Unpacking AI Weather Emulators

What and How Do They Learn?

EB Elizabeth Barnes
Boston University

PH Pedram Hassanzadeh
University of Chicago

TS Tiffany Shaw
University of Chicago

RW Rebecca Willett
University of Chicago





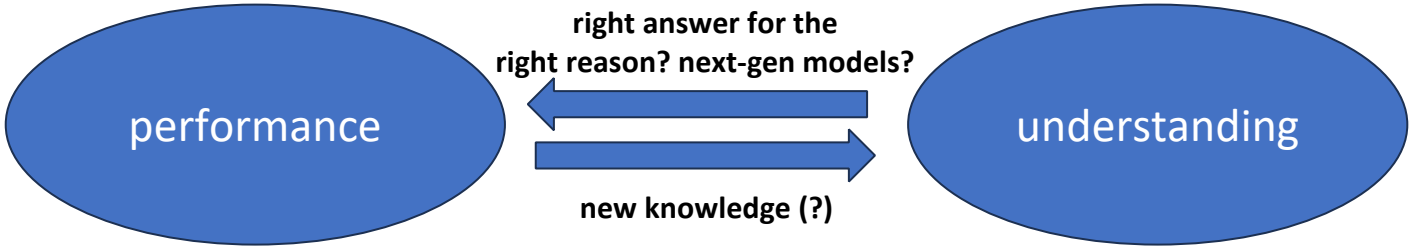
Institute for Mathematical and Statistical Innovation

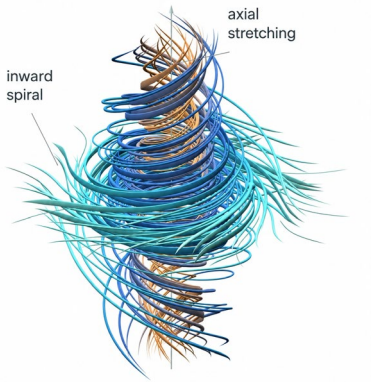


AICE: AI for CLIMATE

THE UNIVERSITY OF CHICAGO DATA SCIENCE INSTITUTE

THE UNIVERSITY OF CHICAGO INSTITUTE FOR CLIMATE & SUSTAINABLE GROWTH



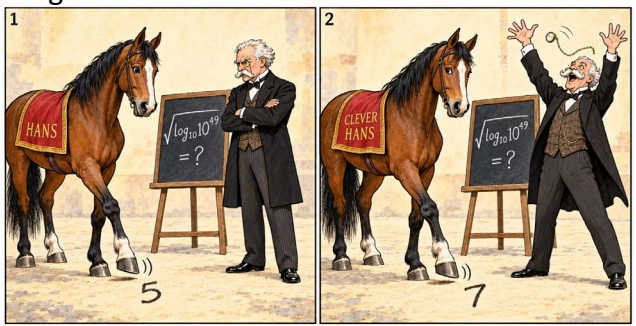


https://openai.com/index/navier-stokes-solution/

”.... solving problems is only a tool and proxy for achieving the primary goal of conceptual understanding and insight.”

Declaration from 25 Field Medalists, Sep. 2206
https://terrytao.wordpress.com/

Image is from ChatGPT



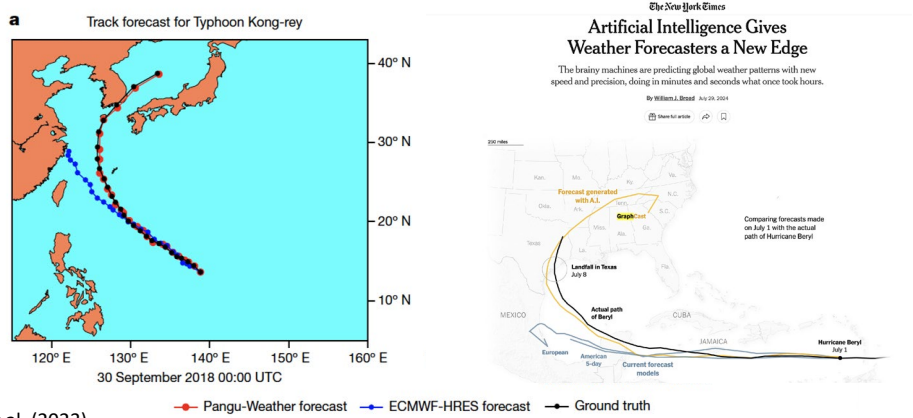
Clever Hans: a horse in Germany in 1890s who could **perform** well at math problems!

But did he **understand** math? Turned out he was good at reading facial cues of the examiners

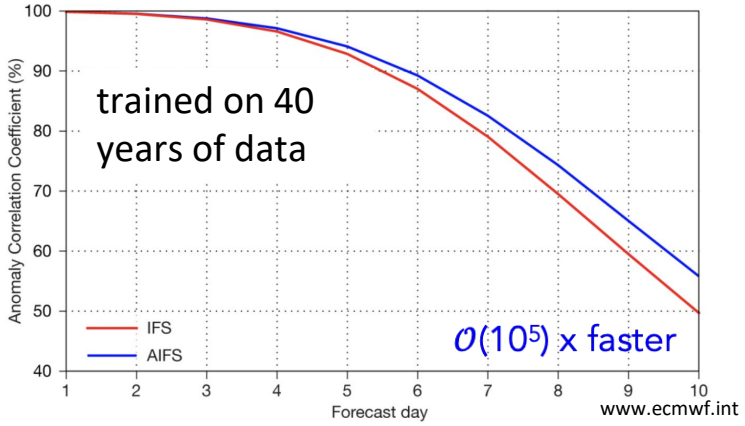
https://en.wikipedia.org/wiki/Clever_Hans

2nd Revolution in 1-15 Day Weather Forecasting (2018-now)

Surprising accuracy: 2019: Oxford workshop, impossible; 2022: FourCastNet (~IFS); 2023: Pangu & GraphCast (>IFS); 2025: operational!

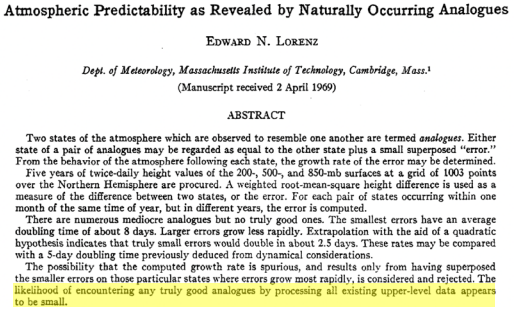


Lam et al. (2023)









Atmospheric Predictability as Revealed by Naturally Occurring Analogues

EDWARD N. LORENZ

Dept. of Meteorology, Massachusetts Institute of Technology, Cambridge, Mass.[1]

(Manuscript received 2 April 1969)

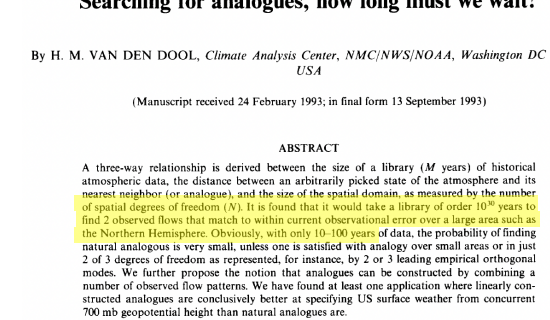
ABSTRACT

Two states of the atmosphere which are observed to resemble one another are termed *analogues*. Either state of a pair of analogues may be regarded as equal to the other state plus a small superposed "error." From the behavior of the atmosphere following each state, the growth rate of the error may be determined.

Five years of twice-daily height values of the 200-, 500-, and 850-mb surfaces at a grid of 1003 points over the Northern Hemisphere are procured. A weighted root-mean-square height difference is used as a measure of the difference between two states, or the error. For each pair of states occurring within one month of the same time of year, but in different years, the error is computed.

There are numerous mediocre analogues but no truly good ones. The smallest errors have an average doubling time of about 8 days. Larger errors grow less rapidly. Extrapolation with the aid of a quadratic hypothesis indicates that truly small errors would double in about 2.5 days. These rates may be compared with a 5-day doubling time previously deduced from dynamical considerations.

The possibility that the computed growth rate is spurious, and results only from having superposed the smaller errors on those particular states where errors grow most rapidly, is considered and rejected. The likelihood of encountering any truly good analogues by processing all existing upper-level data appears to be small.



Searching for analogues, how long must we wait?

By H. M. VAN DEN DOOL, *Climate Analysis Center, NMC/NWS/NOAA, Washington DC 20233, USA*

(Manuscript received 24 February 1993; in final form 13 September 1993)

ABSTRACT

A three-way relationship is derived between the size of a library (M years) of historical atmospheric data, the distance between an arbitrarily picked state of the atmosphere and its nearest neighbor (or analogue), and the size of the spatial domain, as measured by the number of spatial degrees of freedom (N). It is found that it would take a library of order 10^{30} years to find 2 observed flows that match to within current observational error over a large area such as the Northern Hemisphere. Obviously, with only 10–100 years of data, the probability of finding natural analogous is very small, unless one is satisfied with analogy over small areas or in just 2 of 3 degrees of freedom as represented, for instance, by 2 or 3 leading empirical orthogonal modes. We further propose the notion that analogues can be constructed by combining a number of observed flow patterns. We have found at least one application where linearly constructed analogues are conclusively better at specifying US surface weather from concurrent 700 mb geopotential height than natural analogues are.

10^{30} years

https://www.imsi.institute/activities/unpacking-ai-weather-emulators/

Next: **30-min overview** by Elizabeth Barnes

All other talks: 15 min

5 min tech breaks: quick clarifying questions + transition to the next speaker

70-90 min discussion panels: in-depth Q&A, comments/thoughts from the panelists, moderators, audience …

Audience

Main room
Overflow room
Online

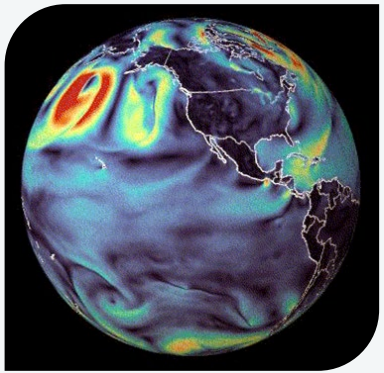
Reception: today at DSI (other side of the campus)

Poster sessions: Tuesday and Thursday

Thursday and **Friday:** Breakout groups and more discussions (more info to follow)

What and how do AI weather emulators learn?

A few explorations and a lot of open questions.

