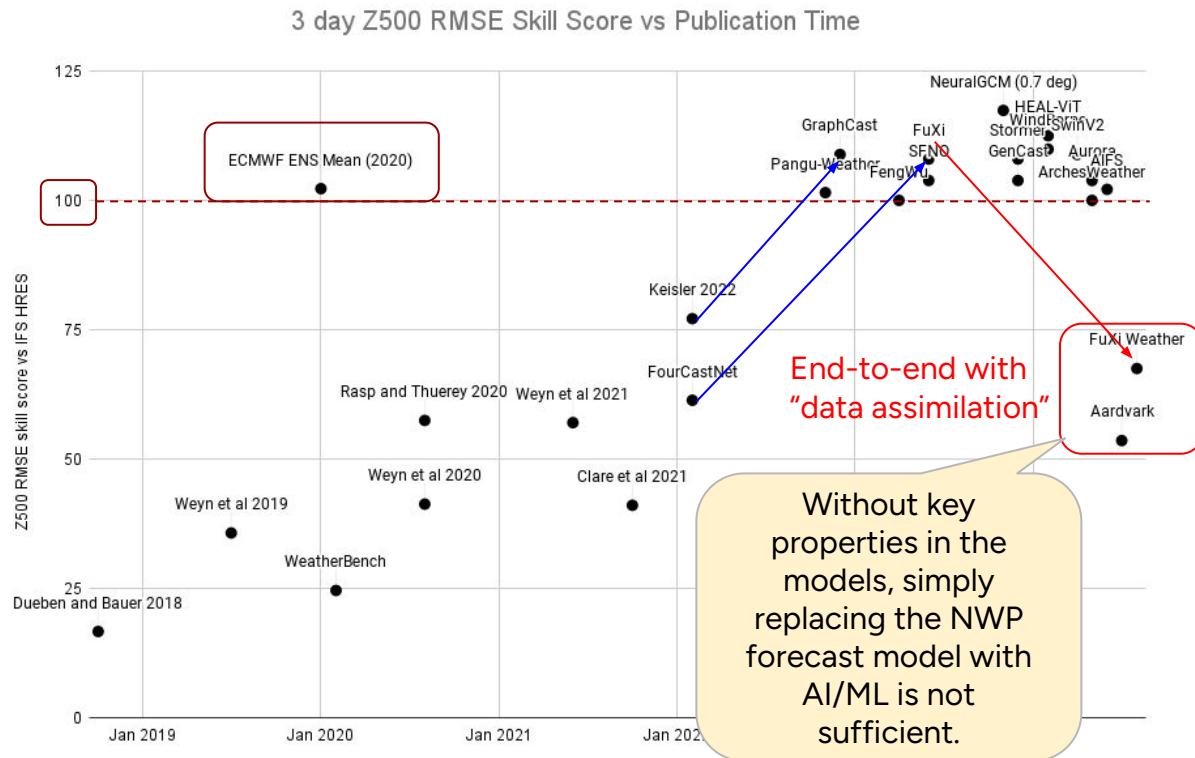
# Machine Learning Weather Prediction (MLWP)

- In the last 5 years, MLWP models have advanced rapidly
- In the last 2 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”



# Machine Learning Weather Prediction (MLWP)

- In the last 5 years, MLWP models have advanced rapidly
- In the last 2 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”



# Some principal challenges in ML Weather Prediction (MLWP) today

1. The “Blurriness” challenge (no, this not a feature)
  - a. Models are trained to score on forecaster metrics (not modeling metrics)
  - b. Models do not reproduce spectral characteristics (this is not physical)
2. The sensitivity challenge: producing accurate Jacobian and ensemble forecast correlation/covariance statistics
3. The Discerned Learning challenge: AI/ML models learn “everything” not just model physics
4. The Evaluation challenge: how to measure whether a forecast is skillful?





# Some principal challenges in ML Weather Prediction (MLWP) today

1. The “Blurriness” challenge (no, this not a feature)
  - a. Models are trained to score on forecaster metrics (not modeling metrics)
  - b. Models do not reproduce spectral characteristics (this is not physical)
2. **The sensitivity challenge:** producing accurate Jacobian and ensemble forecast correlation/covariance statistics
3. **The Discerned Learning challenge:** AI/ML models learn “everything” not just model physics
4. The Evaluation challenge: how to measure whether a forecast is skillful?



# The *sensitivity* challenge

There are multiple sources of uncertainty in forecasts. Ensemble forecasts attempt to estimate these:

- **Initial conditions** (chaotic dynamics)
- Background distribution accuracy (Bayesian prior)
- Observation errors (aleatoric uncertainty)
- Systematic errors in the model (epistemic uncertainty)



# AI/ML Weather models resolve large scales

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>

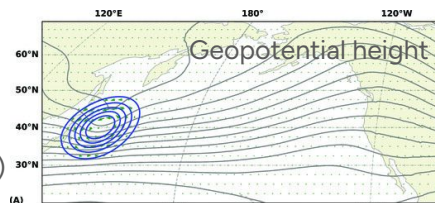
Solution at 500 hPa for a localized disturbance on the DJF atmosphere. The "time evolution of a localized 500-hPa trough at the western end of the North Pacific storm track, which is the canonical initial condition preceding surface cyclogenesis."

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

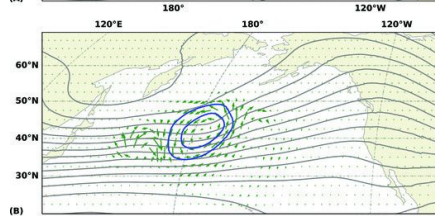
Left - Solution at 500 hPa for a localized disturbance on the DJF-averaged atmosphere state using PanguWeather. Geopotential height is shown by gray lines, every 60 m.

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

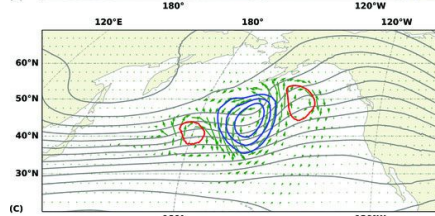
0 days  
(initial condition)



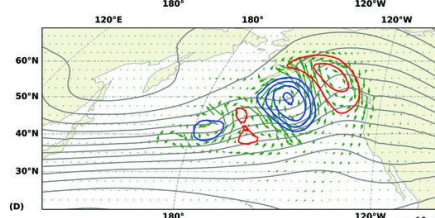
2 days



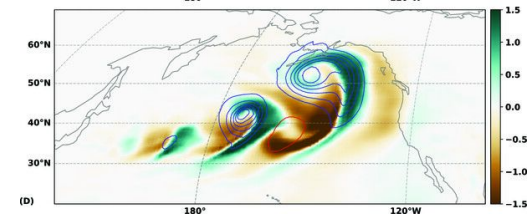
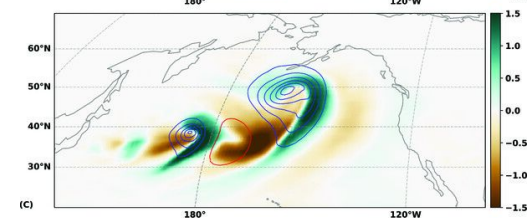
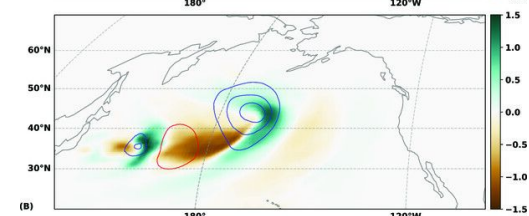
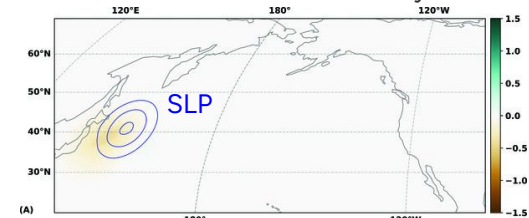
3 days



4 days



specific humidity  
anomaly at 850 hPa



# The ensemble sensitivity challenge

Tian, Holdaway, Kleist (2024) "Exploring the use of Machine Learning Weather models in Data Assimilation"

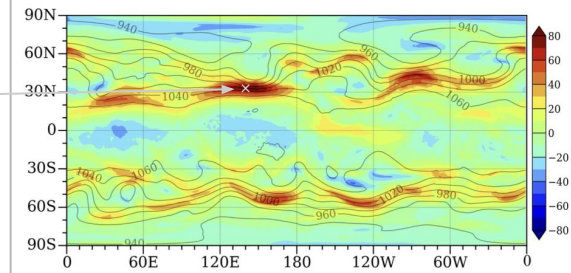
<https://arxiv.org/pdf/2411.14677>

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

Initial perturbation applied here



Graphcast

MPAS-A (reference truth)

