Machine Learning for predictability studies

Steffen Tietsche



ECMWF Training Course: Predictability and ensemble forecast systems

23 Oct 2026

What this lecture is (not) about

Not about:

- Demonstrating that we can use ML to make skilful NWP and S2S forecasts
- explainable AI (XAI) – build trust in ML forecasting by understanding how the model predicts

Rather it is about:

- Using AI/ML as a **tool** to better understand atmospheric / Earth System predictability

Predictability studies aim to
understand the sources and growth of forecast uncertainty and to
identify mechanisms and strategies for reducing forecast errors.
(credit: Guokun Dai)

Timeline of ML-based predictability research

Classical precursors on the left; the data-driven era begins in 2023 when global ML models became skilful enough to use as experimental systems.

	Precursors (≤2022)	2023	2024	2025	2026
Skilful ML models & ensembles	Ham+ 2019 ENSO CNN · IceNet 2021 · S2S AI Challenge 2022	Pangu, GraphCast, FourCastNet ≈ IFS HRES (Ben Bouallègue+ 2024)	NeuralGCM · AIFS-CRPS · FuXi-S2S	GenCast · AIFS-ENS operational · FGN · HENS 7,424 members · ACE2, Samudra, SamudrACE · ACE2 seasonal (Kent+)	AIFS-SUBS: weeks 2–6 at ~200× lower energy · AIFS v2 operational
Sensitivity & optimal perturbations	IFS adjoint and singular vectors		Vonich & Hakim: backprop-optimised ICs cut 10-day heatwave error by >90%	Baño-Medina+: backprop ≈ COAMPS adjoint · Vonich & Hakim: skill beyond 30 days · Winkler & Denhard: Arnoldi SVs for Pangu	Tian+: limits of ML tangent-linear and adjoint properties
Sources: XAI, regimes, causal discovery,	Kretschmer+ 2016/17 causal networks · Runge+ 2019 PCMCI · Nowack+ 2020 · Toms+ 2020 XAI · Mayer & Barnes 2021 forecasts of opportunity	Runge+ causal-inference review · Miloshevich+, Lopez-Gomez+ heatwave precursors	Shin+: moisture as primary MJO predictability source		Hassanzadeh+: ML backcasting · Spuler+: causal DAG-VAE · Pérez-Carrasquilla+: causal origins of extremes
Relaxation & counterfactuals	Jung+ 2010, 2014 · Hoskins+ 2012 · Magnusson 2017	Hakim & Masanam: imposed heating → Gill response in Pangu		Duan+ nudged NeuralGCM storylines · PGW attribution (Baño-Medina+) · Perkan+ relaxation + autodiff	Li+ (WCD): tropical relaxation in Pangu & NeuralGCM · Huang+: Green's functions · IFS nudged to AIFS (Polichtchouk+)
Limits & physical consistency	Identical-twin intrinsic-limit experiments in NWP	Selz & Craig: no butterfly effect in Pangu · Bonavita: effective resolution, balance	Rackow+: skilful in pre-industrial and +2.9 K climates · WeatherBench 2	Sun+: no gray-swan TCs beyond training · Brenowitz+ lagged-ensemble benchmark	Selz & Craig II: coarse-grained training data · Chen+: DKE spectra of AIFS-ENS, GenCast, FCN3 · Wu+: SSW propagation

Boiling it down to two basic approaches

1) “learned-sources” approach:

Modelling predictable signals directly
with only 10s/100s of trainable parameters

- Regimes / classifiers
- causal inference
- explainable predictors and forecasts of opportunity

Pros	Cons
Targeted, cheap, flexible	Very strong inductive bias
Direct and clear understanding	Risk of oversimplification

2) “forecast-experimentation” approach:

Use/adapt existing data-driven forecast models
with millions of trainable parameters

- initial-condition sensitivity
- constraint-response studies
- brute-force ensembling

Pros	Cons
Robustness and generalization	Increased cost and complexity
path towards forecast improvements	More difficult to interpret the outcomes

Question: Can you think of other pros and cons for the two approaches?

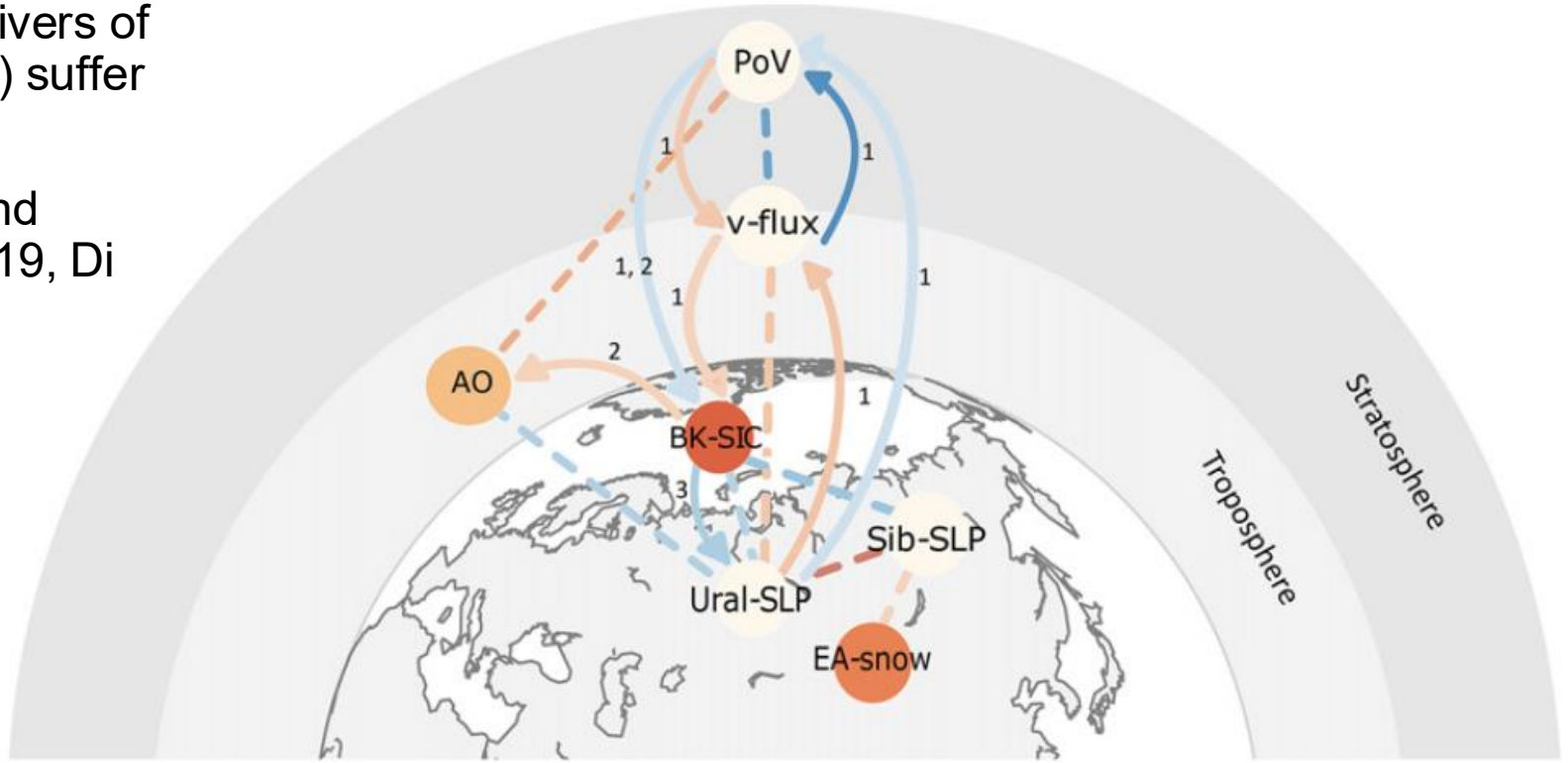
First approach: learned sources of predictability