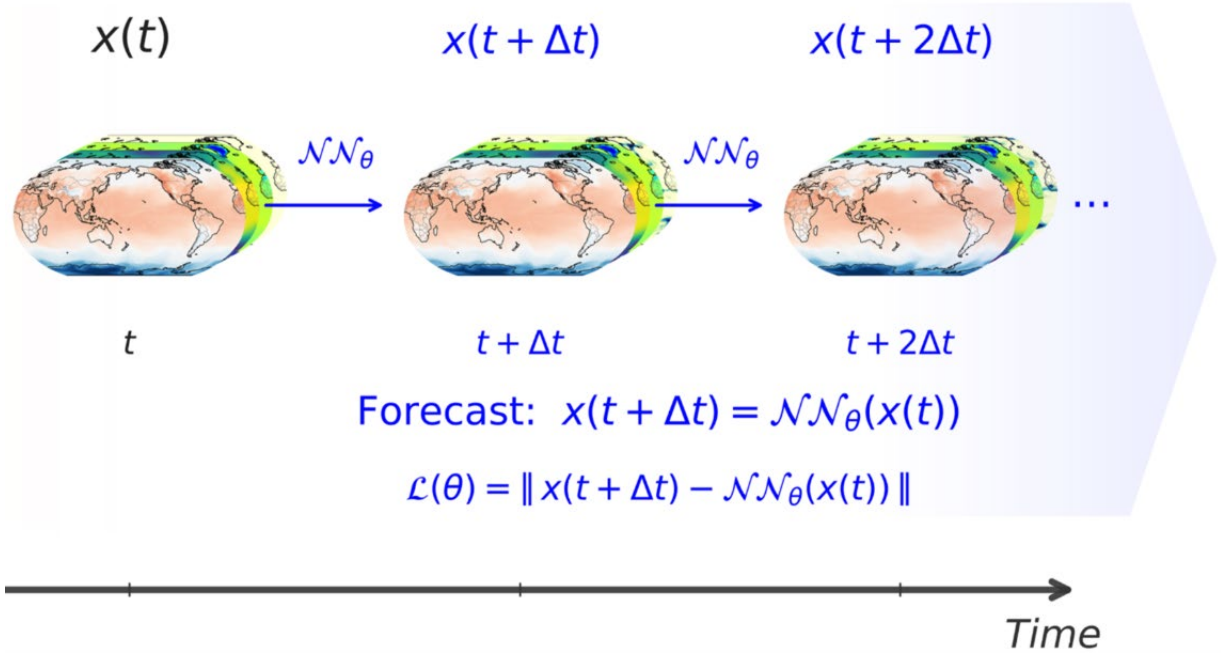


Dr. Elizabeth A. Barnes
Dalton Family Chair in Environmental Data Science & Sustainability
Professor of Computing & Data Sciences
Professor of Earth & Environment
Boston University

Unpacking AI Weather Emulators: What and How Do They Learn?
IMSI Workshop

September 28, 2026

Autoregressive AI weather & climate models



Adapted from Hassanzadeh et al. (2026)

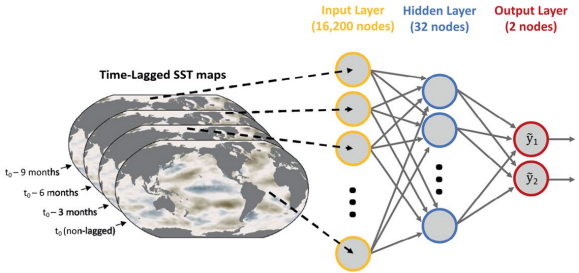
AI for weather and climate has arrived...

AI for weather and climate has arrived...

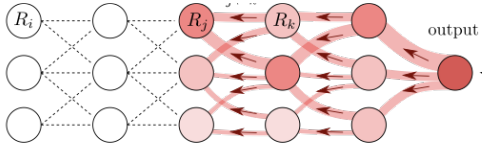
But new scientific insight has lagged behind.

XAI has been applied successfully to simpler climate prediction networks

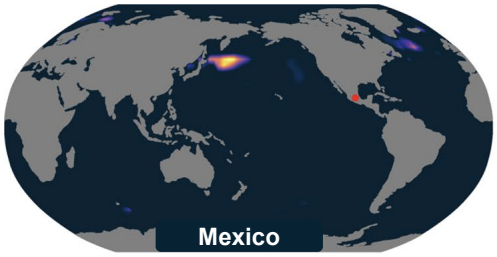
Predicting 5-year average surface temperature at each grid point

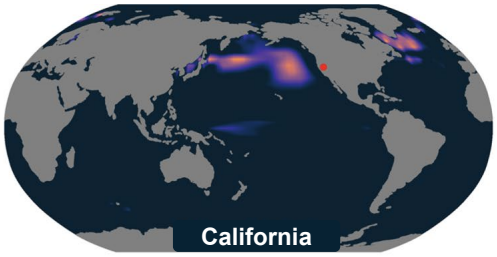


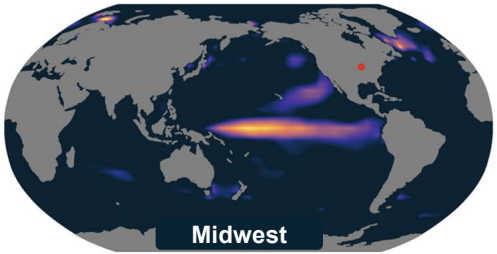
Prediction of average land surface temperature anomaly over the next 5 years



XAI





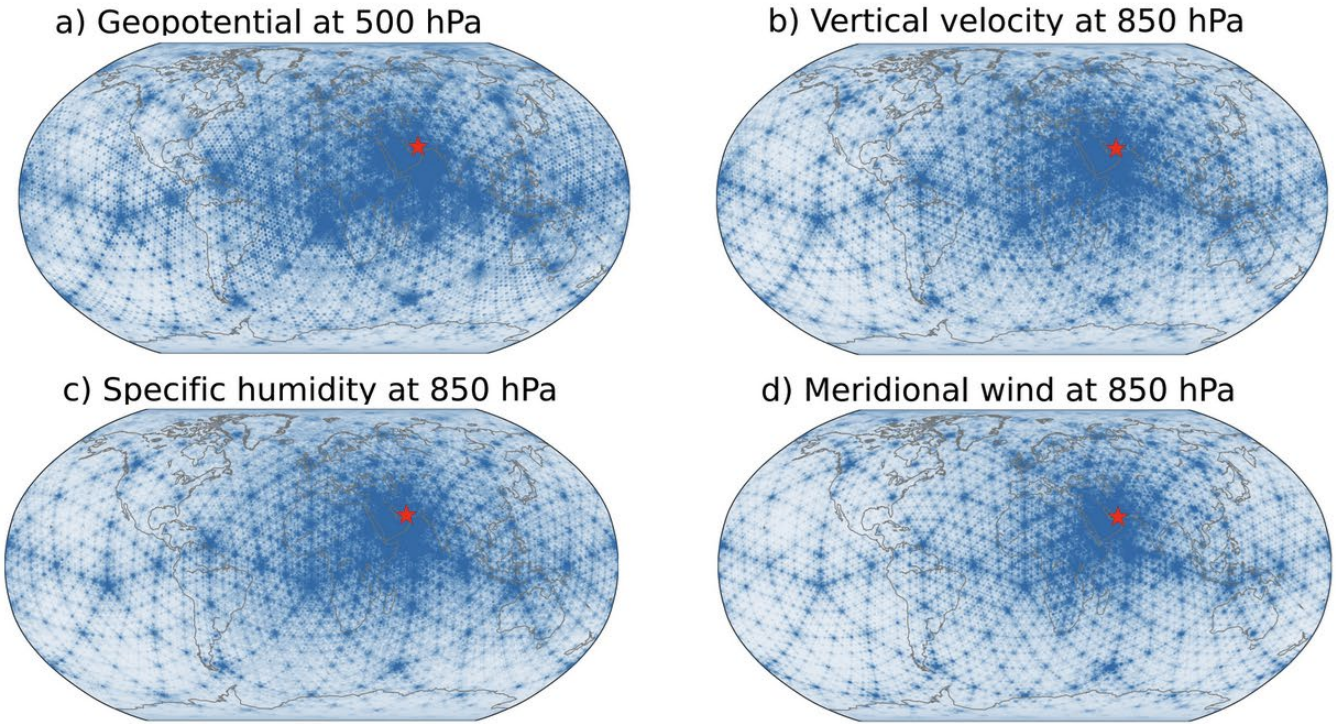






Benjamin Toms et al. (2021; GRL)

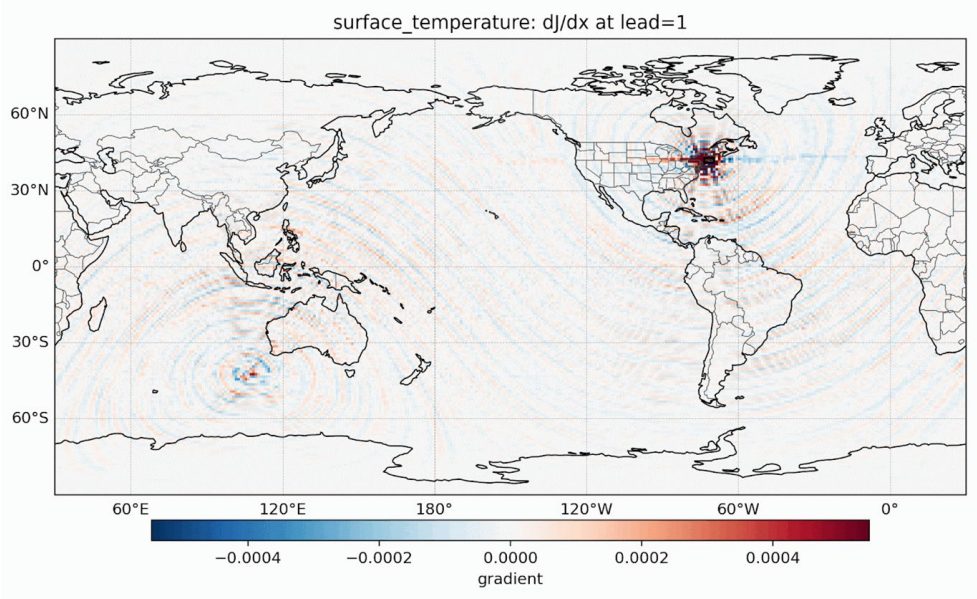
But XAI on autoregressive emulators is messy



Effective receptive field of precipitation in Graphcast for the Dubai floods.
Sun et al. (2025)

Still, we have had some recent success with long-rollouts of gradients *(post-hoc XAI)*

Backpropagating gradients through 200+ autoregressive steps (~50+ days) reveals **physical sources of information** used by the emulator.