Horizontal

Vertical

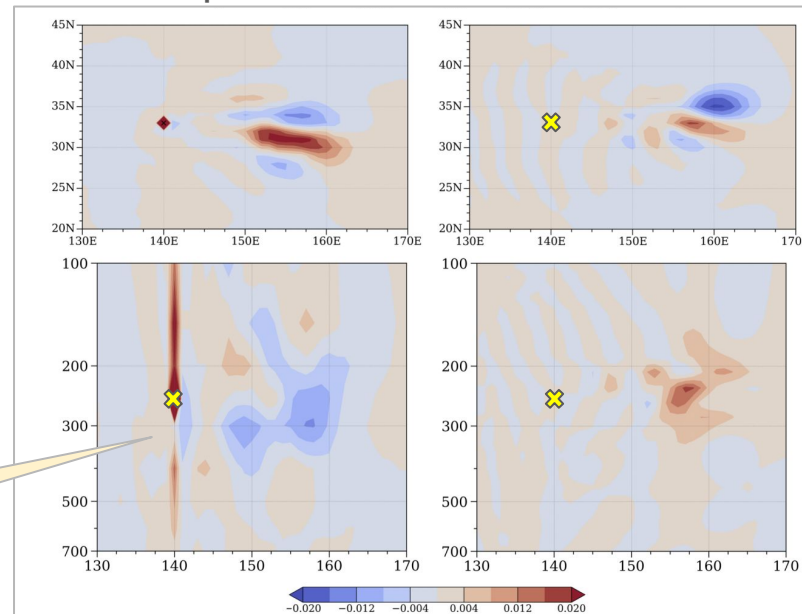


Figure 2: (a) Background geopotential heights (contoured) and zonal wind (shaded) at 00 UTC on January 1, 2022. The cross marks the location of the imposed perturbation. (b)-(c) Horizontal distribution of the TL response in zonal wind 6 hours into the forecast to a zonal wind perturbation at the initial 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).

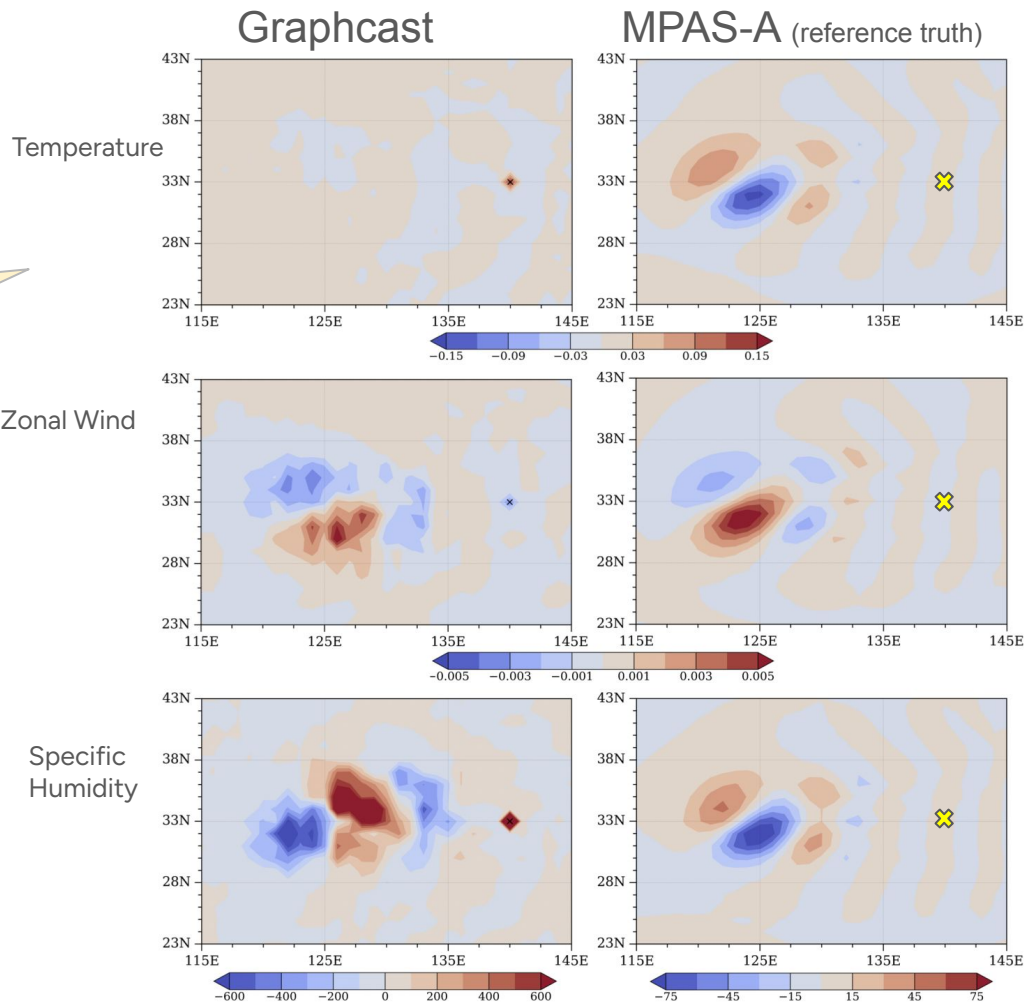


# The ensemble sensitivity challenge

Tian, Holdaway, Kleist (2024) "Exploring the use of Machine Learning Weather models in Data Assimilation"

<https://arxiv.org/pdf/2411.14677>

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





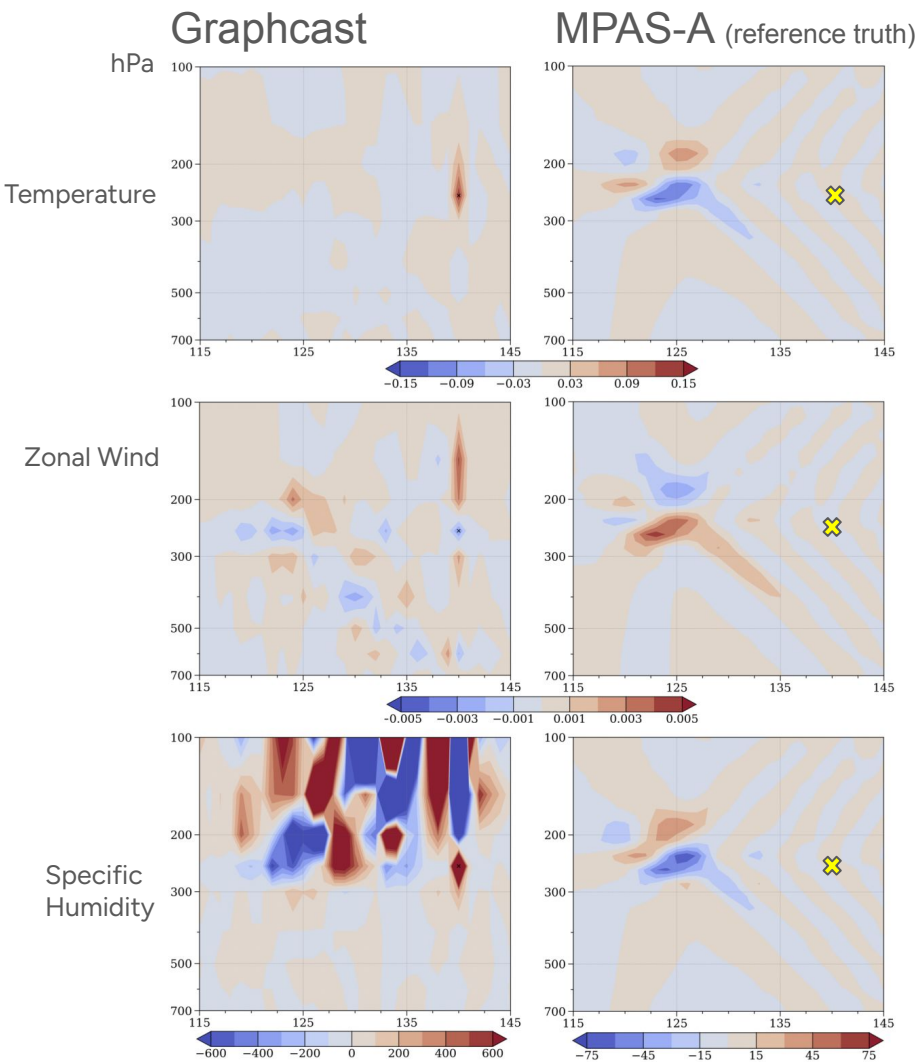
# The ensemble sensitivity challenge

Tian, Holdaway, Kleist (2024) "Exploring the use of Machine Learning Weather models in Data Assimilation"

<https://arxiv.org/pdf/2411.14677>

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.