This is a short section only – the lecture will focus on the second approach.

Causal discovery and inference

- Classical techniques to understand drivers of predictability (e.g. correlation analysis) suffer from confounding factors
- Causal inference techniques detect and scrutinize causal links (e.g. Runge 2019, Di Capua 2020)

Example on the right: detecting the influence of sea ice in the Barents and Kara Seas on Ural Blocking and cold Eurasian winters on seasonal time scales

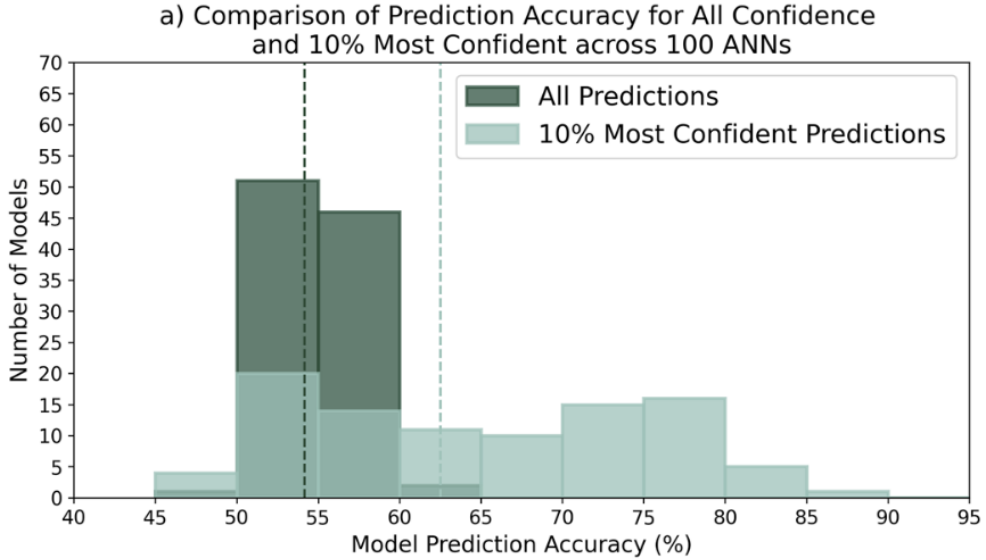


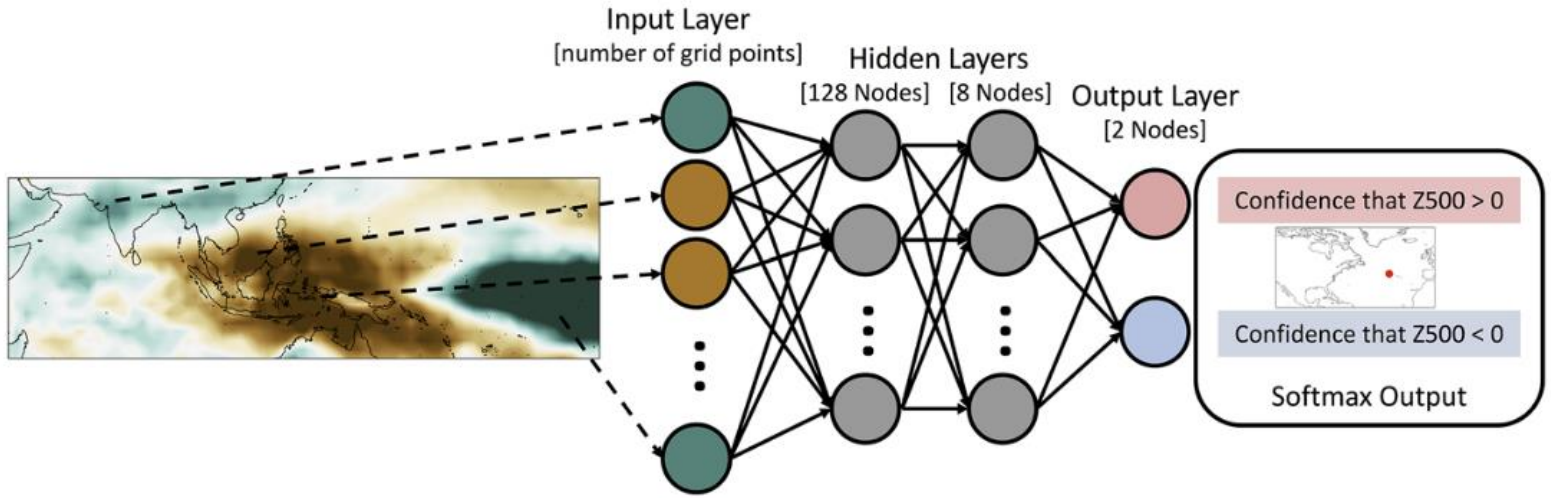
Kretschmer et al. 2016

Explainable predictors and forecasts of opportunity

- train small neural network (10s / 100s parameters) to extract drivers from data and make predictions