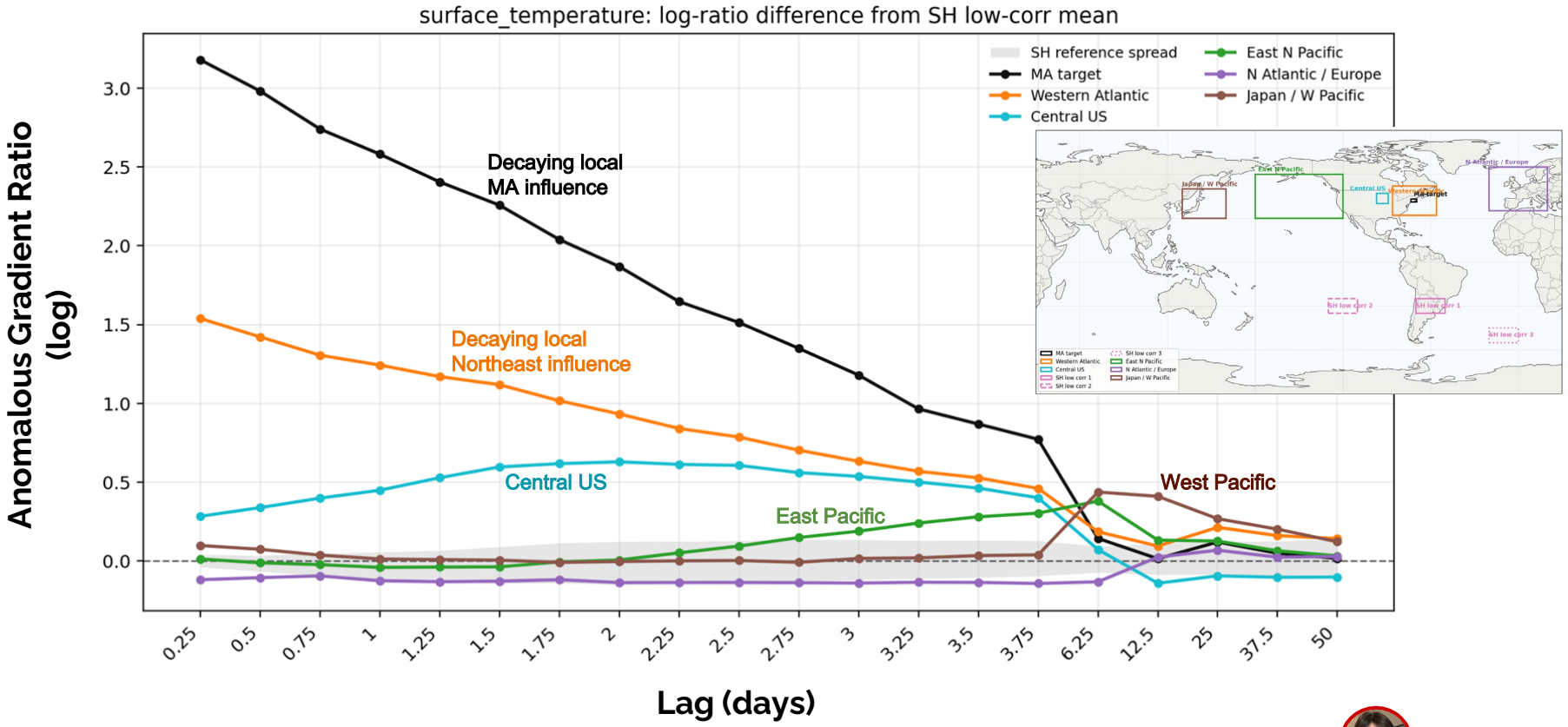


Output J -> Massachusetts surface temperature at T_0



Ziqi Tang et al. (in prep)

Different sources act at different lags





Ziqi Tang et al. (in prep)

Predictions tell us what the model **can do**.

Post-hoc XAI tells us **where they look**.

But *training* tells us **how it thinks.**

We don't have to open the black box after the fact. We can watch it being built.

There Will Be a Scientific Theory of Deep Learning
Simon, Kunin et al. (Apr. 23, 2026)

"the emerging theory is concerned with the dynamics of the training process"

Sudden Drops in the Loss: Syntax Acquisition, Phase Transitions, and Simplicity Bias in MLMs; Chen et al. (2023)

LLMs abruptly acquire capabilities at specific training moments — and these can be manipulated to improve the final result.

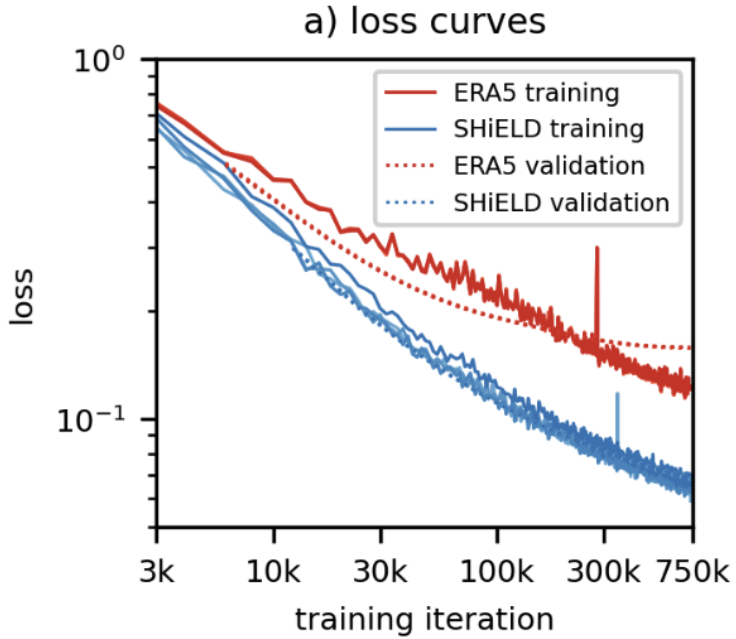
Let's bring this lens to AI weather and climate models

What happens during training?

This is what we usually look at.

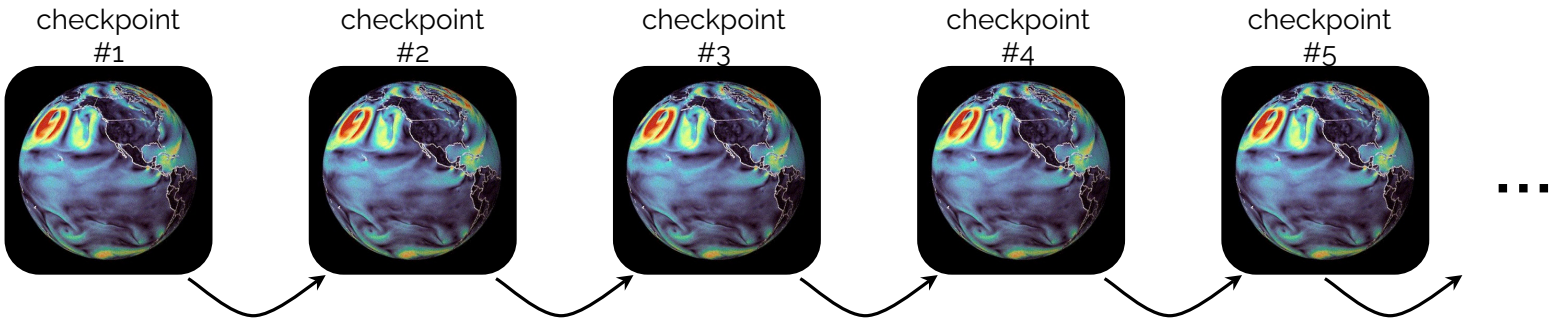
Loss goes down.

Looks good.



Watt-Meyer et al. (2025); ACE2

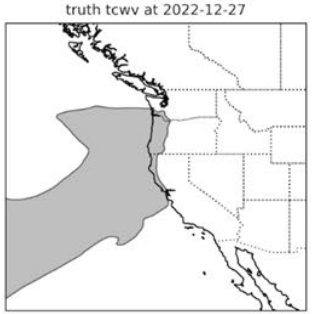
Turning training into an experiment



Use the AI model to make the **same forecast over and over**,
and **watch how it learns**

FourCastNet v2 (SFNO)
Created by NVIDIA [link]

…or watch them forget

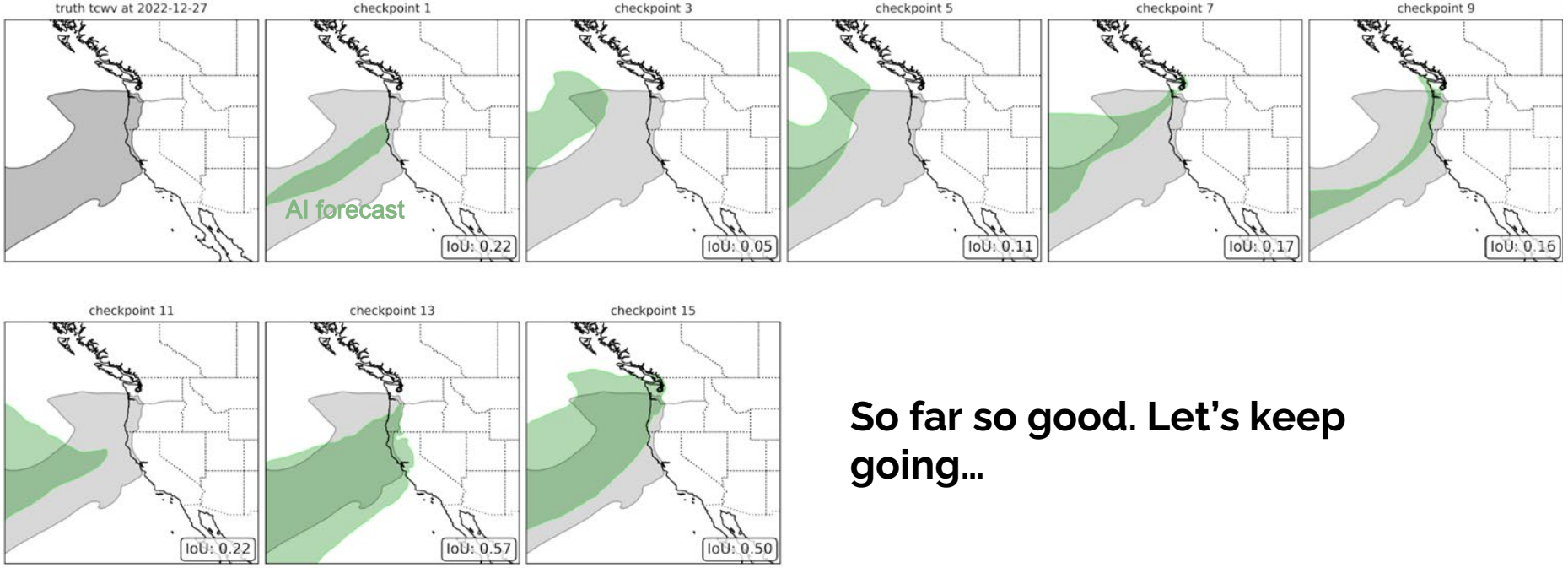


Task the AI model to **forecast** this atmospheric river event **7 days ahead**.