# The ensemble sensitivity challenge

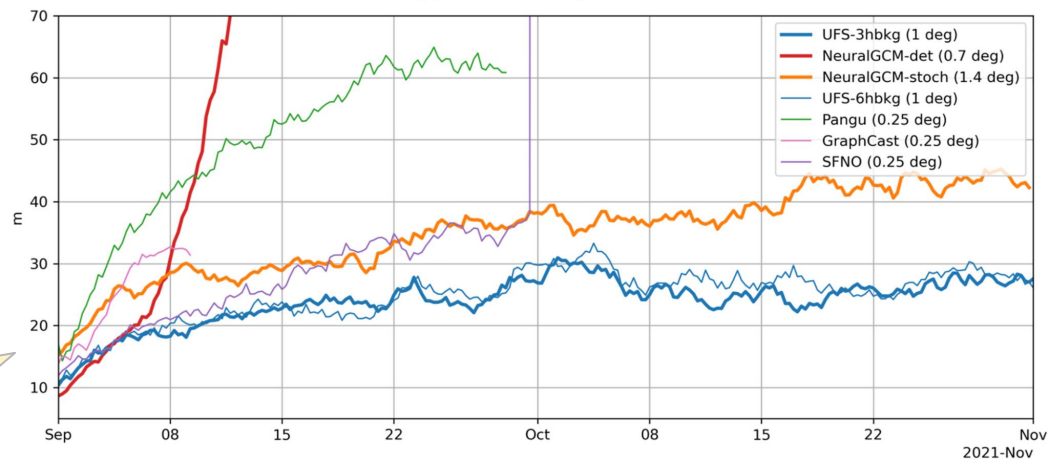
Slivinski et al. (2025) "Assimilating Observed Surface Pressure Into ML Weather Prediction Models"

<https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2024GL114396>

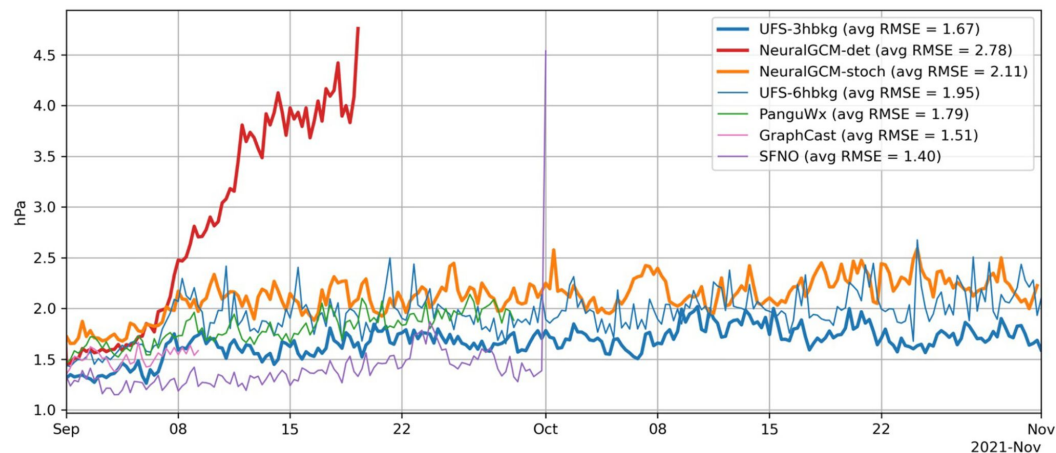
All of the MLWP models underperform the 1-degree NOAA Unified Forecast System (UFS) NWP model when applied within an 80-member ensemble Kalman filter.

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

(a) NH Z500 analysis RMSE



(b) Global background RMS fit to observations



# 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 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.”



# We “fight chaos” by understanding error growth

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

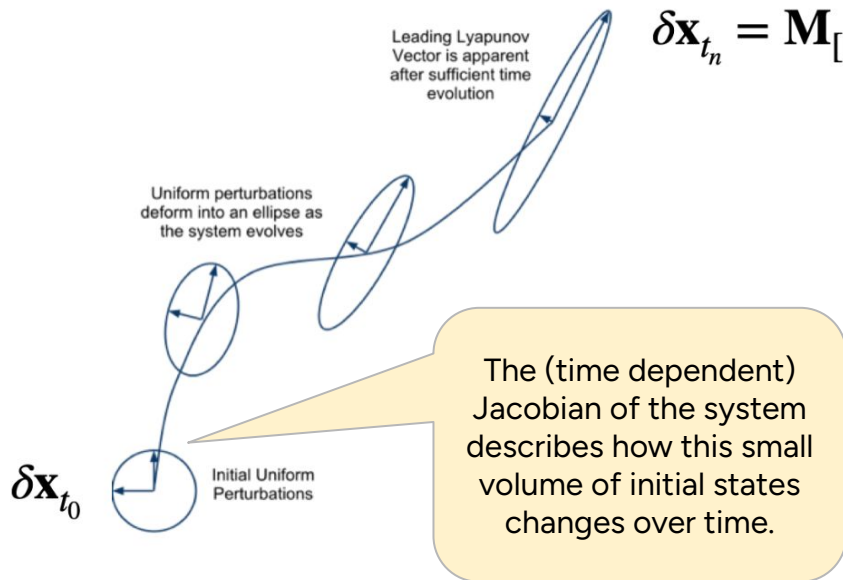
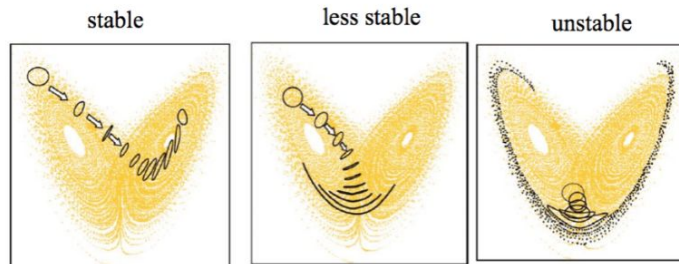


Image Source: Kalnay (2003)

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

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

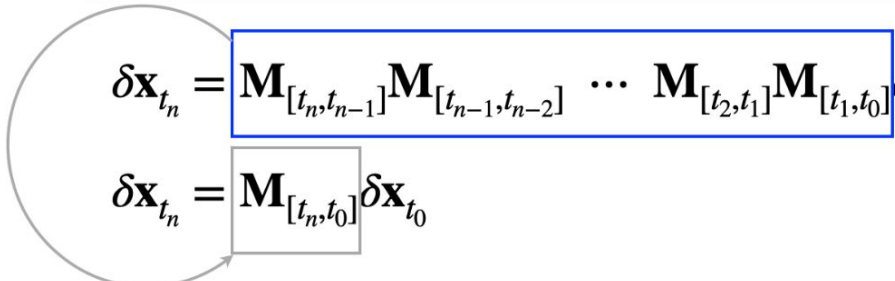


This Lorenz model is  
3-Dimensional

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:

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





# 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:

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



# 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:

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

"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

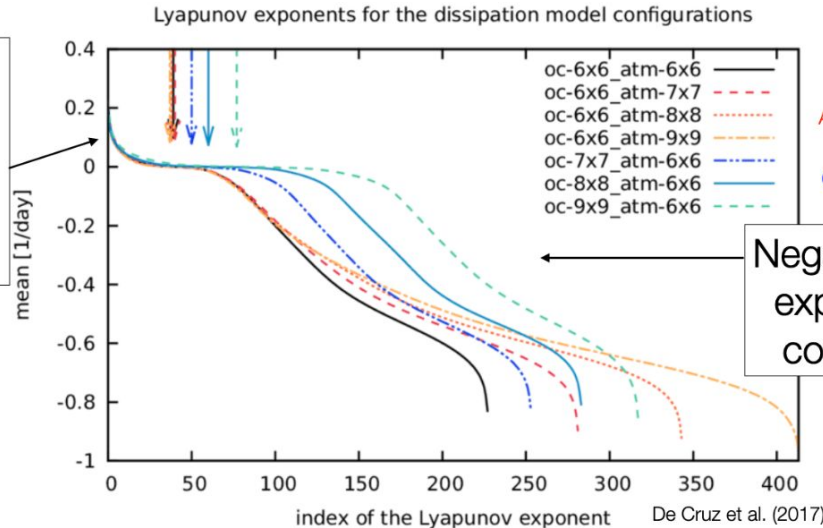
$$\delta \mathbf{x}_{t_n} = \mathbf{M}_{[t_n, t_{n-1}]} \mathbf{M}_{[t_{n-1}, t_{n-2}]} \cdots \mathbf{M}_{[t_2, t_1]} \mathbf{M}_{[t_1, t_0]} \delta \mathbf{x}_{t_0}$$

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

Lyapunov exponents are eigenvalues of:

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

Positive exponents indicate exponential error growth in corresponding directions



Negative exponents indicate exponential error decay in corresponding directions



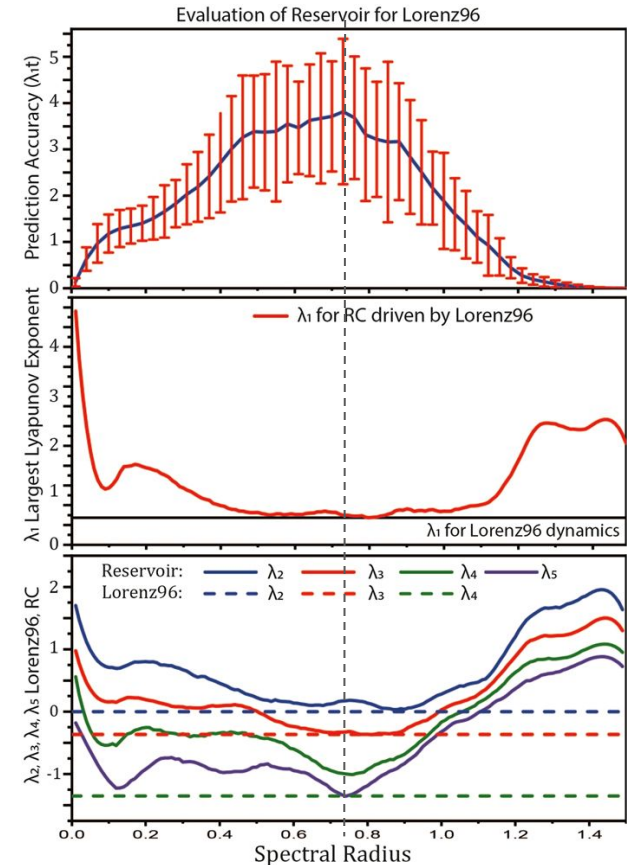
# Why does this matter?