- can be combined with tools like layer-wise relevance propagation → “explainable AI” (not covered here)

Example shown: predicting the NAO from tropical OLR fields. Confirms importance of MJO and identifies ”forecasts of opportunity” with enhanced skill





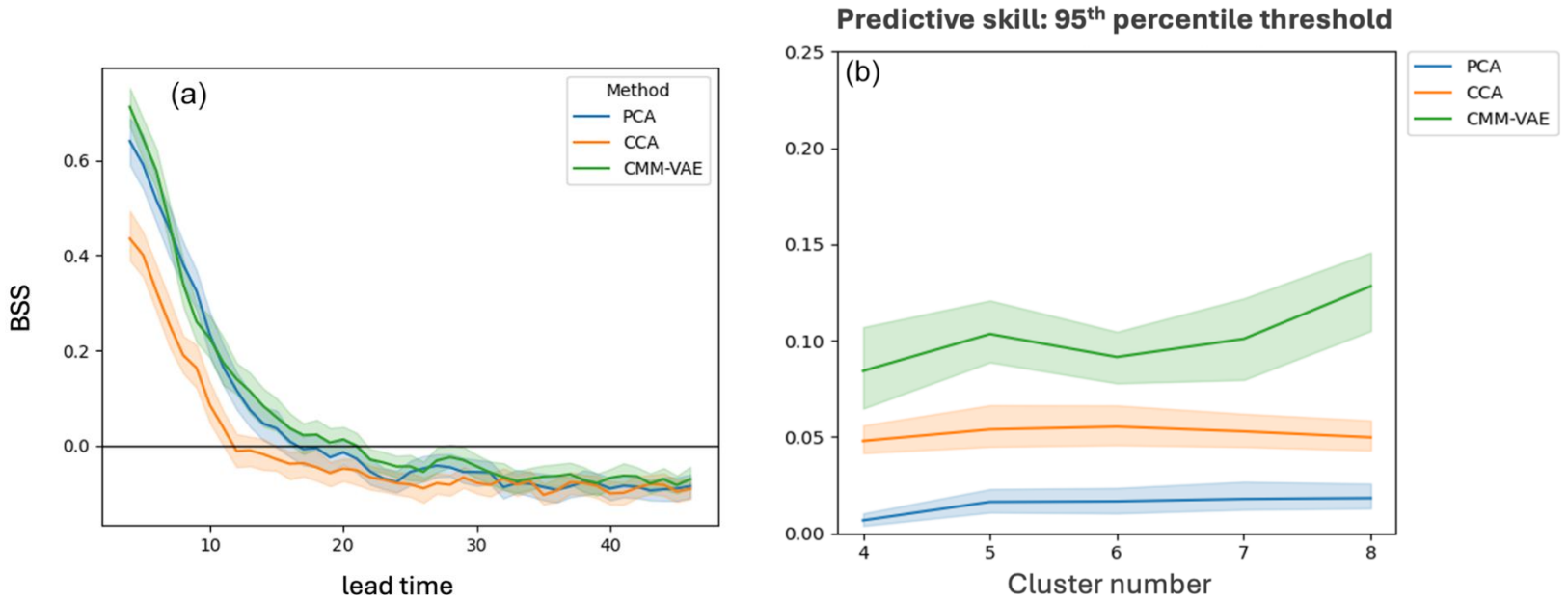
Mayer & Barnes 2021

Learning dynamical drivers of extreme events

- Choose event of interest and build small neural networks to improve on classical (often linear) approaches to understanding their precursors and drivers
- Interpretability through ablation and XAI techniques (e.g. LRP, backpropagation)

For example, Spuler et al. 2025: identification of predictable and informative circulation regimes in East Atlantic and Mediterranean that drive extreme precipitation over Marocco