Annabel Wade et al. (in prep)

…or watch them forget

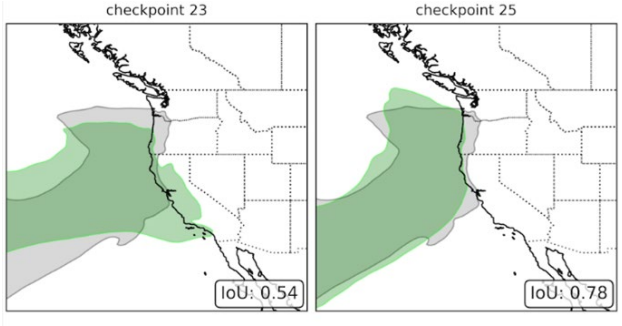


So far so good. Let's keep going…



Annabel Wade et al. (in prep)

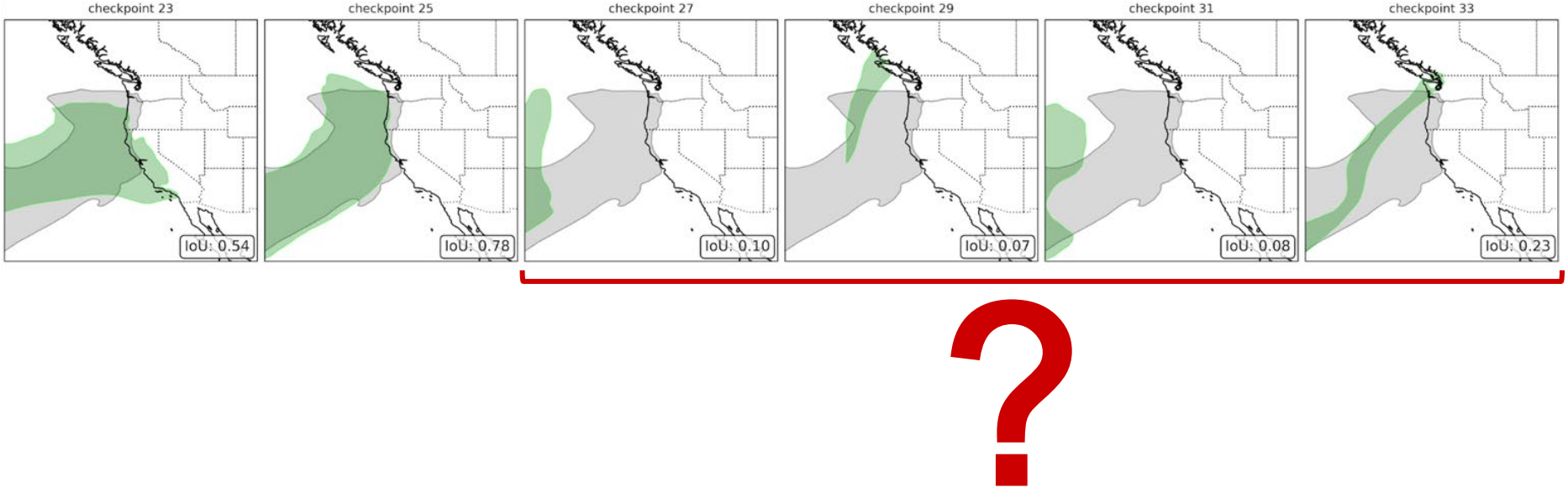
…or watch them forget





Annabel Wade et al. (in prep)

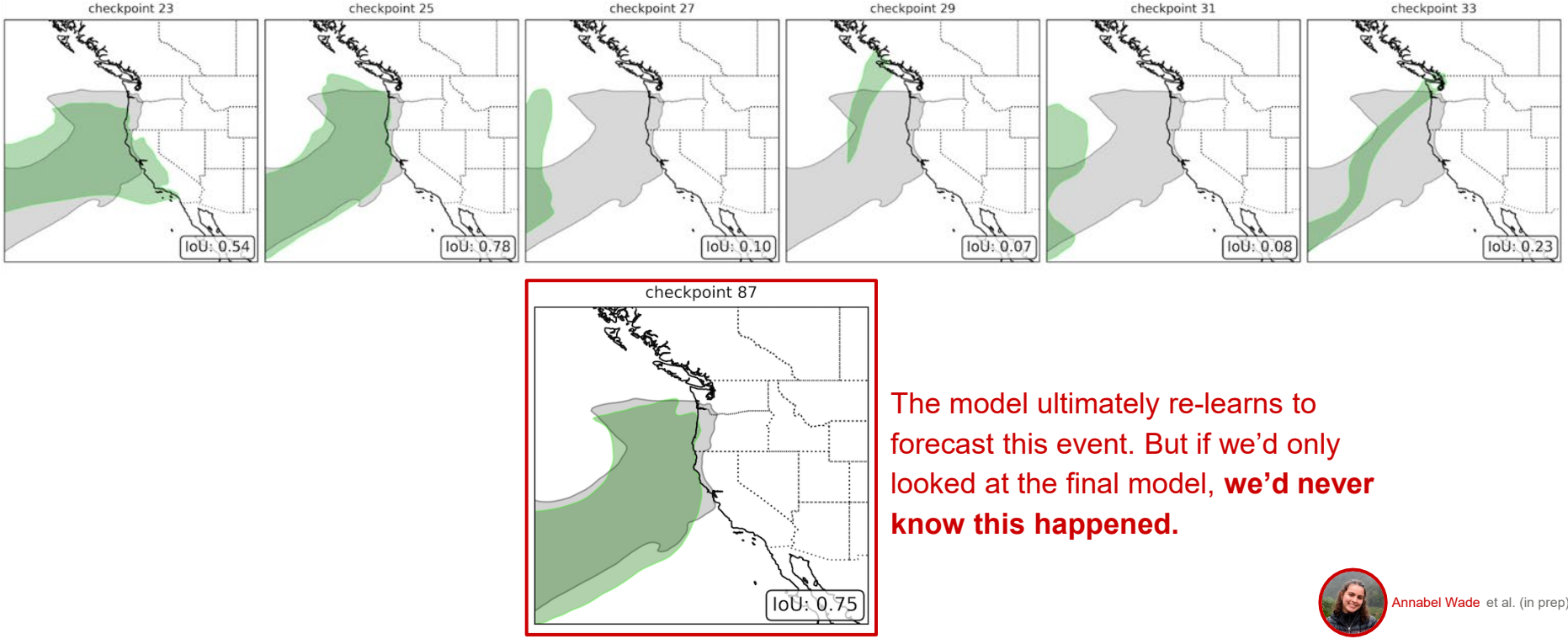
…or watch them forget





Annabel Wade et al. (in prep)

…or watch them forget

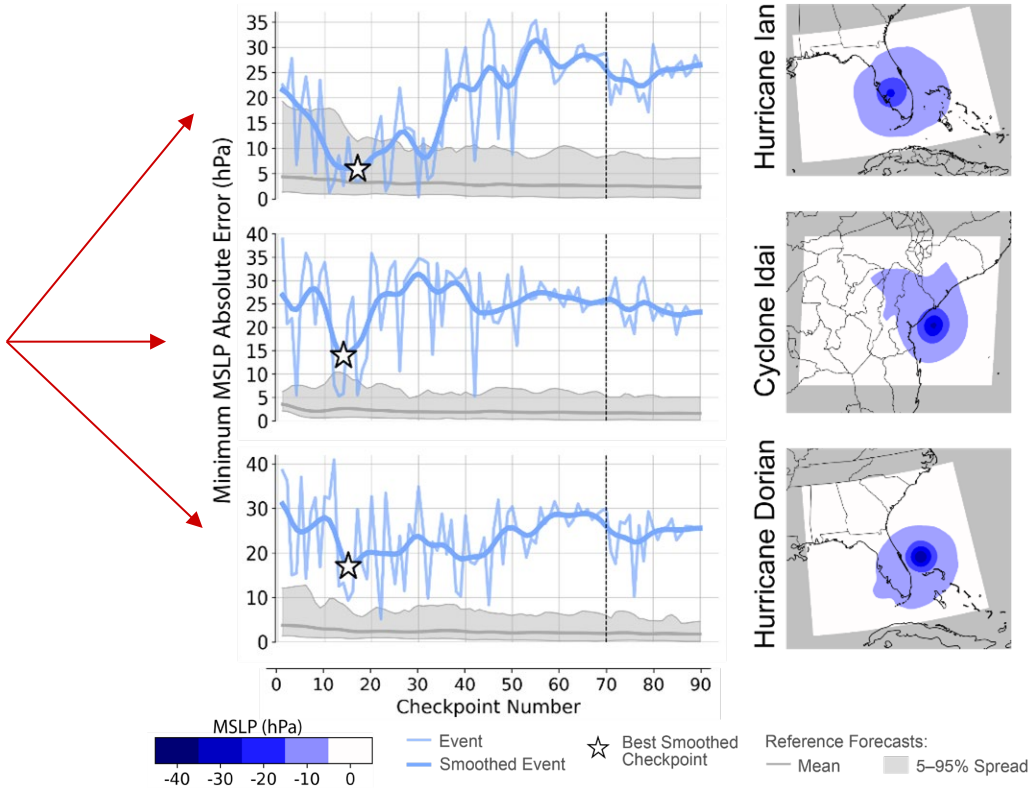


The model ultimately re-learns to forecast this event. But if we'd only looked at the final model, **we'd never know this happened.**

Annabel Wade et al. (in prep)

In some cases, the model never recovers from forgetting

Model learns to forecast peak storm intensity and then "forgets"; **never recovering** by the end of training.



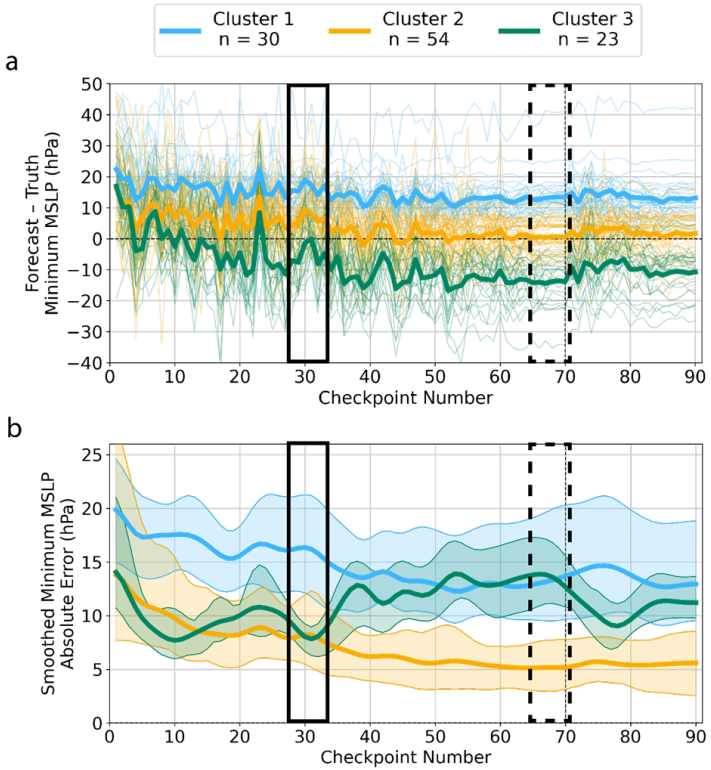


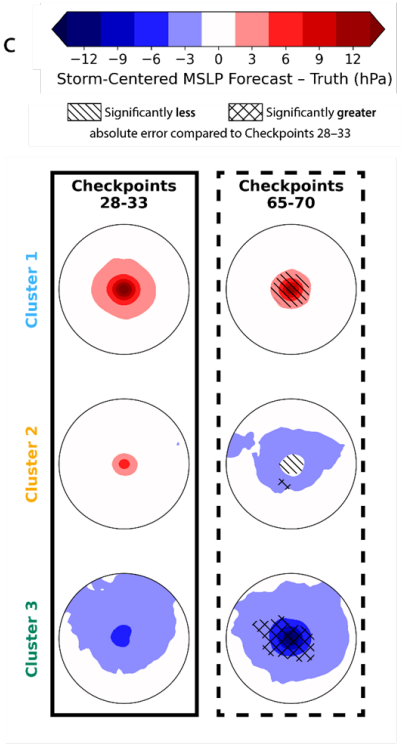
Rebecca Baiman et al. (2025, 2026)

Clustering reveals distinct learning behaviours

Clustering TC forecast errors across training checkpoints reveals three distinct learning behaviors.