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

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

Why?

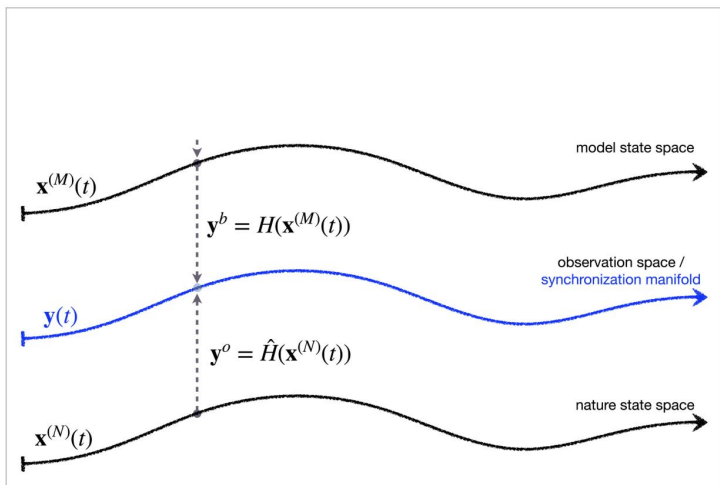
Otherwise, errors grow exponentially in the dimensions that are not resolved by the model.



# Are large ML ensembles useful?

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>

## Typical NWP data assimilation:



This is the typical process of data assimilation for NWP

Figure 1

[Open in figure viewer](#) | [PowerPoint](#)

Data assimilation is applied to update the hidden/reservoir state space  $\mathbf{s}(t)$ . Observations  $\mathbf{y}^o$  are sampled from the nature state space  $\mathbf{x}^{(N)}(t)$ , while the composition  $H \circ W_{out}$  is used as an observation operator to map the hidden/reservoir state space to an equivalent representation that can be used to form the innovations  $\mathbf{d}(t) = \mathbf{y}^o(t) - H \circ W_{out}(\mathbf{s}(t))$  in the observation space.

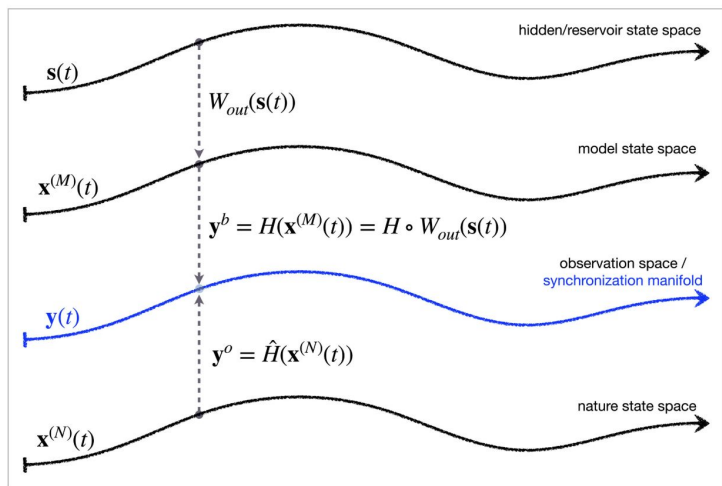




# Are large ML ensembles useful?

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>

## Augmented data assimilation:

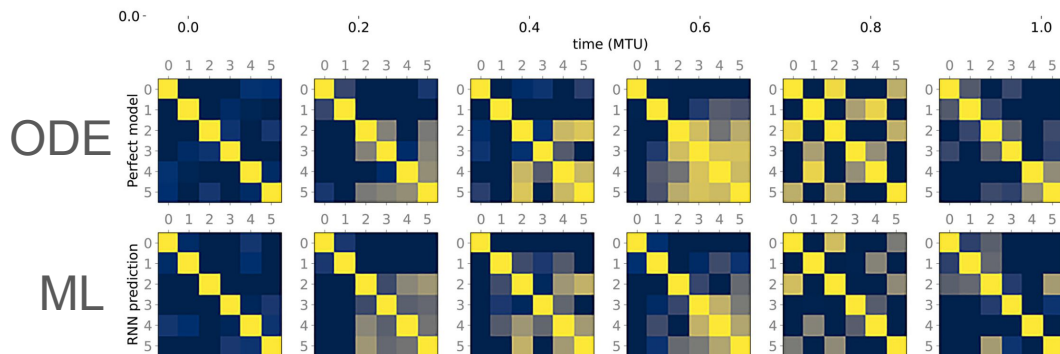


Here we add an additional abstraction layer and compute in the hidden/reservoir/latent space.

Figure 1

[Open in figure viewer](#) | [PowerPoint](#)

Data assimilation is applied to update the hidden/reservoir state space  $\mathbf{s}(t)$ . Observations  $\mathbf{y}^o$  are sampled from the nature state space  $\mathbf{x}^{(N)}(t)$ , while the composition  $H \circ W_{out}$  is used as an observation operator to map the hidden/reservoir state space to an equivalent representation that can be used to form the innovations  $\mathbf{d}(t) = \mathbf{y}^o(t) - H \circ W_{out}(\mathbf{s}(t))$  in the observation space.



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



# Are large ML ensembles useful?

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>

## Augmented data assimilation:

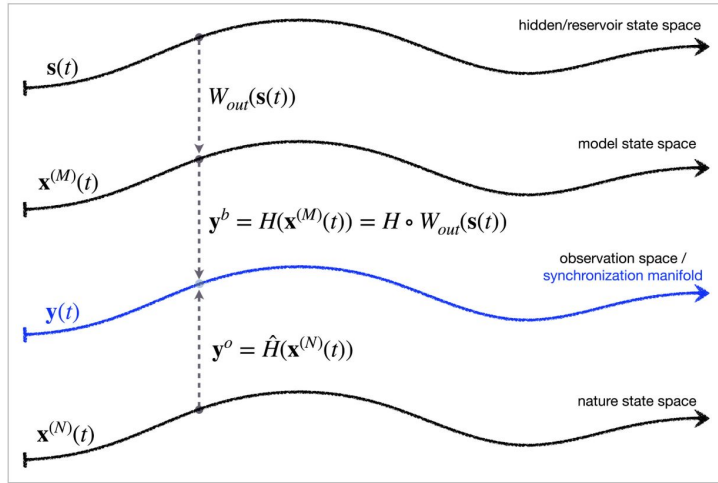
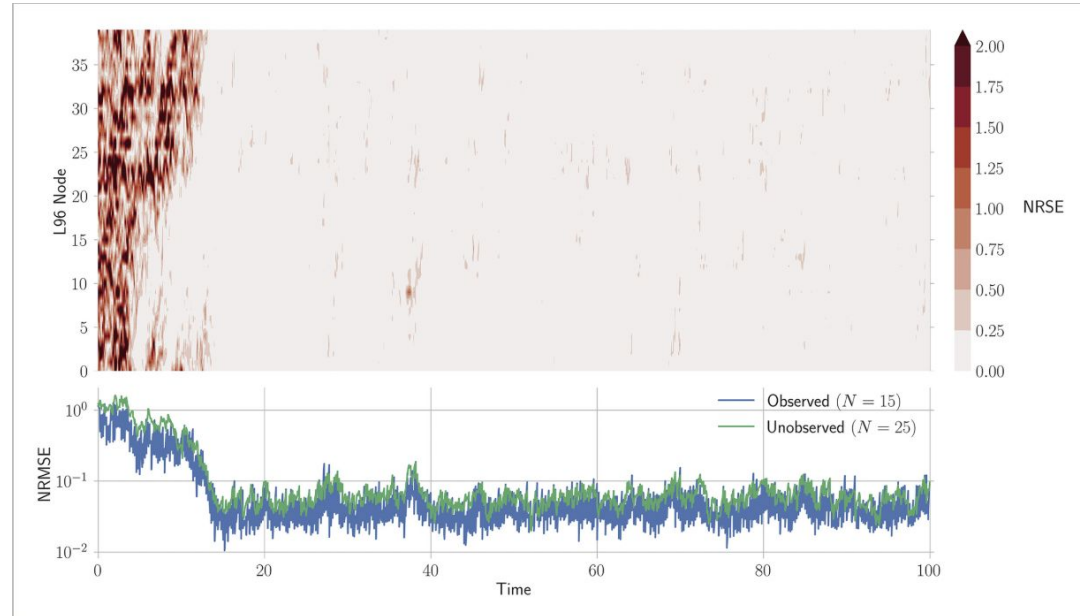


Figure 1

[Open in figure viewer](#) | [PowerPoint](#)

Data assimilation is applied to update the hidden/reservoir state space  $\mathbf{s}(t)$ . Observations  $\mathbf{y}^o$  are sampled from the nature state space  $\mathbf{x}^{(N)}(t)$ , while the composition  $H \circ W_{out}$  is used as an observation operator to map the hidden/reservoir state space to an equivalent representation that can be used to form the innovations  $\mathbf{d}(t) = \mathbf{y}^o(t) - H \circ W_{out}(\mathbf{s}(t))$  in the observation space.