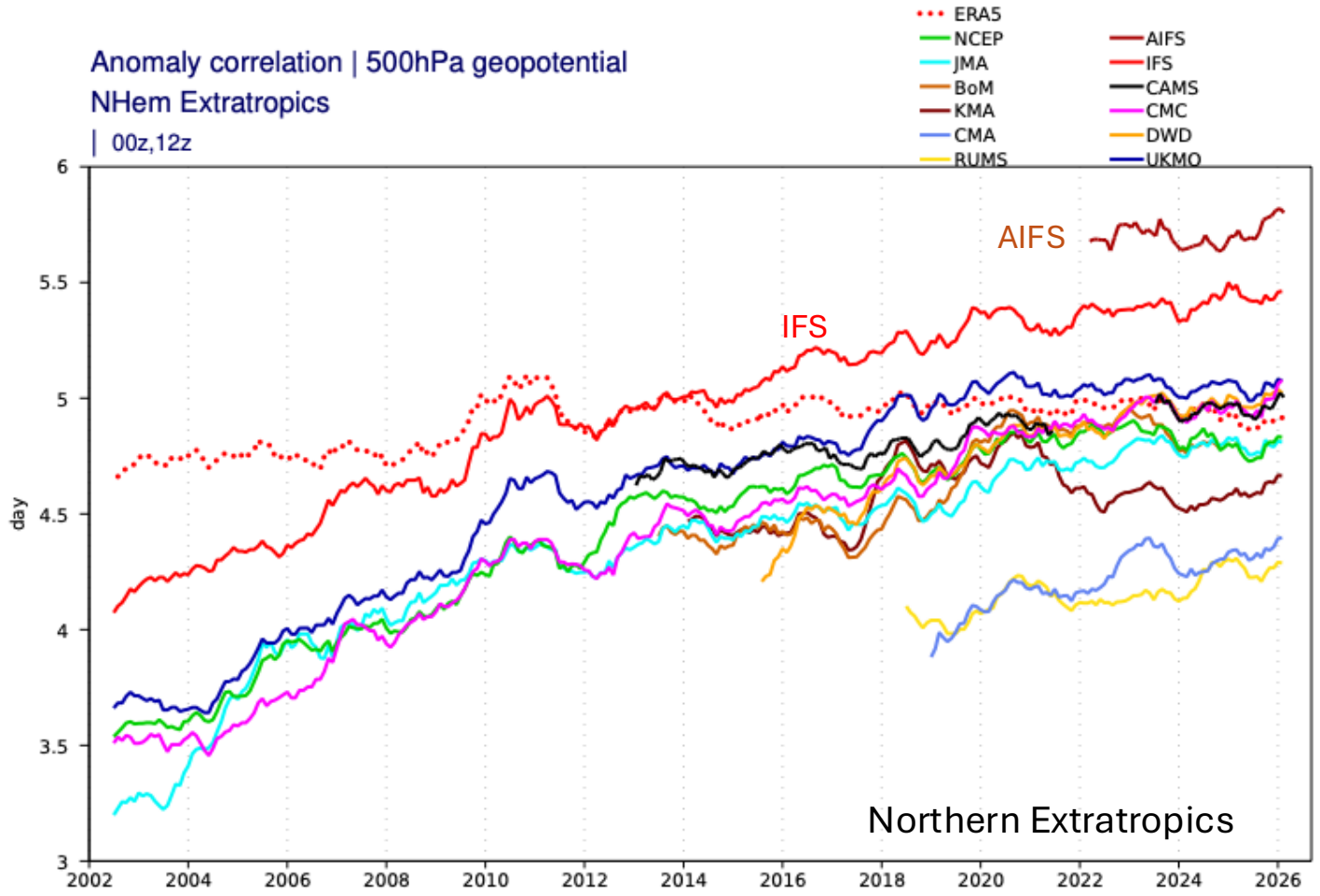
In comparison with classic circulation regimes (PCA, CCA), the NN finds regimes that are… → as predictable… → … but more informative for precip!



Second Approach:
ML forecast models as a laboratory

Predictable signals as emergent properties

Where we are in 2026: ML medium-range forecasts



Skill of operational weather forecasts:

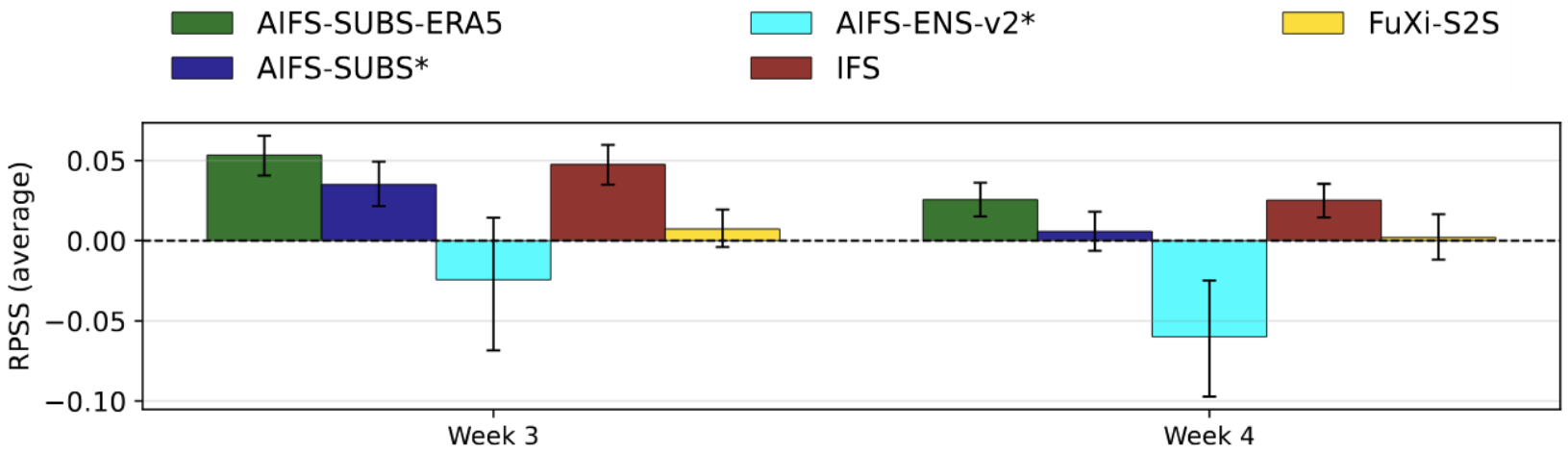
- ML (AIFS) clearly outperforms physics-based IFS – at a fraction of the cost!

Where we are in 2026: ML sub-seasonal forecasts

ML forecasts (AIFS) competitive with physical forecasts (AIFS)* - at a fraction of the cost

*Side node: ML postprocessing enables subseasonal forecasts that are clearly better than any physical model.