One subset of storms is **actively unlearned** in the second half of training.

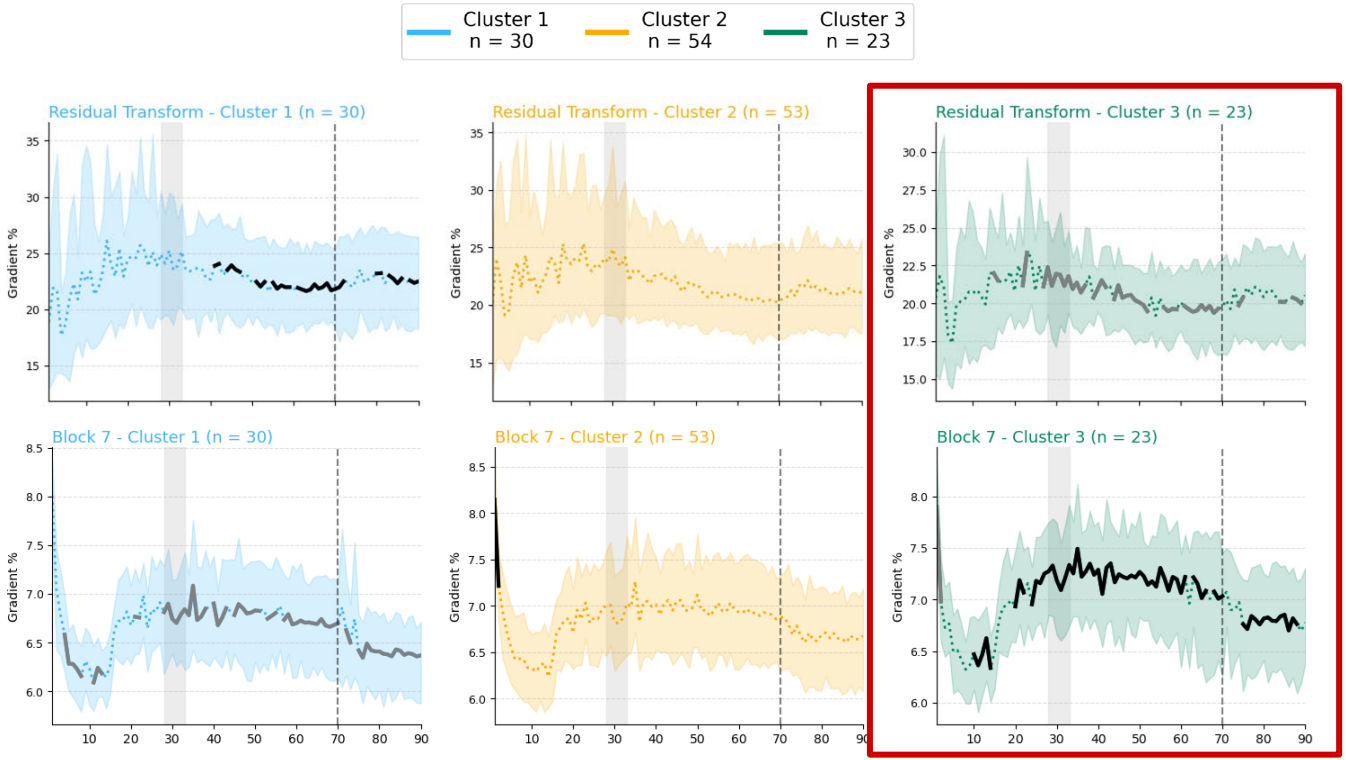






Rebecca Baiman et al. (2025, 2026)

Layer sensitivity differences for unlearned storms



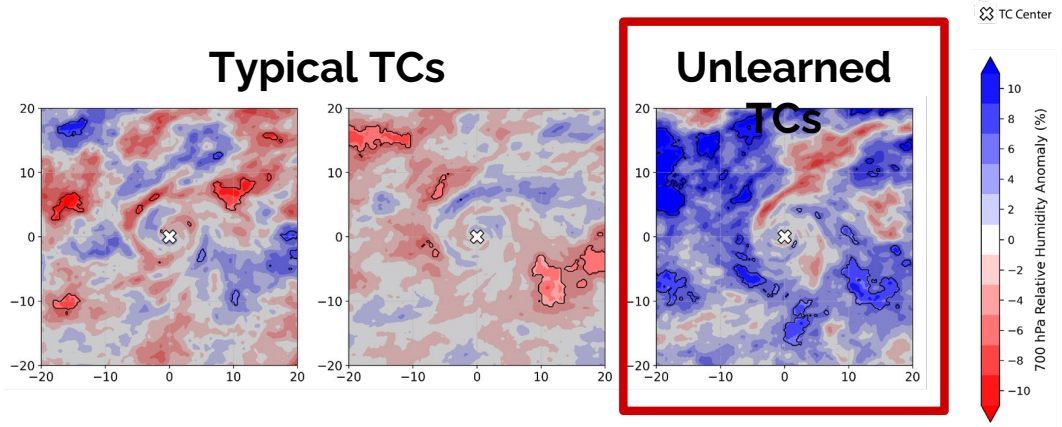


Rebecca Baiman et al. (in prep)

Unlearned storms live in more moist environments

The subset of TCs that the model unlearns have **significantly more moist environments**, although it can forecast the moisture correctly!

This suggests that the model finds **moisture - driven intensification hardest to retain.**

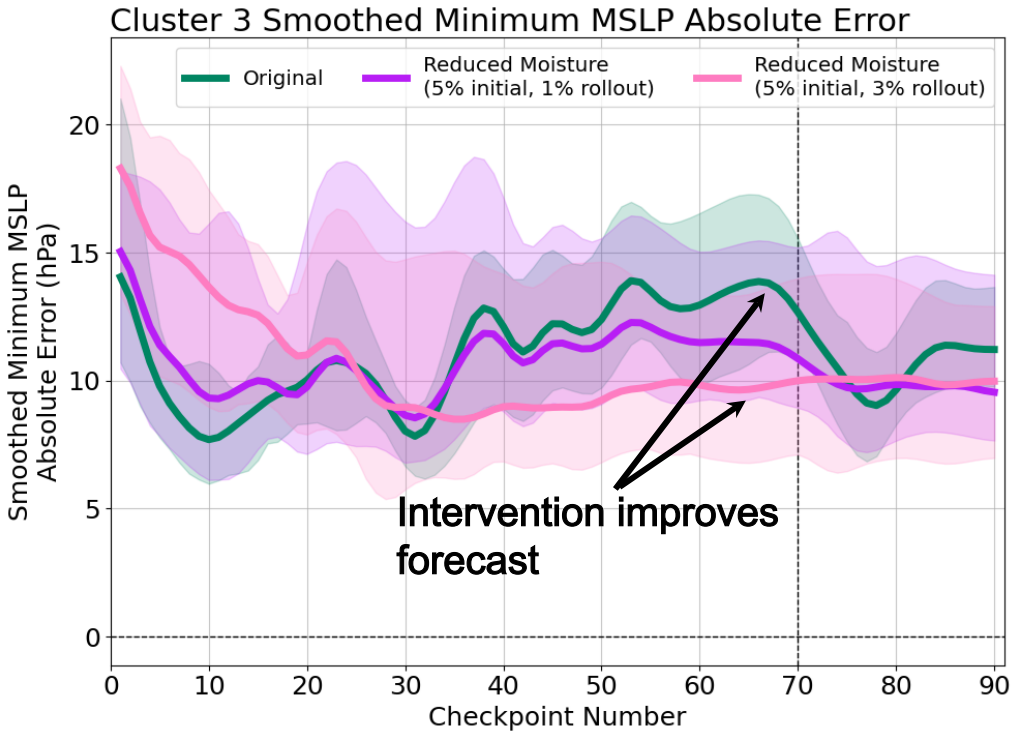




Rebecca Baiman et al. (2025, 2026)

Reducing moisture during inference fixes the issue

Intervention: if we reduce moisture through the rollout, *the unlearning trend disappears and the forecast improves!*

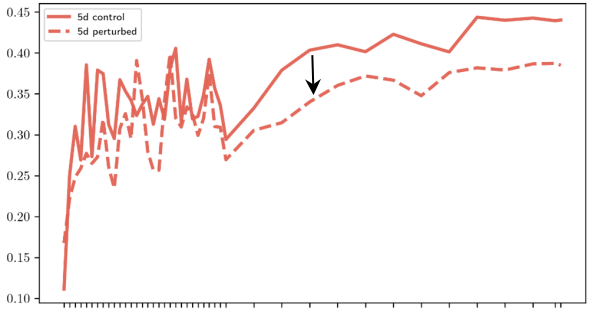




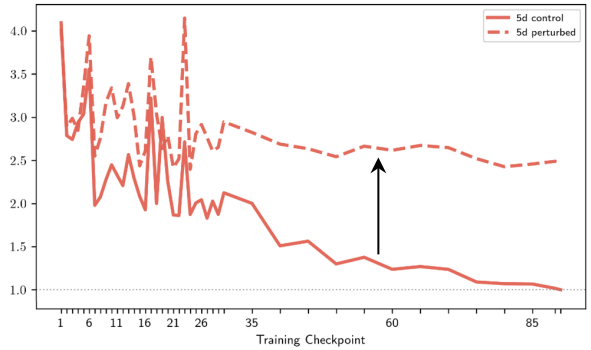
Rebecca Baiman et al. (2026)

Input perturbations reveal when the model learns cross-variable relationships

By replacing the winds in the inputs with average fields we identify **when the model learns physics!**



Storm Location IoU

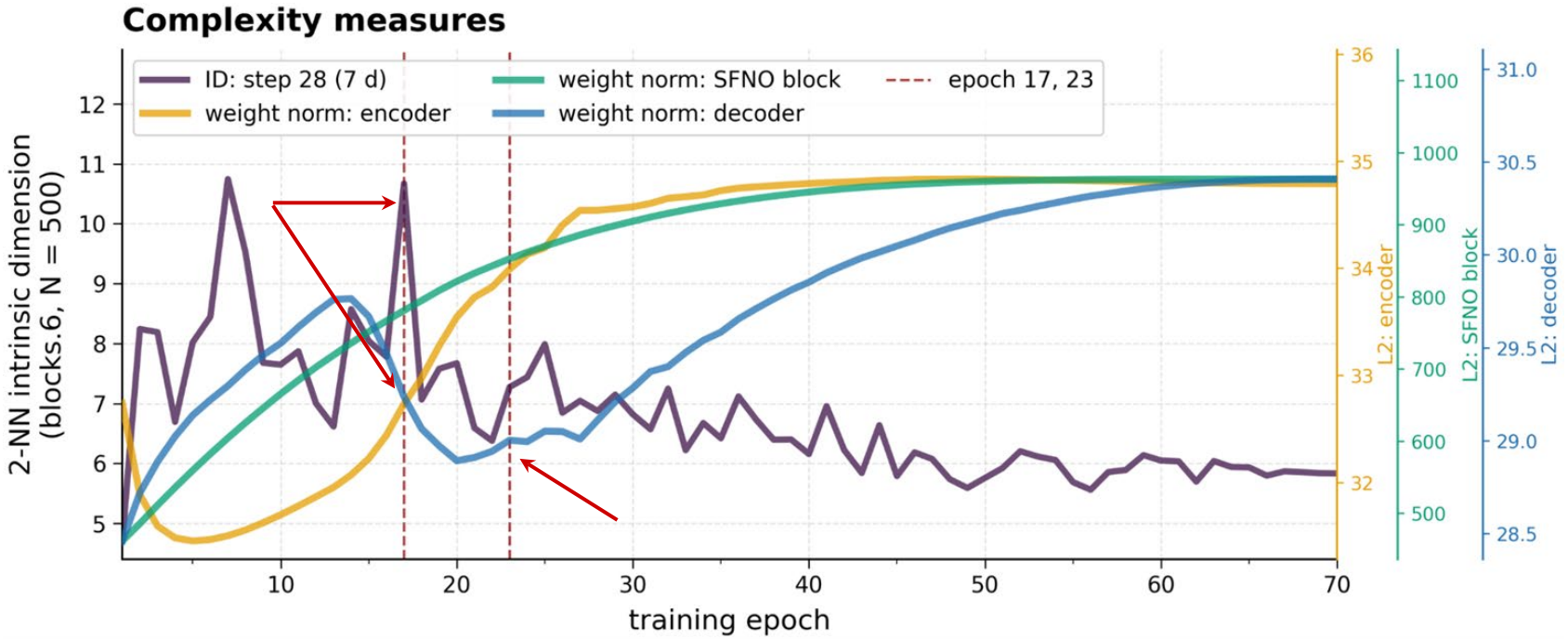


Storm Intensity Error



Annabel Wade et al. (in prep)