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



# Summary: The *sensitivity* challenge

Why is the system Jacobian so important?

- 1) The system Jacobian produces the tangent linear model, adjoint, and describes ensemble response to perturbations in initial conditions at a given point in state space. This is critical to get the correct ensemble spread and statistics.
- 2) The Lyapunov vectors and Lyapunov exponents are produced by integrating the Jacobian over time - getting the Lyapunov spectrum correct is a requirement for producing an accurate forecast model of any chaotic dynamical system (like the weather). This is foundational.

A key test for the ability of MLWP models to reliably produce usable statistics is thus to validate that they produce the correct state-dependent Jacobians.

Today that is not the case.

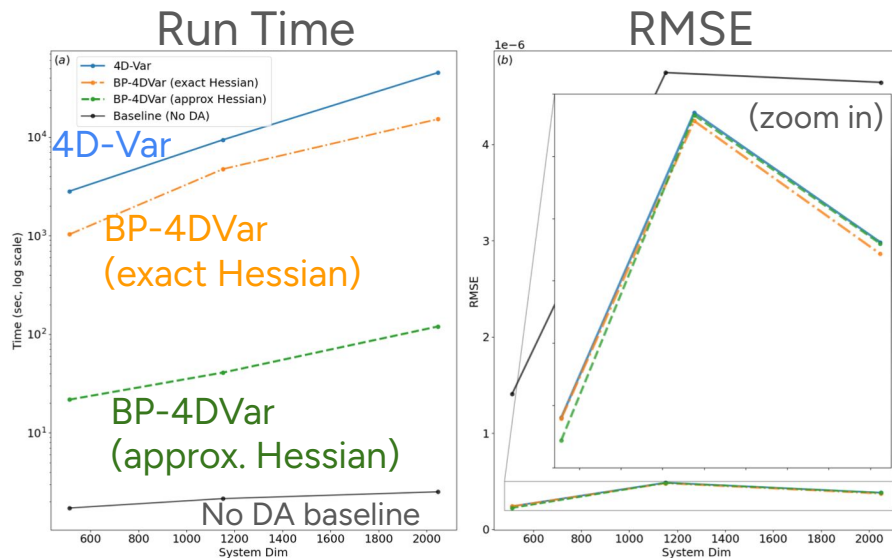


# Employ advanced data assimilation (DA) methods

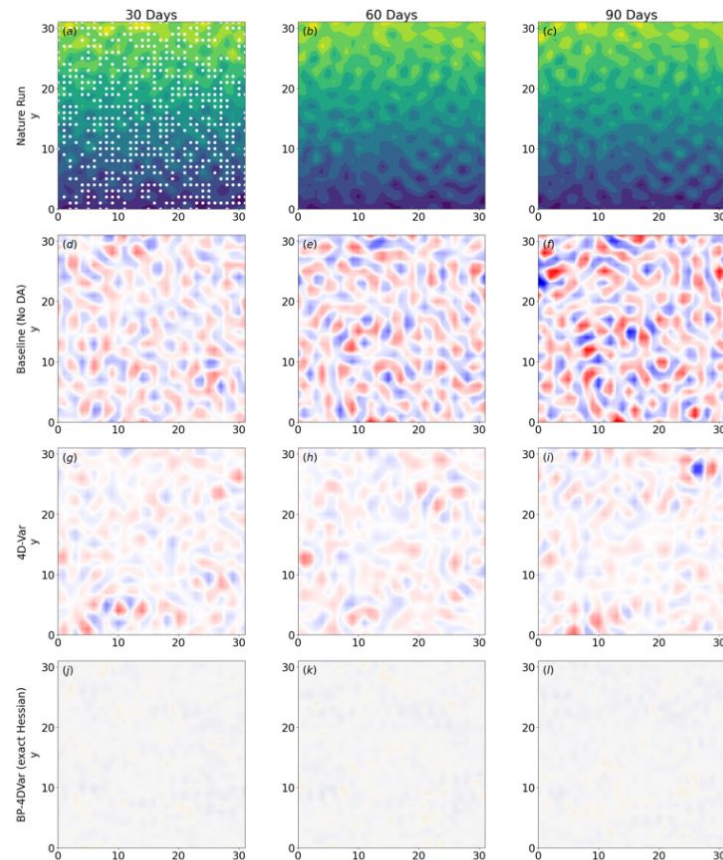
Solvik, Penny, and Hoyer (2025) "4D-Var Using Hessian Approximation and Backpropagation Applied to Automatically Differentiable Numerical and Machine Learning Models" <https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004608>

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

Time (sec, log scale)



**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.



# DataAssimBench

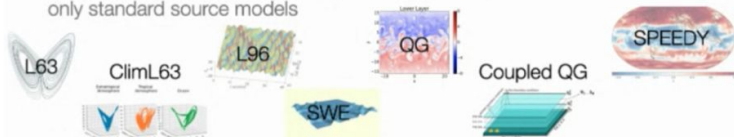
Google-funded development of a utility for integrating AI/ML with Data Assimilation (with JAX support) **benchmarking** and “**work bench**”, inspired by WeatherBench

<https://github.com/StevePny/DataAssimBench>

<https://github.com/StevePny/DataAssimBench-Examples>

Where are the benchmark training sets?

- We generate them from known dynamical systems, there is no standard set, only standard source models



- This is a stepping stone toward the use of more realistic models, real real-world observational data

NOAA UFS      FV3-GFS      MOM6      IFS      Hycom      COAMPS      NEMO      Wavewatch3      Navgen

Contact:

[steve.penny@sofarocean.com](mailto:steve.penny@sofarocean.com)

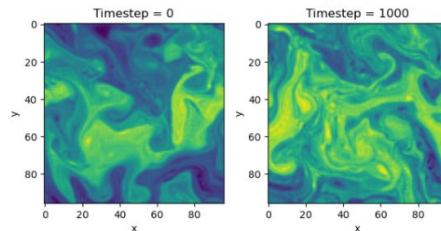
Steve Penny, Kylen Solvik (CU),  
Stephan Hoyer (Google),  
Tim Smith (CIRES/NOAA),  
Tse-Chun Chen (PNNL)

```
In [1]: from dabench.data import sqgturb
import matplotlib.pyplot as plt

In [2]: model_obj = sqgturb.DataSQGturb()
model_obj.generate(n_steps = 1000)
gridded_vals = model_obj.to_original_dim()
```

```
In [3]: fig, ax = plt.subplots(1, 2)
fig.suptitle('SQG Turbulence Model, Potential Vorticity (PVU)')
ax[0].imshow(gridded_vals[0, 1])
ax[0].set_title('Timestep = 0')
ax[0].set_xlabel('x'); ax[0].set_ylabel('y')
ax[1].imshow(gridded_vals[-1, 1])
ax[1].set_title('Timestep = 1000')
ax[1].set_xlabel('x'); ax[1].set_ylabel('y')
fig.tight_layout()
fig.subplots_adjust(top=1.2)
plt.show()
```

SQG Turbulence Model, Potential Vorticity (PVU)

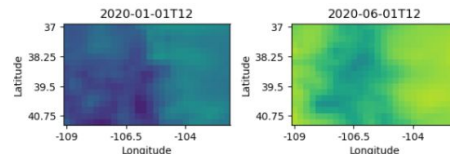


```
In [1]: from dabench.data import aws
import matplotlib.pyplot as plt
import numpy as np

In [2]: data_obj = aws.DataAWS(variables = ['air temperature_at_2_metres'],
years = [2020, 2021],
min_lat = 36.992426, max_lat = 41.003444,
min_lon = -109.060253, max_lon = -102.041524
)
data_obj.load()
gridded_values = data_obj.to_original_dim()
```

```
In [3]: fig, ax = plt.subplots(1, 2)
fig.suptitle('Air Temp at 2 Metres (K), Colorado')
ax[0].imshow(gridded_values[12], vmin=250, vmax=300)
ax[0].set_title(np.datetime.as_string(data_obj.times[12], unit='h')); ax[0].set_xlabel('Longitude')
ax[0].set_ylabel('Latitude')
ax[1].imshow(gridded_values[3660], vmin=250, vmax=300)
ax[1].set_title(np.datetime.as_string(data_obj.times[3660], unit='h')); ax[1].set_xlabel('Longitude')
ax[1].set_ylabel('Latitude')
fig.tight_layout()
fig.subplots_adjust(top=1.4)
plt.show()
```