Skill averaged over several surface variables from the AI Weather Quest
https://aiweatherquest.ecmwf.int/



Schloer 2026

Data-driven forecast models as a laboratory for predictability research

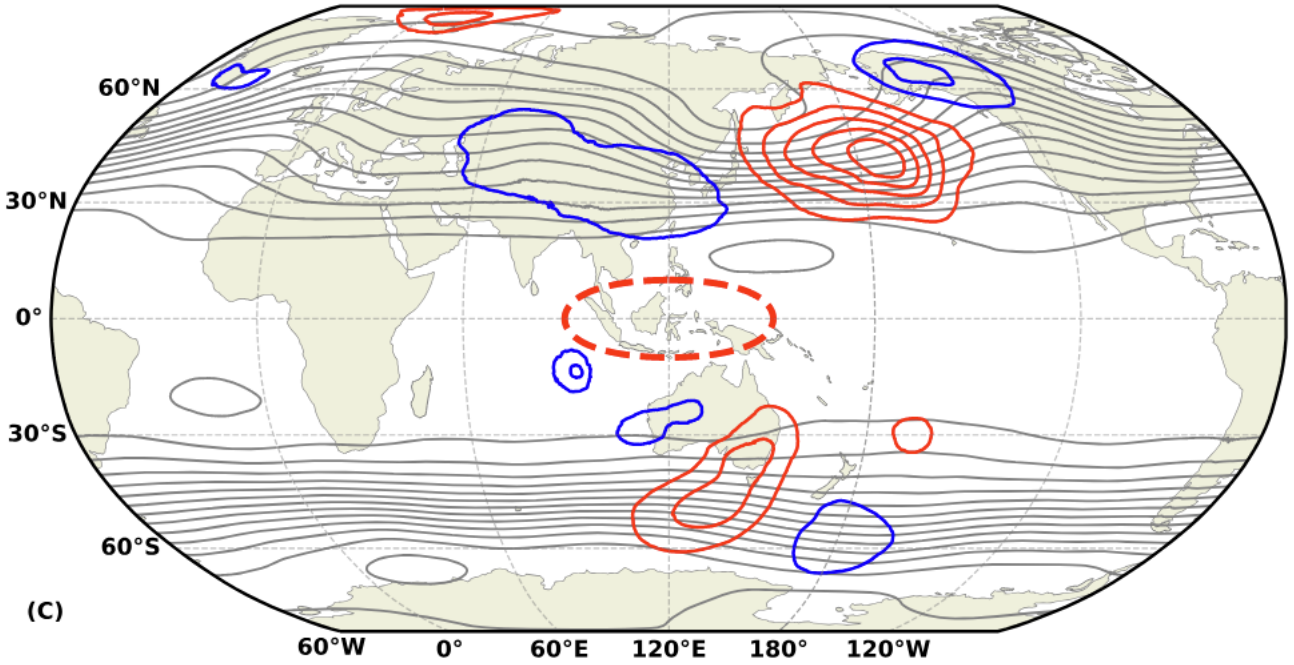


FIG. 1. Response in DJF 500-hPa geopotential height to steady tropical heating of 0.1 K day^{-1} within the region outlined by the dashed red line. The DJF-averaged geopotential height is shown by gray lines every 60 m, and anomalies are shown by red (positive) and blue (negative) lines; the zero contour is suppressed. Solutions are shown for (a) 5 days (contours every 0.3 m); (b) 10 days (contours every 2 m); and (c) 20 days (contours every 20 m).

Hakim & Masanam (2024): response of Pangu-Weather (first-generation MLWP) to tropical heating over Maritime Continent

→ text-book response of extratropical wave train

Idealized study is far outside the distribution of training data

→ ML models can encode enough atmospheric physics to serve as a laboratory

What makes MLWP models a great tool for predictability studies?

1. **SKILFUL**
   - NWP and S2S forecasts of comparable or better quality than with physics-based models
   - good generalization properties (e.g. for out-of-distribution initialization)

2. **CHEAP**
   - orders of magnitude less compute to produce a forecast (e.g. AIFS-SUBS: 200x less)
   - easy to set up and run

3. **DIFFERENTIABLE**
   - nonlinear adjoint is part of the model (needed for backpropagation of errors during training)
   - straightforward to compute sensitivity of forecast to any quantity of interest (e.g. initial conditions)

4. **OPEN**
   - MLWP model weights are usually freely available (e.g. AIFS)
   - underlying source code and training configuration often freely available

Data-driven forecast models as a laboratory for predictability research

Data-driven models are skilful, cheap, differentiable and open. That turns classic predictability questions (how fast do errors grow, where does skill come from, what are the limits) into experiments you can run on a single GPU.

4 instruments to use:

Sensitivity for free

Backpropagation as a nonlinear adjoint

→ see lecture on initial-condition uncertainty

Ensembles at scale

Thousands of members or cases

→ see lectures on model uncertainty and verification

Sources of predictability

Explainable predictors, causal discovery, backcasting?