Complexity through training reveals a potential phase transition

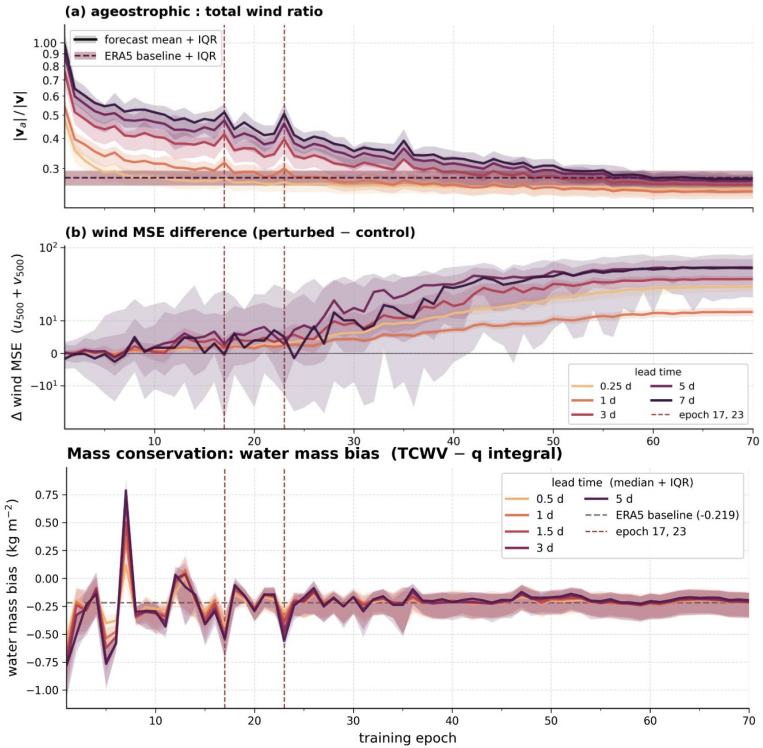




Annabel Wade et al. (in prep)

And this transition is timed with the learning of physics

- Geostrophic balance and mass conservation are learned following the potential phase transition about ⅓ of the way into training.

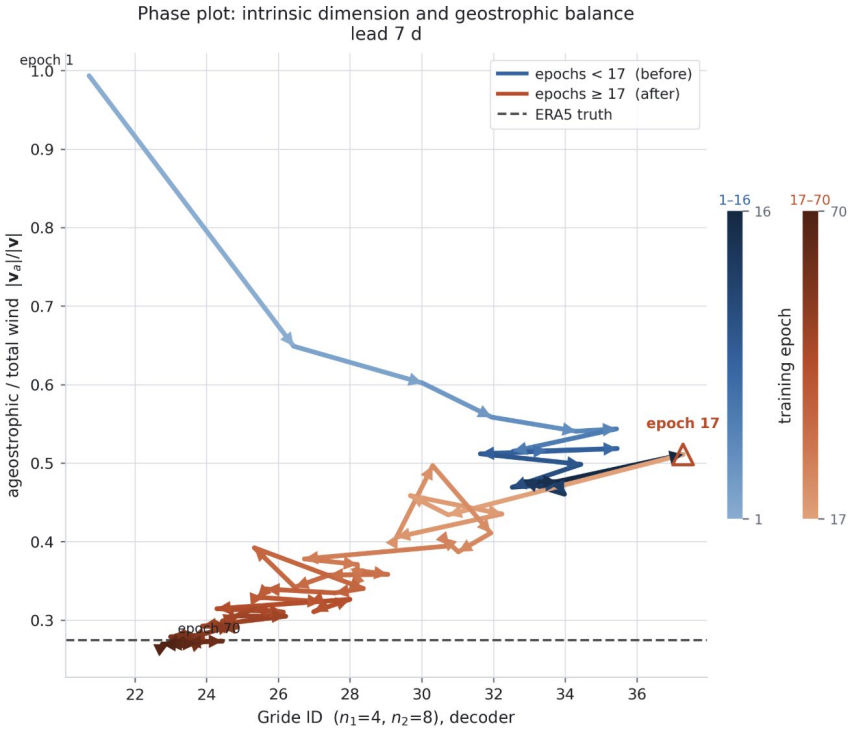




Annabel Wade et al. (in prep)

And this transition is timed with the learning of physics

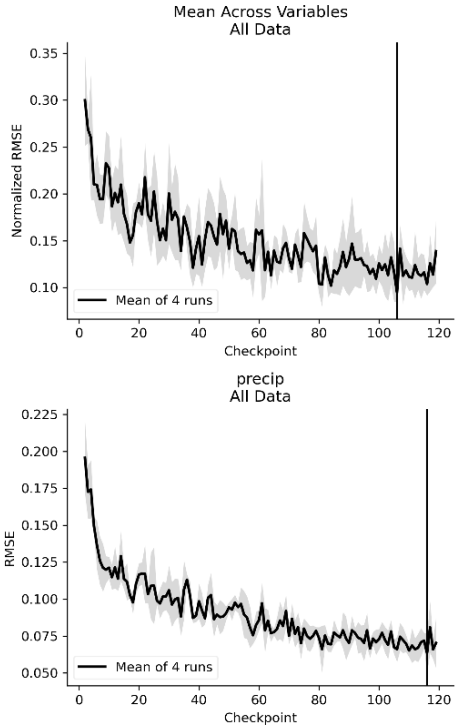
- Geostrophic balance and mass conservation are learned following the potential phase transition about ⅓ of the way into training.





Annabel Wade et al. (in prep)

We can gain insights on climate timescales too [ACE2]

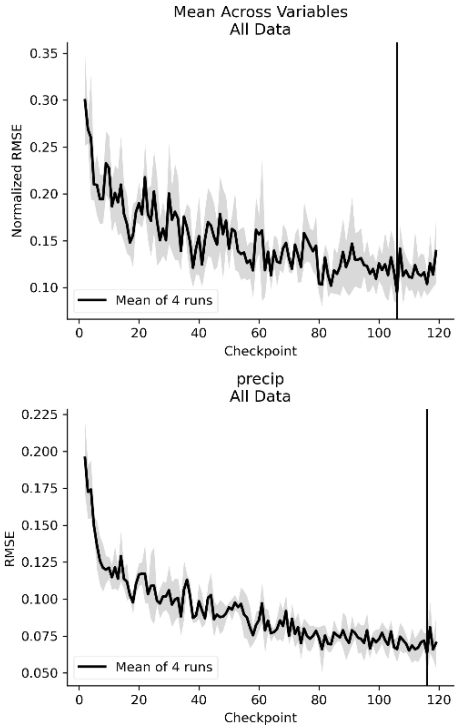


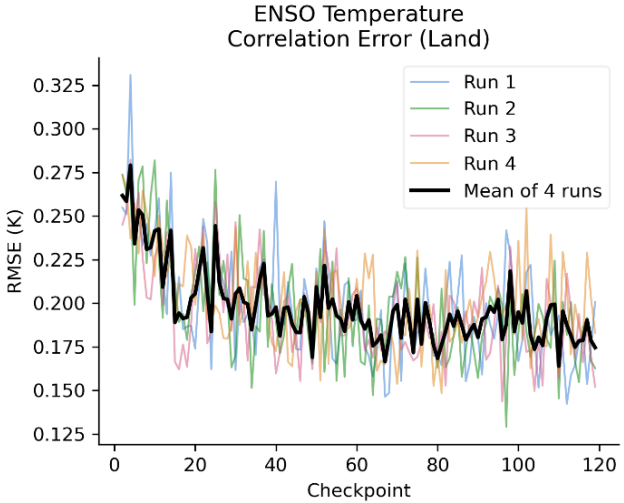
As expected, 6 - hr RMSE decreases with training over all variables…but what about climate phenomena?

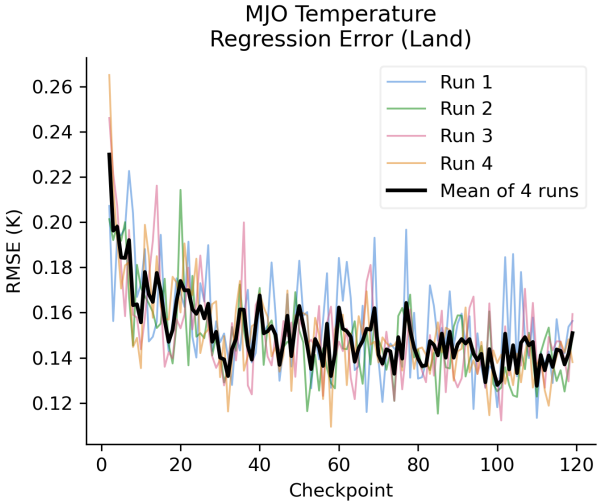


Landsberg et al. (in prep)

We can gain insights on climate timescales too [ACE2]





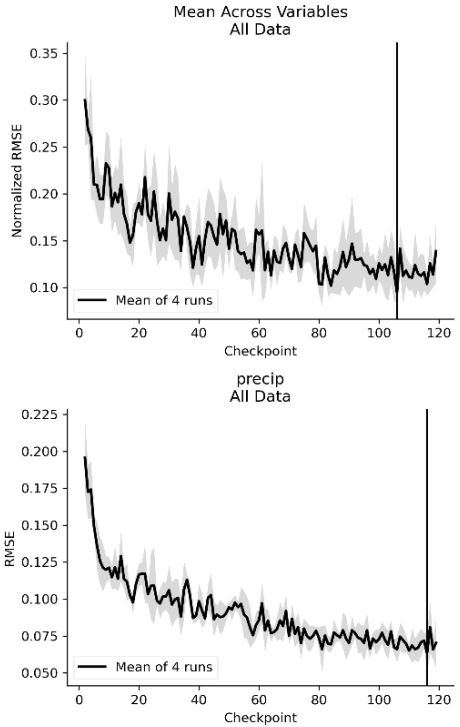


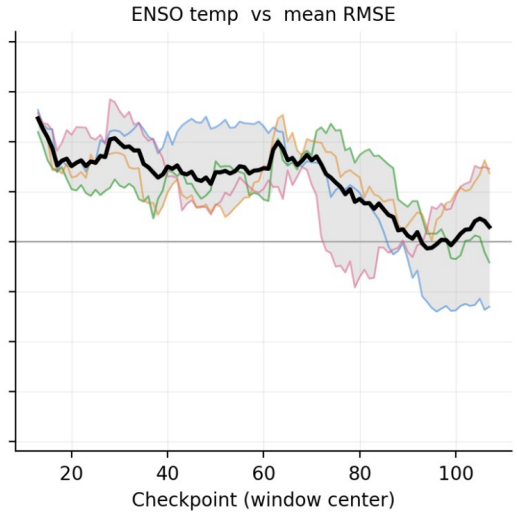
ENSO and MJO teleconnections to land temperature are learned much earlier in training, stabilizing ~¼ of the way in.