Air Temp at 2 Metres (K), Colorado



# The Discerned Learning Challenge

What we'd like: the underlying system dynamics of the real-world/nature

# The Discerned Learning Challenge

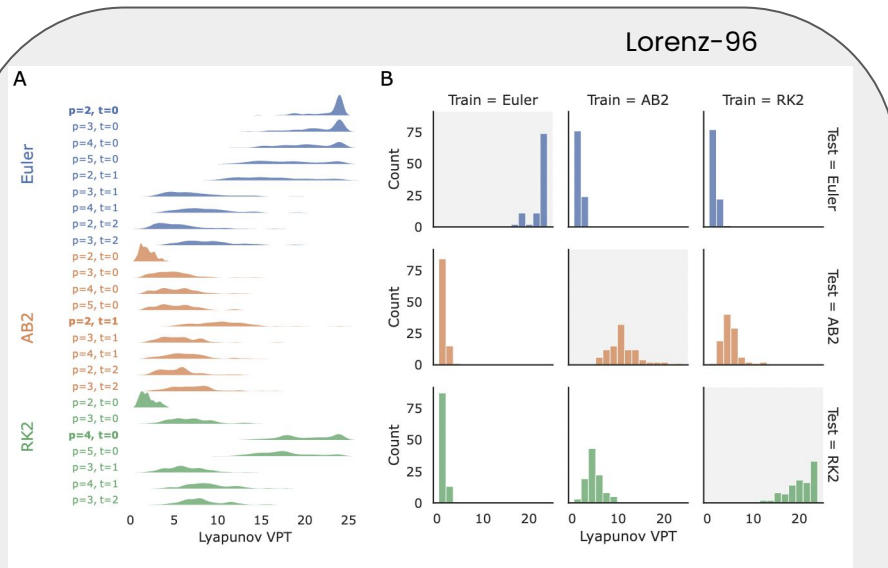
What we'd like: the underlying system dynamics of the real-world/nature

Actual Answer: Everything it can from the training data, even things you may not want



# The Discerned Learning Challenge

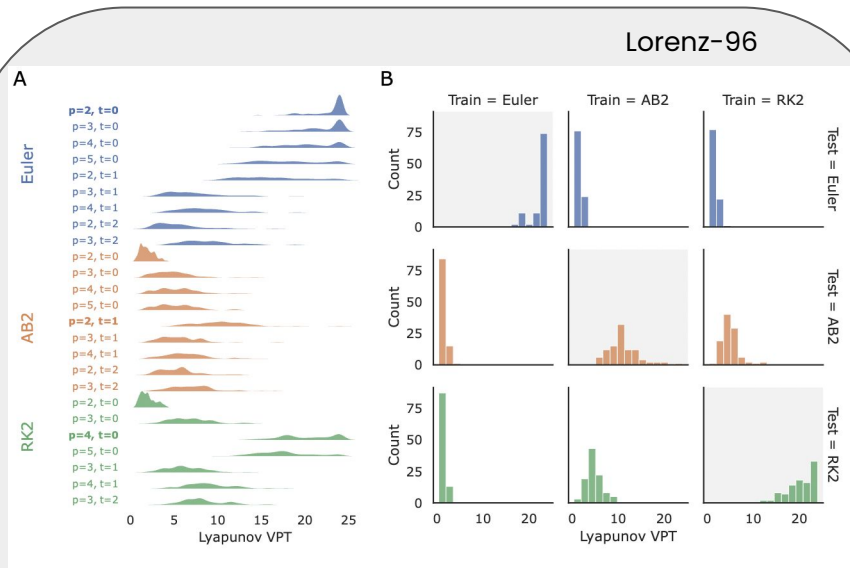
Chen, Penny, Smith, Platt (2025) "Machine Learned Empirical Numerical Integrator from Simulated Data" <https://doi.org/10.1175/AIES-D-23-0088.1>



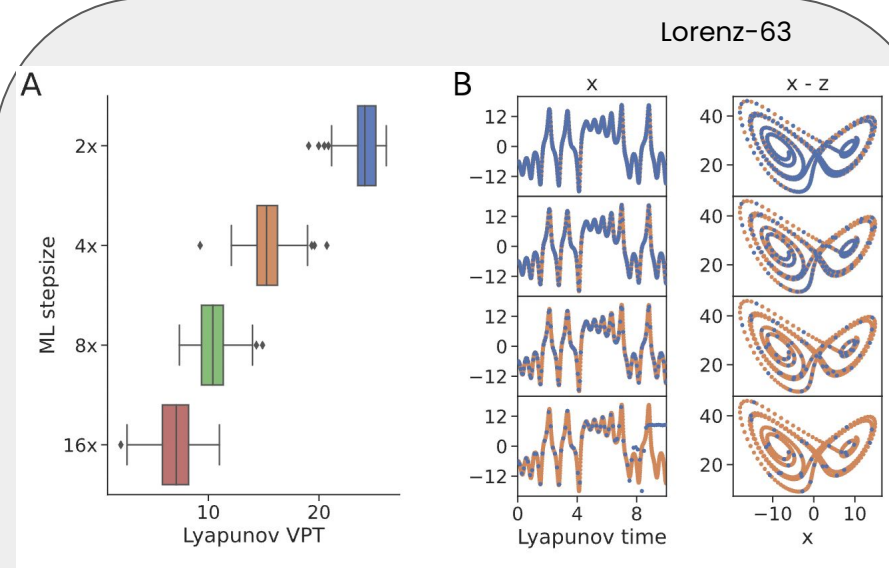
- The best models trained on datasets generated by different numerical integrations methods cross validated
- Prediction skill is *significantly worse* when predicting across integration methods
- Highlights dangers in using reanalysis datasets as "truth" for ML training

# The Discerned Learning Challenge

Chen, Penny, Smith, Platt (2025) "Machine Learned Empirical Numerical Integrator from Simulated Data" <https://doi.org/10.1175/AIES-D-23-0088.1>



- The best models trained on datasets generated by different numerical integrations methods cross validated
- Prediction skill is *significantly worse* when predicting across integration methods
- Highlights dangers in using reanalysis datasets as "truth" for ML training



- A 16x subsampling of training data *significantly reduces* forecast skill
- E.g. ERA5 6-hour timestep versus IFS with a 720s timestep is 30x subsampled, or versus FV3 nonhydrostatic 150s timestep is 144x.
- Highlights dangers in temporal subsampling in reanalysis datasets for ML training

# The Discerned Learning Challenge

A reanalysis dataset like ERA5 has many sources of hidden 'noise', for example:

- Choice of Model(s)
  - Integration method for the dynamics
  - Model time step and output frequency
  - Physics parameterizations
  - Data assimilation analysis method
    - Ensemble behavior
    - Tangent linear / adjoint
  - Static background error estimate
  - Dynamic background error estimate
  - Hybrid weighting choices
- Set of observation platforms assimilated
  - Observation network spatial structure
  - Observation quality control
  - Observation instrument errors
  - Observation representativeness errors
  - Observation error correlations
  - Observation platform biases
  - Observation bias correction method

# Summary: The Discerned Learning Challenge

- AI/ML models learn everything and do not filter physics from numerics and other assumptions/choices made in the training set
- To build true observation-to-forecast capabilities, all of the considerations that have gone into products like ERA5 and operational forecast systems must be accounted for within MLWP models



# In conclusion: What does DA bring to AI/ML?

- Awareness of the **weather as a chaotic dynamical system** - a feature that has been leveraged by the most successful DA approaches (e.g. 4D-Var, EnKF)
- Can **deal with sparse observations** to produce full atmospheric state estimates - *a key test*
- Treatment as a **Bayesian inference problem**, building a trajectory from past states to future states
- Careful **accounting of the errors** in observations, in the forecast model, and in system state estimates
- A focus on **observations as a measure of truth** - not reanalysis datasets, which are themselves a product of DA
- The expert knowledge that goes into building reanalysis systems that must be incorporated into the next generation of observation-to-forecast AI/ML approaches (there are no datasets to *machine learn* this from today)



Fin



## Stephen G. Penny

Head of Weather

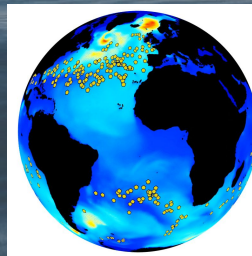
+1.415.230.2299

[steve.penny@safaroocean.com](mailto:steve.penny@safaroocean.com)

[www.safaroocean.com](http://www.safaroocean.com)

Access our global Spotter  
sensor weather network:

[weather.safaroocean.com](http://weather.safaroocean.com)



# Appendix



# ML obs-to-forecast examples

- “Data Assimilation” (ML) Interfaced ML forecast model
  - FuXi-Weather: Sun et al. (2024) <https://arxiv.org/abs/2408.05472>
- “Data Fusion” - (multiple observation types mapped to an analysis)
  - Manshausen et al. (2025) <https://arxiv.org/pdf/2406.16947>
  - Maddy et al. (2024) <https://ieeexplore.ieee.org/document/10520901>
- “Multivariate Autoregression” (multiple observation types mapped to a forecast)
  - Aardvark: Allen et al. (2025) <https://www.nature.com/articles/s41586-025-08897-0>
  - McNally et al. (2024) <https://arxiv.org/pdf/2407.15586>
  - Alexe et al. (2024) <https://arxiv.org/pdf/2412.15687>