→ see lectures on coupled variability

Relaxation & counterfactuals

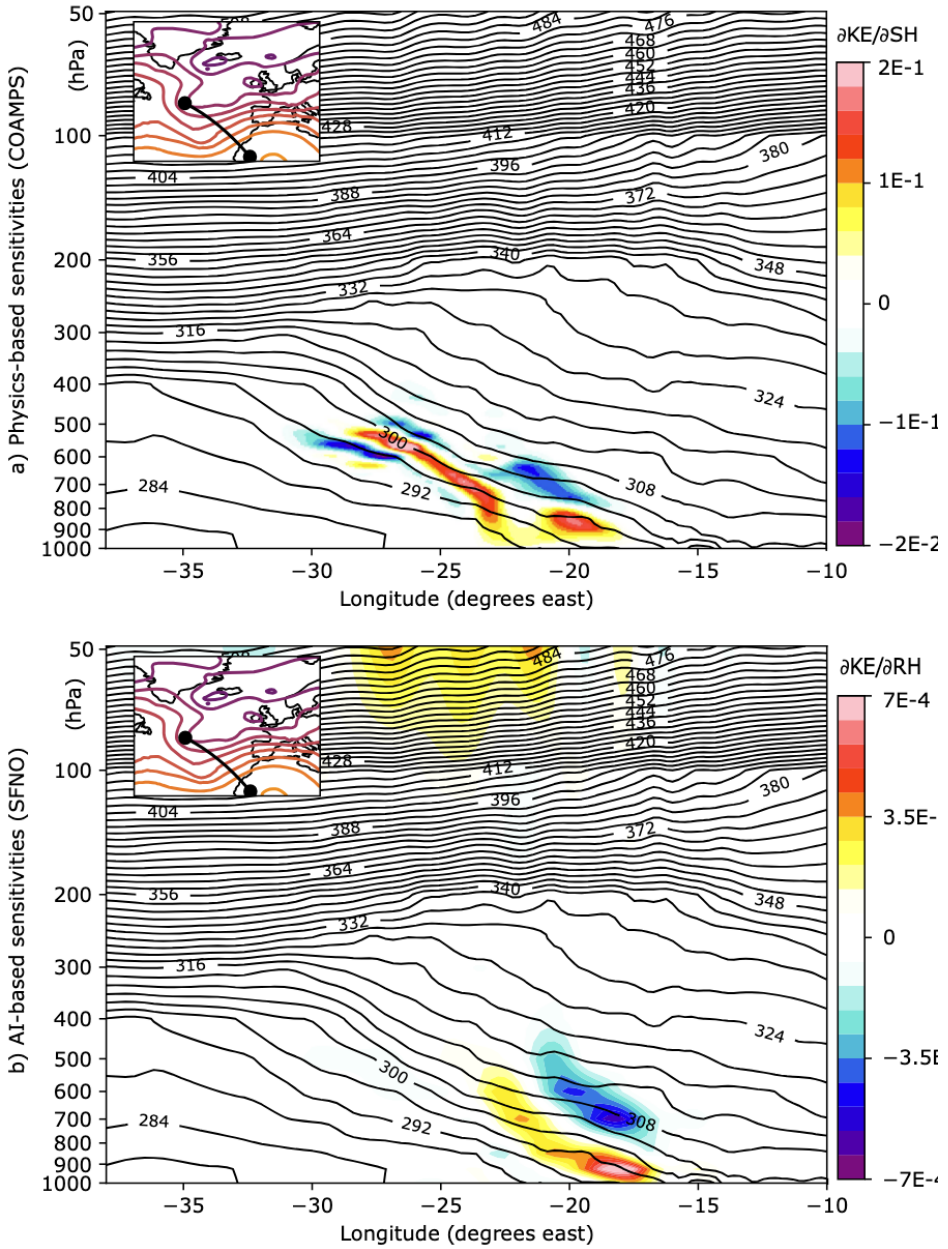
Constrain one region, read the remote response

→ cp. IFS relaxation studies

Some limitations: Current ML models trained on reanalysis do not reproduce the butterfly effect and can have issues with physical and cross-variable consistency. Their capability to extrapolate and generalize is astonishing but has limits.

Initial-state sensitivity

- NWP: tangent-linear and adjoint models, valid for ~2–3 days, expensive to maintain
- ML: gradient of any metric w.r.t. the initial state from one backward pass, at any lead time, for the full nonlinear model
- Validating example (right): ML backprop sensitivities ≈ COAMPS adjoint for cyclone Xynthia (Baño-Medina et al. 2025)
- Caveat: ML sensitivities can be noisy and unphysical in places (Tian et al. 2026)



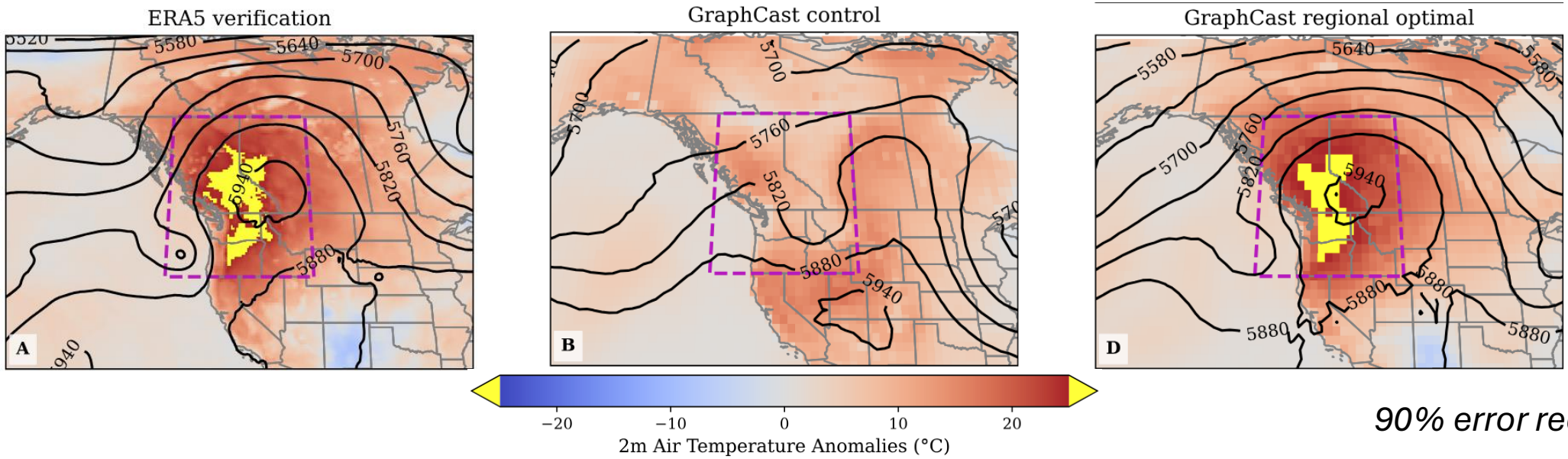
vertical cross-section of sensitivity in initial relative humidity for cyclone Xynthia (Baño-Medina 2025)

Optimal initial conditions

- Vonich & Hakim 2024: predictability limit of the 2021 Pacific Northwest Heatwave (30K anomaly!)
- Exploit differentiability of ML forecast model to find optimal initial condition for predicting this extreme

1. Produce a forecast at a chosen lead time, initializing from ERA5 initially, then from optimisation iterations
2. Compute the loss function (= forecast error)
3. Calculate gradient of the loss w.r.t. to inputs
4. Update inputs using the loss gradient
5. Repeat (1–4) until the loss plateaus or 100 iterations is reached. An example of the loss as a function of iteration