Landsberg et al. (in prep)

We can gain insights on climate timescales too [ACE2]





ENSO teleconnections are tightly correlated with model RMSE early in training

ENSO and MJO teleconnections to land temperature are learned much earlier in training, stabilizing ~¼ of the way in.



Landsberg et al. (in prep)

Training dynamics can reveal how AI models “think”. And when to intervene.

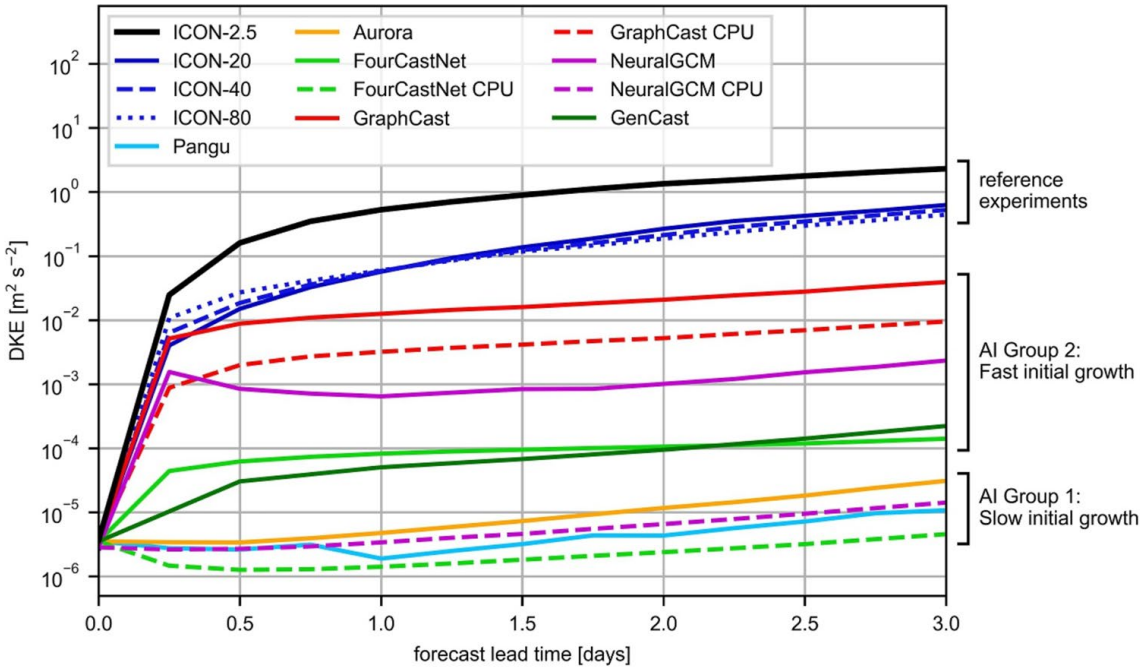
- Many extreme events → models learn, forget, and recover
- Multiple AI models →pattern holds across weather and climate
- Latent representations →what models encode changes throughout training
- Cross-variable relationships →models learn key physics at identifiable moments

Training dynamics can reveal how AI models “think”. And when to intervene.

- Many extreme events → models learn, forget, and recover
- Multiple AI models →pattern holds across weather and climate
- Latent representations →what models encode changes throughout training
- Cross-variable relationships →models learn key physics at identifiable moments

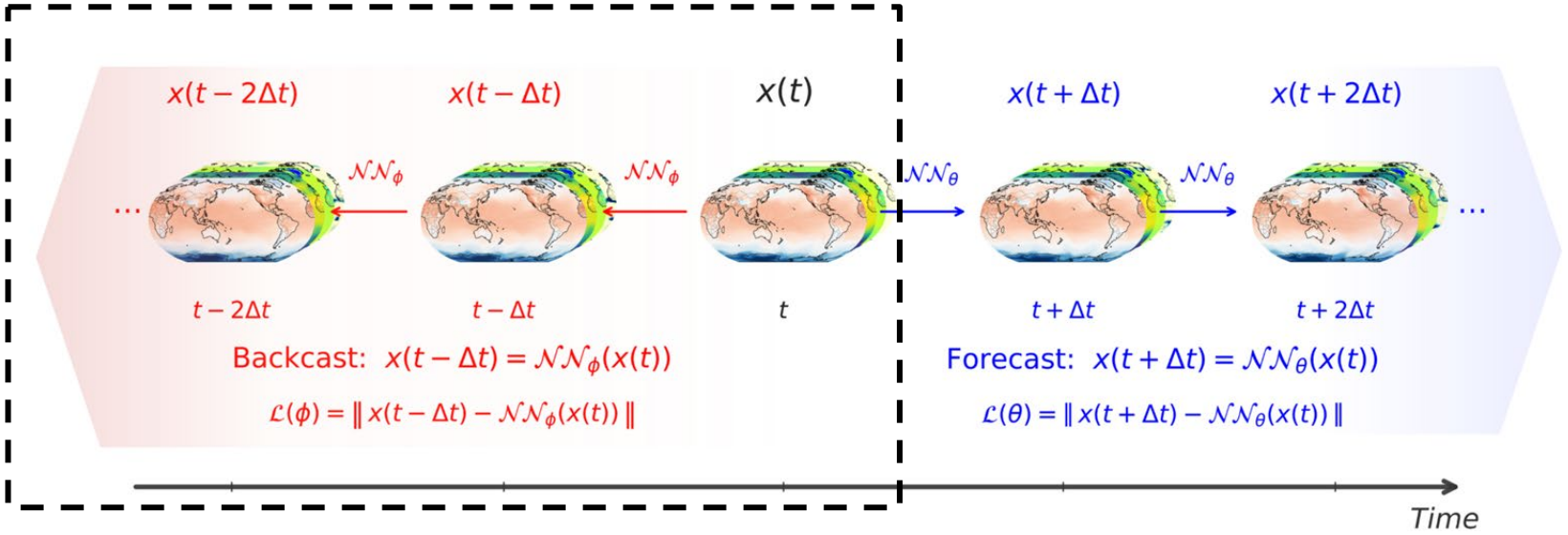
Watch → Locate → Intervene

These models are skillful, but lack fundamental physics



Selz et al. (2026)

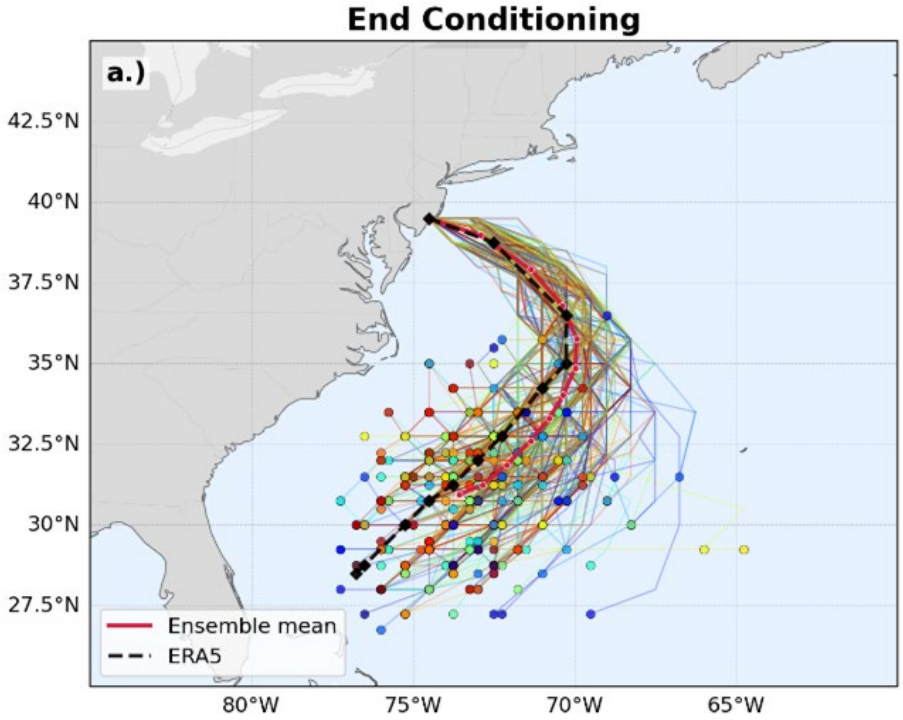
These models are skillful, but lack fundamental physics…allowing us to run them backward!



Hassanzadeh et al. (2026)

1000 ways to make landfall

1000 AI - generated atmospheric trajectories that all lead to Hurricane Sandy making landfall

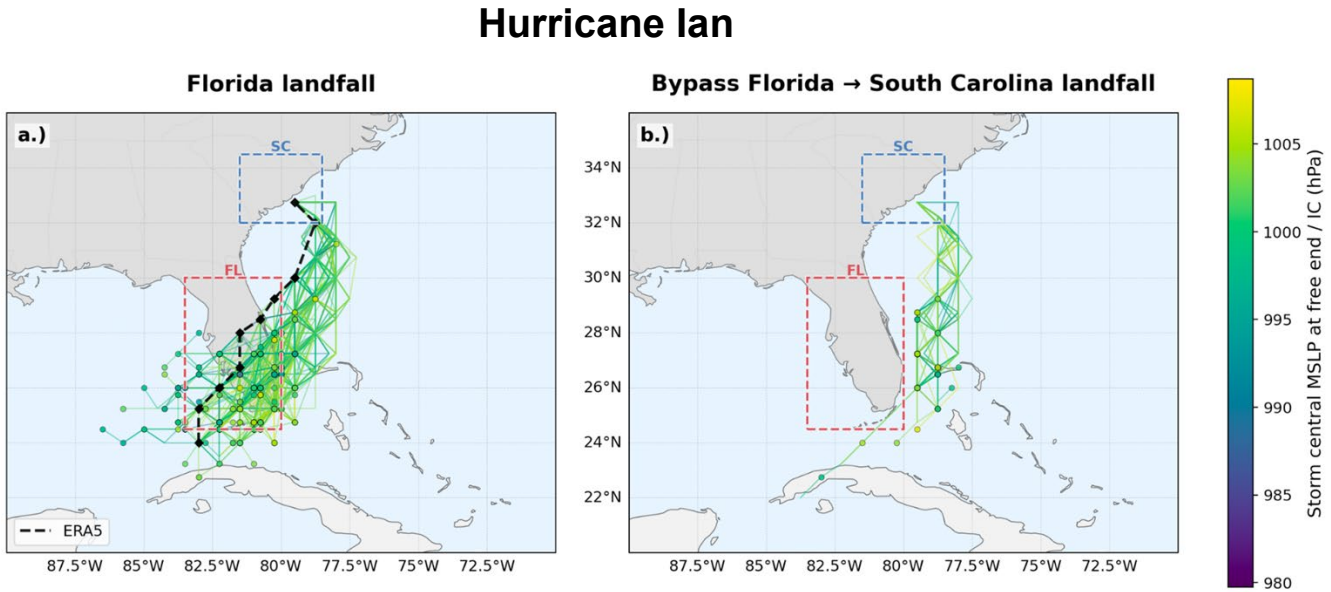




Lin et al. (2026)

Learning large-scale precursors to “gray-swans”

We are starting to explore the state space of precursors to the Rewind trajectories.