# Details of Aardvark

- It's not cycled
  - No recognition or addressing of chaotic dynamics
  - No evolution of dynamical forecast errors
  - No Bayesian Inference (in time)
- Uses a 24-hour analysis window but does not use a background field
- It's not clear how the temporal and spatial relationships of the observations is handled

## Issues:

- ERA5 is used for training and as “ground truth” for evaluating forecasts
- “Outperforms” the GFS, except for z500, versus ERA5  
(GFS is closer to GFS analysis, and ECMWF IFS is closer the ECMWF analysis, so this is not a meaningful result)



# Details of FuXi Weather

- Cycles, using background and observations in 8-hour window (+)
- It's not clear how the temporal and spatial relationships of the observations is handled during the FuXi-DA step

## Issues:

- “the analysis fields from FuXi-DA are less accurate than ERA5, resulting in a marked degradation in forecast performance”
- “FuXi Weather consistently outperforms ECMWF HRES in observation sparse regions, such as central Africa”



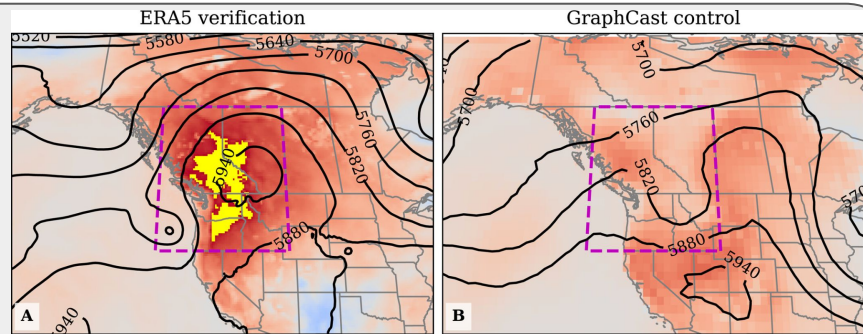
# AI/ML Weather models resolve large scales

Vonich and Hakim (2024): "Predictability Limit of the 2021 Pacific Northwest Heatwave From Deep-Learning Sensitivity Analysis" <https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2024GL110651>

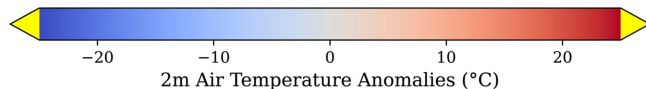
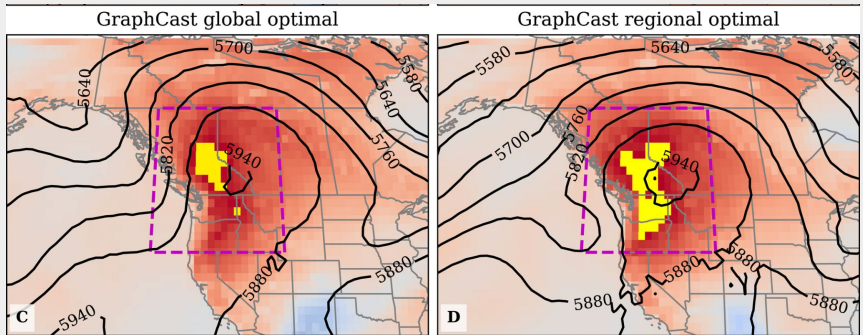
AI/ML "Weather" models seem to recover planetary-scale dynamics at subseasonal timescales

MLWP models do have sensitivity to initial conditions at large spatial scales (e.g. > 500-700km) and long time scales (e.g. > 5-10 days).

Original ERA5 analysis versus 1.0° GraphCast forecast from ERA5 initial conditions



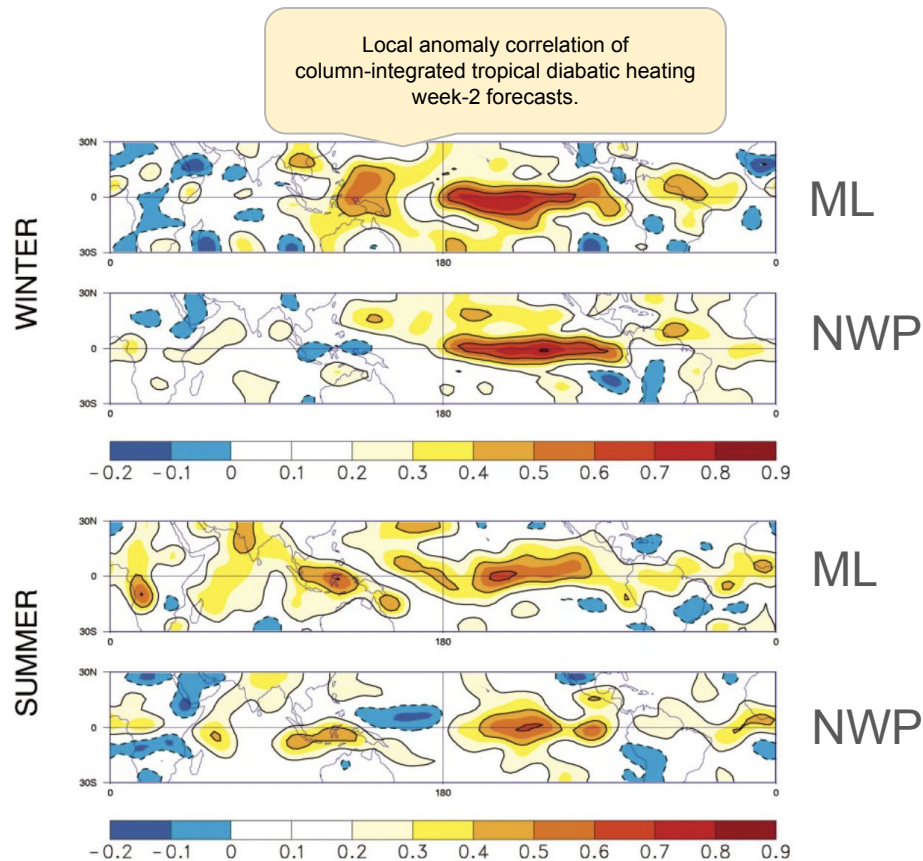
GraphCast forecast after optimizing initial conditions (globally or regionally)





# AI/ML Weather models resolve large scales

ML models can capture similar patterns of anomaly correlation as NWP models



# AI/ML Weather models resolve large scales

Newman et al. (2003) "A Study of Subseasonal Predictability" <https://doi.org/10.1175//2558.1>

ML models can capture similar patterns of anomaly correlation as NWP models - the linear inverse model (LIM) from 1989:

$$\mathbf{X} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_m]$$

$$\mathbf{X}' = [\mathbf{x}_2 \quad \mathbf{x}_3 \quad \cdots \quad \mathbf{x}_{m+1}]$$

$$\mathbf{X}' = \mathbf{A}\mathbf{X}$$

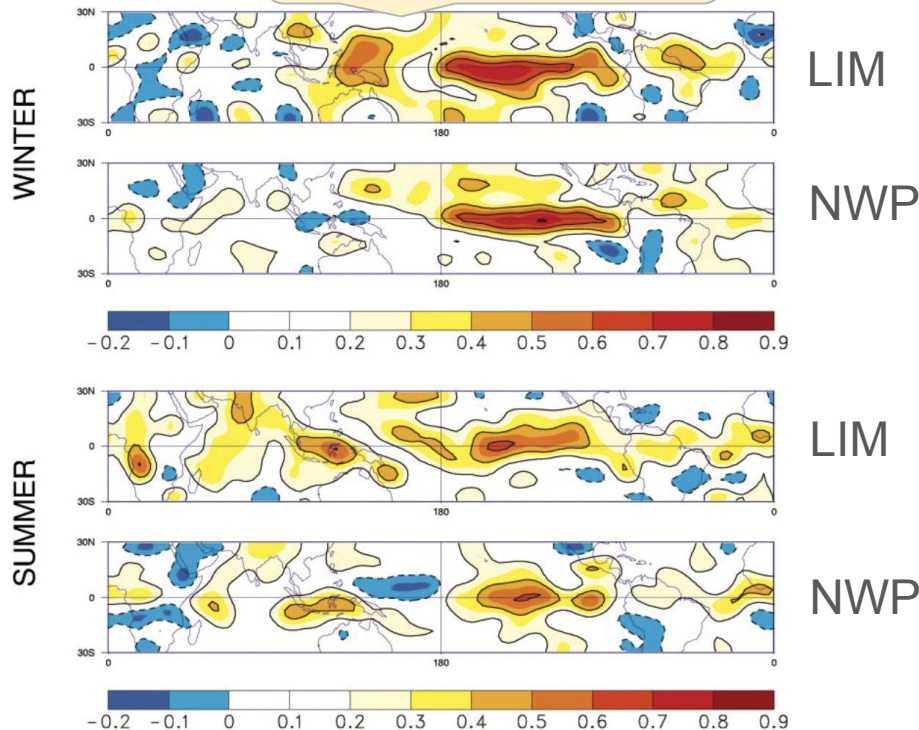
$$\mathbf{B} = \exp(\mathbf{A})$$

$$\frac{d\mathbf{x}}{dt} = \mathbf{B}\mathbf{x} + \mathbf{F}_s$$

Penland, C. (1989) "Random forcing and forecasting using principal oscillation pattern analysis"

[https://doi.org/10.1175/1520-0493\(1989\)117<2165:RF AFUP>2.0.CO;2](https://doi.org/10.1175/1520-0493(1989)117<2165:RF AFUP>2.0.CO;2)

Local anomaly correlation of column-integrated tropical diabatic heating week-2 forecasts.