500 hPa geopotential height and 2m air temperature at 0 UTC on 30 June 2021 in reanalysis (left), a 10 day forecast from reanalysis (middle), and from optimized initial conditions (right)

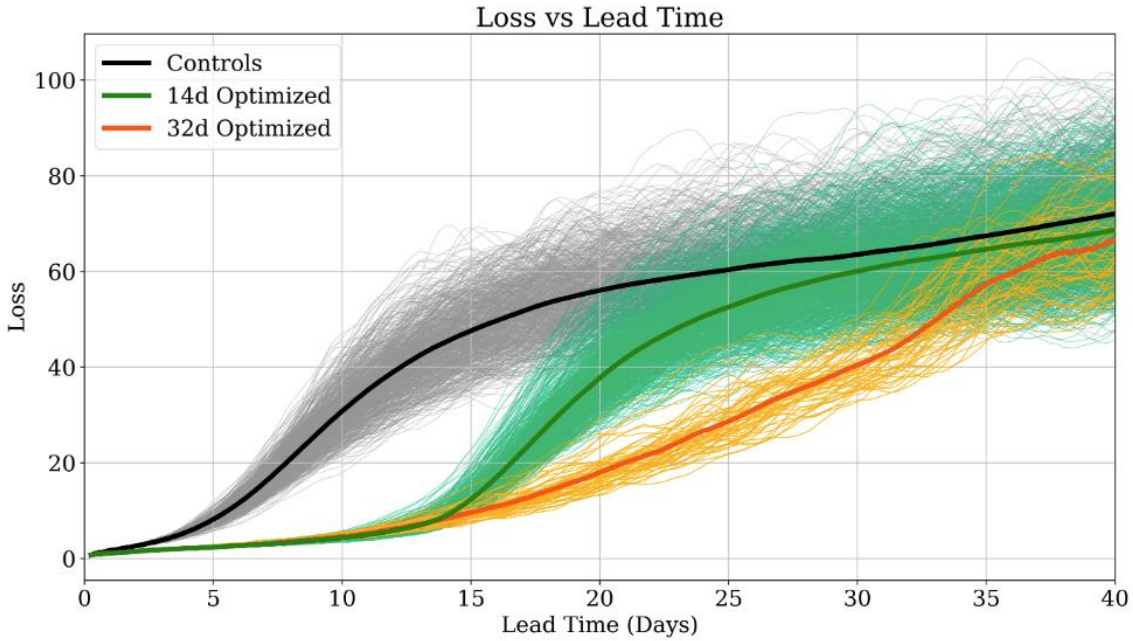


90% error reduction!

Optimal initial conditions cont.

Vonitch & Hakim 2026:
initial-condition optimisation always works.

→ proves the existence of initial conditions that considerably extend the current established limits of predictability.



global MSE for 732 twice-daily forecasts during 2020

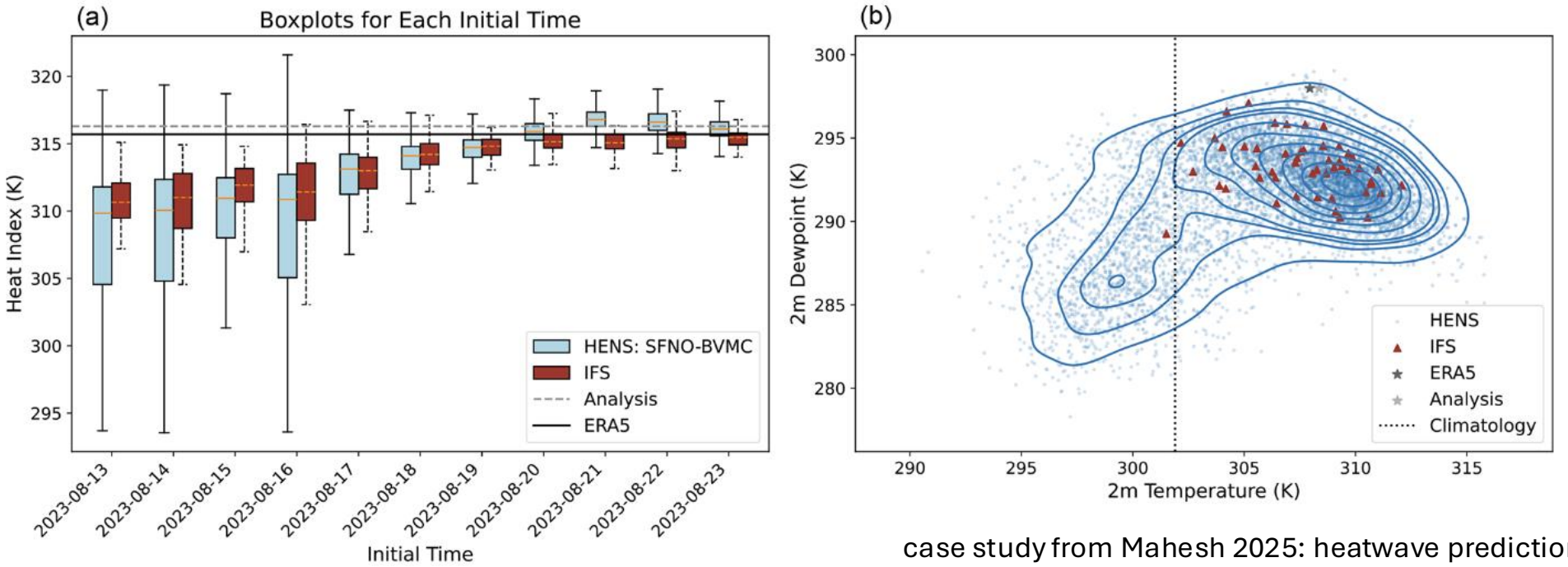
Question: skill beyond 30 days with optimised initial conditions — what do you think this means? A statement about practical predictability, intrinsic predictability, or about the model?