Jerry Lin et al. (2026)

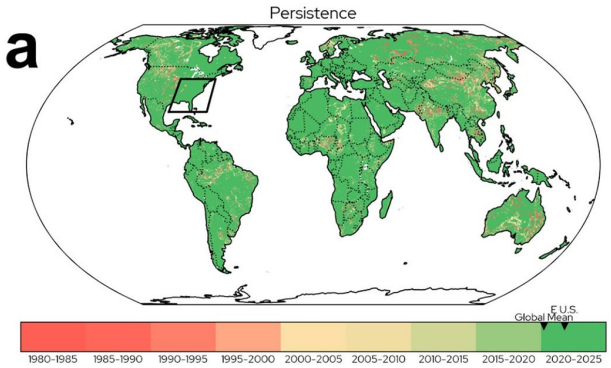
Bug or feature?

*"As coarse - graining is reduced, a butterfly - like effect emerges and forecast skill degrades...the more completely the training data represent the true dynamical system spatially and temporally, **the more the AI model behaves like the system, but the worse it forecasts the large scales,** which are often the desired target of prediction and most verification metrics."*

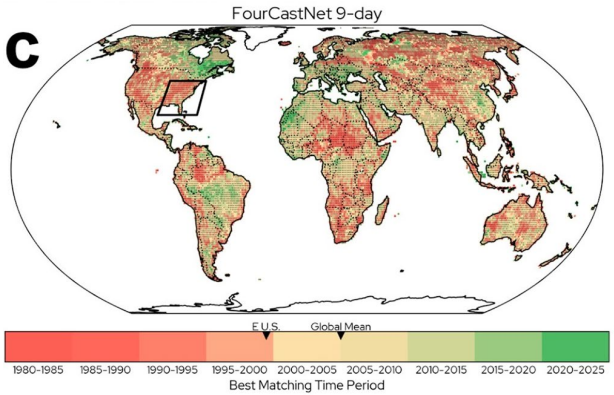
*"Our results prompt a question: **do we actually want AIWP models to fully behave like physics - based models even if restoring the physics costs performance, e.g., accuracy of practical forecasting.** "*

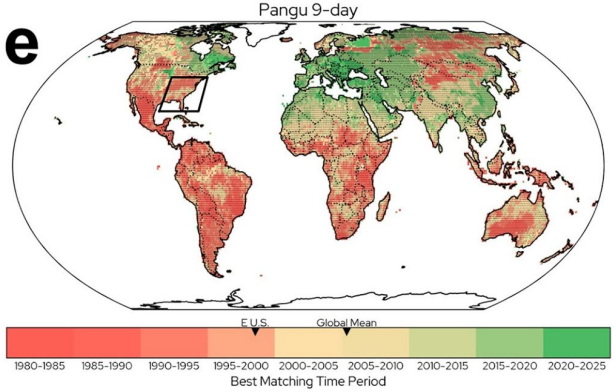
- *Hassanzadeh et al. (2026)*

AI models forecast the future with yesterday's climate



Colors show which historical period AI models drift toward when run forward 9 days.



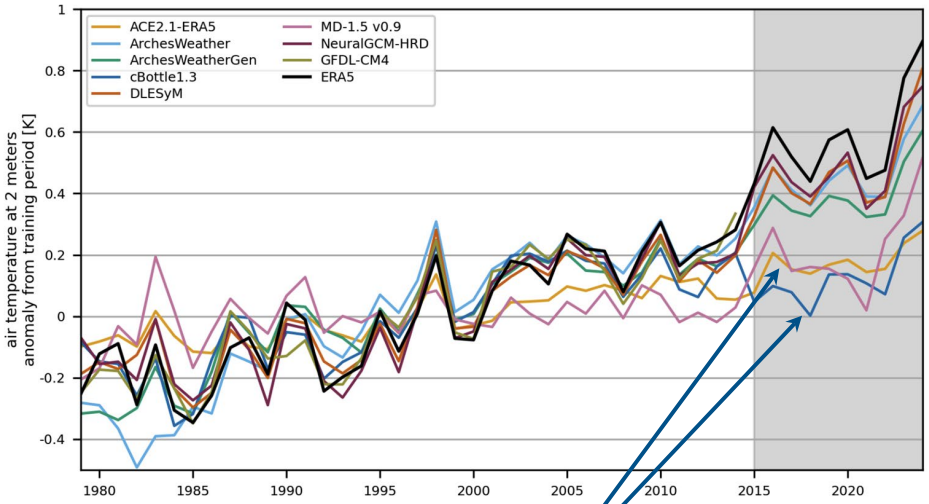




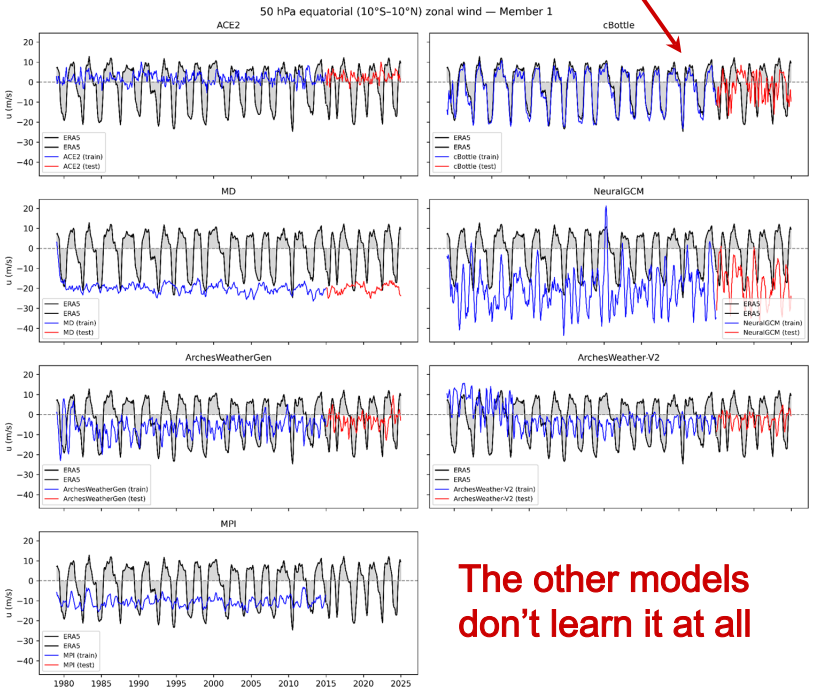
Jacob Landsberg & Barnes (2026)

Memorization & Colinearity Hazards

cBottle memorizes the QBO



Lack of explicit CO_2 in the input appears to reduce the historical trend



The other models don't learn it at all



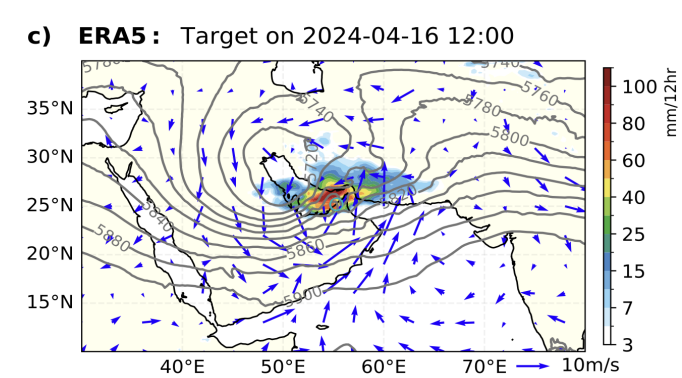
Jerry Lin et al. (2026)
Mu-Ting Chien et al. (in prep)

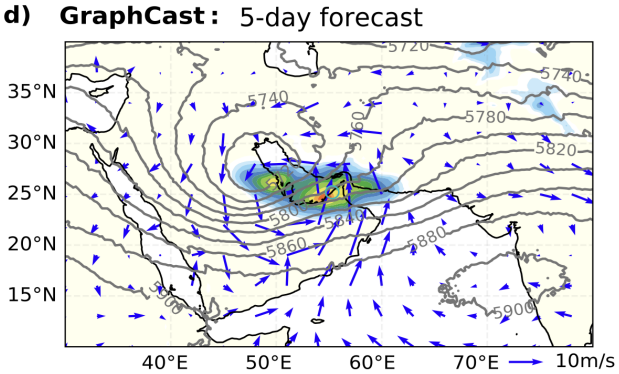
When they forecast an unseen extreme – how?

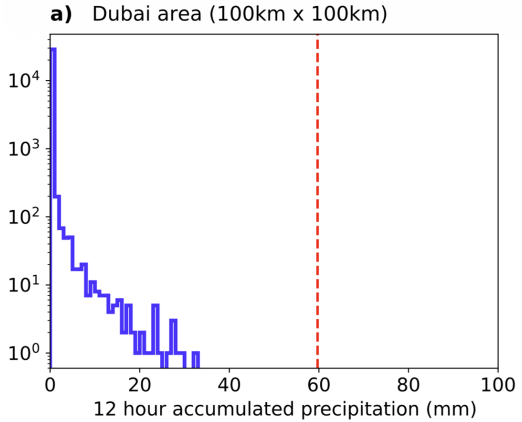
Extrapolation?

Translocation?

The physics itself?







Sun et al. (2026)

Tools I would love to have

1. **Training dynamics theory:** when and how are the physics learned? Are there optimal times we can disrupt the training process?

2. **Mechanistic interpretability:** where in the model are mechanisms and features learned? Can we force them to learn concepts in a way that we can later intervene on them? [e.g. MacMillan & Ouellette 2025]

3. **A handle on non - stationarity:** how can we track its impacts? Rigorously estimate uncertainty? Are there training samples from early in the record that disproportionately hurt (or help)?

This week, we are not asking “are the forecasts good enough” but **“what have they learned and how?”**.

What tools are we missing? Which can be adapted and which need to be developed?

What needs to be learned and when? Are there bugs we might turn into features?

Thank you.

Elizabeth A. Barnes, PhD
Dalton Family Chair in Environmental Data Science & Sustainability
Professor, Computing & Data Sciences
Professor, Earth & Environment
Boston University
eabarnes@bu.edu | www.barnes-research.com
https://github.com/eabarnes1010

BOSTON UNIVERSITY