How can we leverage AI/ML in operational forecasting today?



# How can we leverage AI/ML in operational forecasting today?

Build operational forecast systems that are world-leading *today* while leveraging what AI/ML has to offer:

Step 1: Focus AI/ML to enhance NWP. This includes mapping, averaging, and combining forecasts, observations, and simulations efficiently and effectively

Step 2: Modernize physics-based forecast models. Ensure they  
(a) run on GPUs, (b) support automatic differentiation, and  
(c) can interface with ML models and ML optimization

Step 3: Employ advanced data assimilation (DA) methods. The aim is to simultaneously leverage as many observations as possible, and eliminate dependency on external NWP analysis products (e.g. ECMWF HRES initial conditions)

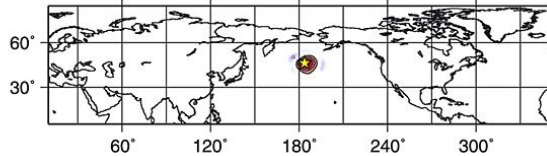
# Are large ensembles useful?

Miyoshi et al. (2014) The 10,240-member ensemble Kalman filtering with an intermediate AGCM  
<https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2014GL060863>

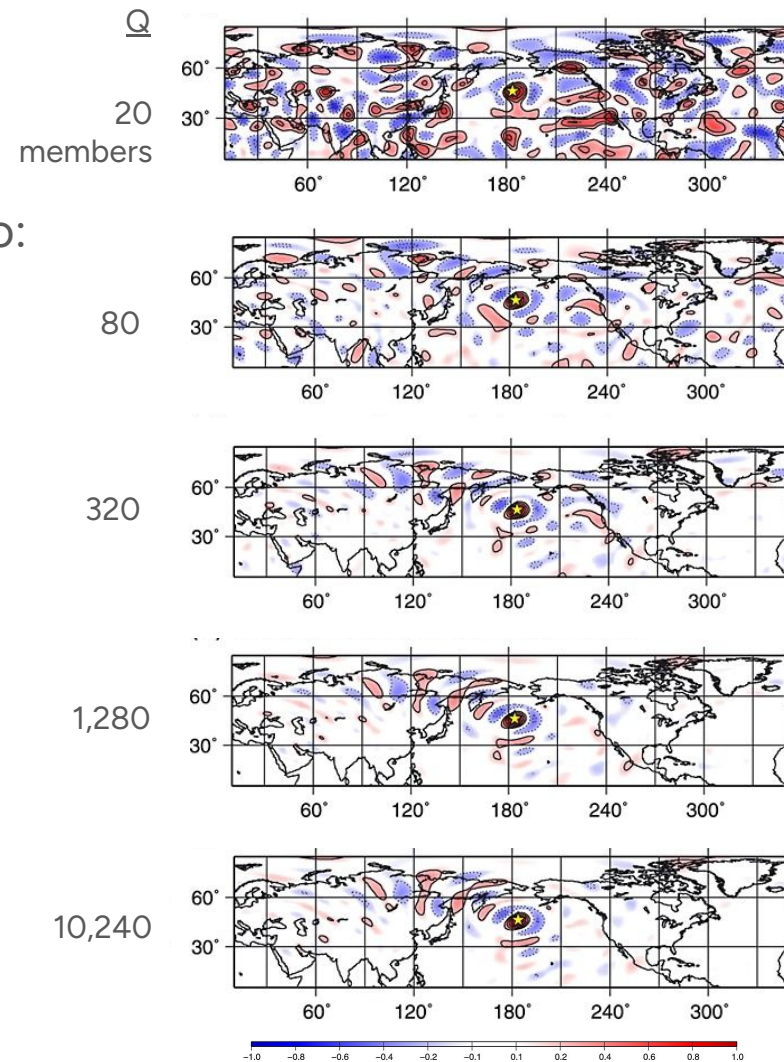
Yes, we know that large ensembles can be used to:

- Reduce the need for ad hoc horizontal localization schemes and produce more accurate analyses
- Eliminate the need for vertical localization, which is challenging to apply with vertically integrated observations (e.g. AVHRR/VIIRS radiometers)

(b) 20 members w/ 700-km localization



Ensemble error correlations with respect to the center point at 46.4°N, 176.3°W (yellow star) computed with 3.75° x 3.75° SPEEDY model.



# Are large ensembles useful?

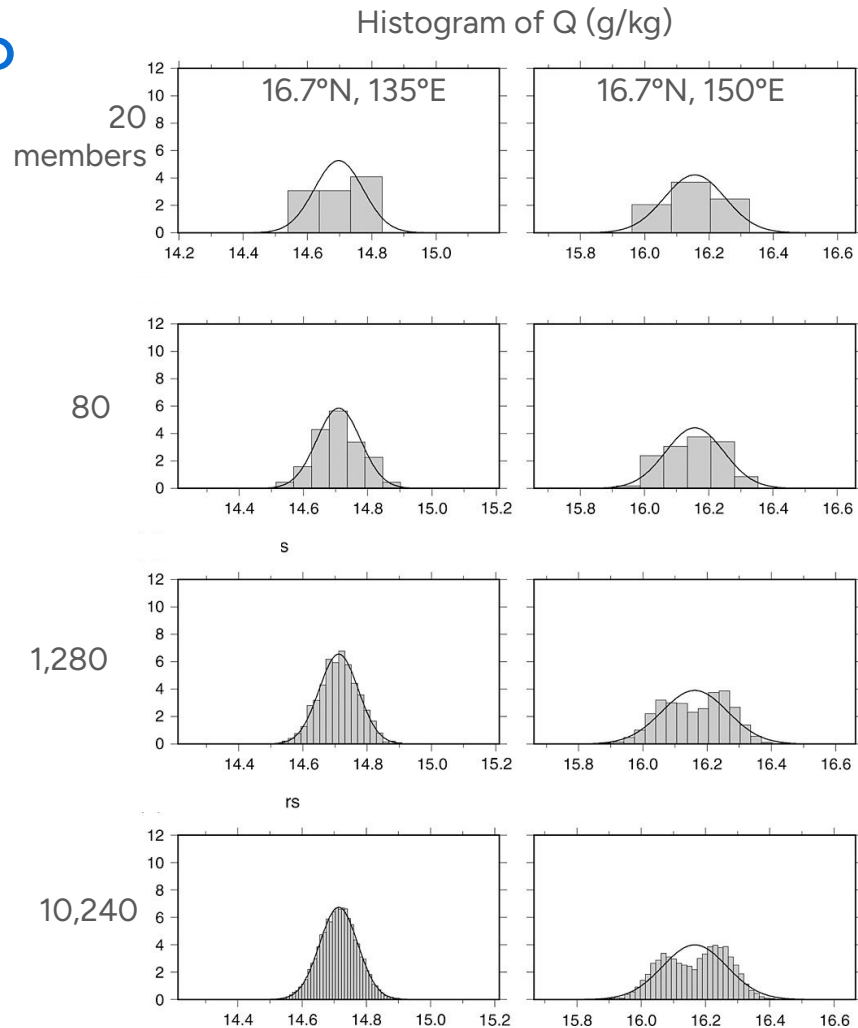
Miyoshi et al. (2014) The 10,240-member ensemble Kalman filtering with an intermediate AGCM

<https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2014GL060863>

Yes, we know that large ensembles can be used to:

- Resolve non-Gaussian forecast distributions.

Even at low resolution (i.e.  $3.75^\circ \times 3.75^\circ$ ), bimodal distributions arise in 6-hour forecasts for fields such as the specific humidity ( $Q$ ) with ensemble sizes  $> 1,000$ .



# Are large MLWP ensembles useful?

Answer: No, not yet.

Three examples to test validity of model response to perturbations in the state:

1. Tangent Linear Model (TLM) tests show the impact of a small perturbation to the initial conditions.
2. Adjoint model tests show the “upstream” changes to a model needed to produce the resulting
3. A cycled Ensemble Kalman Filter (EnKF) quantifies the accuracy of forecast error covariance statistics, which describe the first order relationships between all variables both locally and at a distance.



# Applying DA with AI/ML

From Olivier Talagrand, one of the most influential figures in DA in the last 40 years categorizes methods based on statistical estimation theory:

“Assimilation of meteorological or oceanographical observations can be described as the process through which all the available information is used in order to estimate as accurately as possible the state of the atmosphere or oceanic flow. The available information essentially consists of the observations proper, and the physical laws that govern the evolution of the flow. The latter are available in practice under the form of a numerical model. The existing assimilation algorithms can be described as either sequential or variational.” (*Talagrand, 1997*)