Pushing ensembles to the extreme

Computing cost for a 6-week forecast:
e.g. 1.5 hours on 256 CPUs for IFS (physical) *versus* <1 minute on 1 GPU with AIFS (ML)

One way to exploit this: huge ensembles to study precursors, drivers and statistics of extreme events

→ e.g. Mahesh 2025: weather forecasting with 7000 (!) members

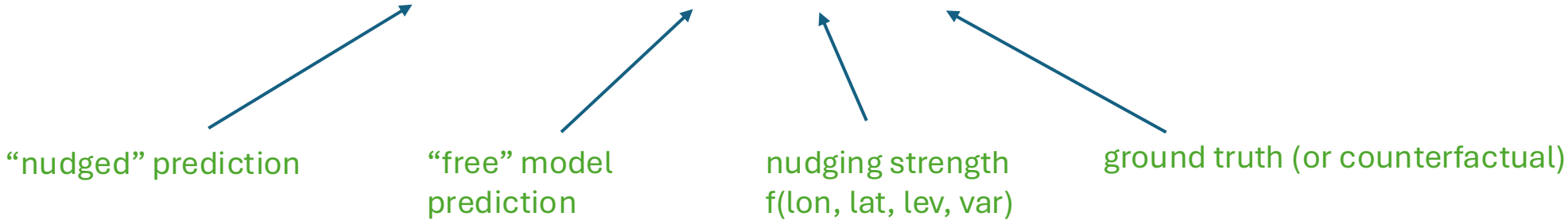


case study from Mahesh 2025: heatwave prediction for Kansas

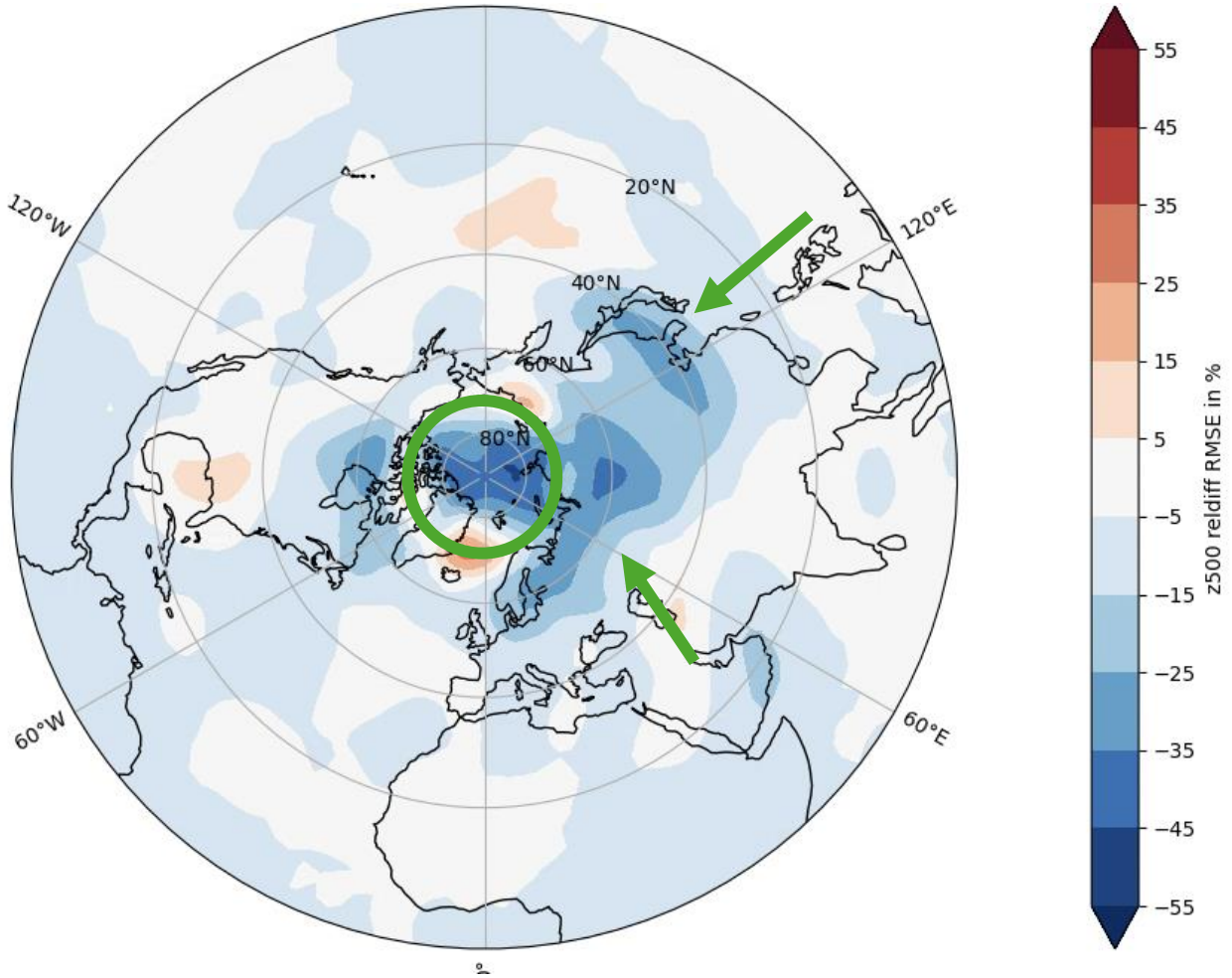
Relaxation experiments (constrain-and-response)

- Also known as "*nudging*" experiments – widely used to study teleconnections in physical modelling (e.g. Jung, GRL 2014; Hitchcock, GMD 2022)
- Idea: during a forecast, constrain *a subset* of regions/levels/variables and study impact on the unconstrained parts of the model

$$x_{t+1} = (1 - G) \cdot M(x_t) + G \cdot O_{t+1}$$



Example: Arctic-midlatitude linkages



Constrain winds and temperature north of 70N in AIFS to ground truth (ERA5)

- response in mid-latitudes: reduced forecast errors
- direct evidence of "information transport" from one region to another (aka teleconnection)

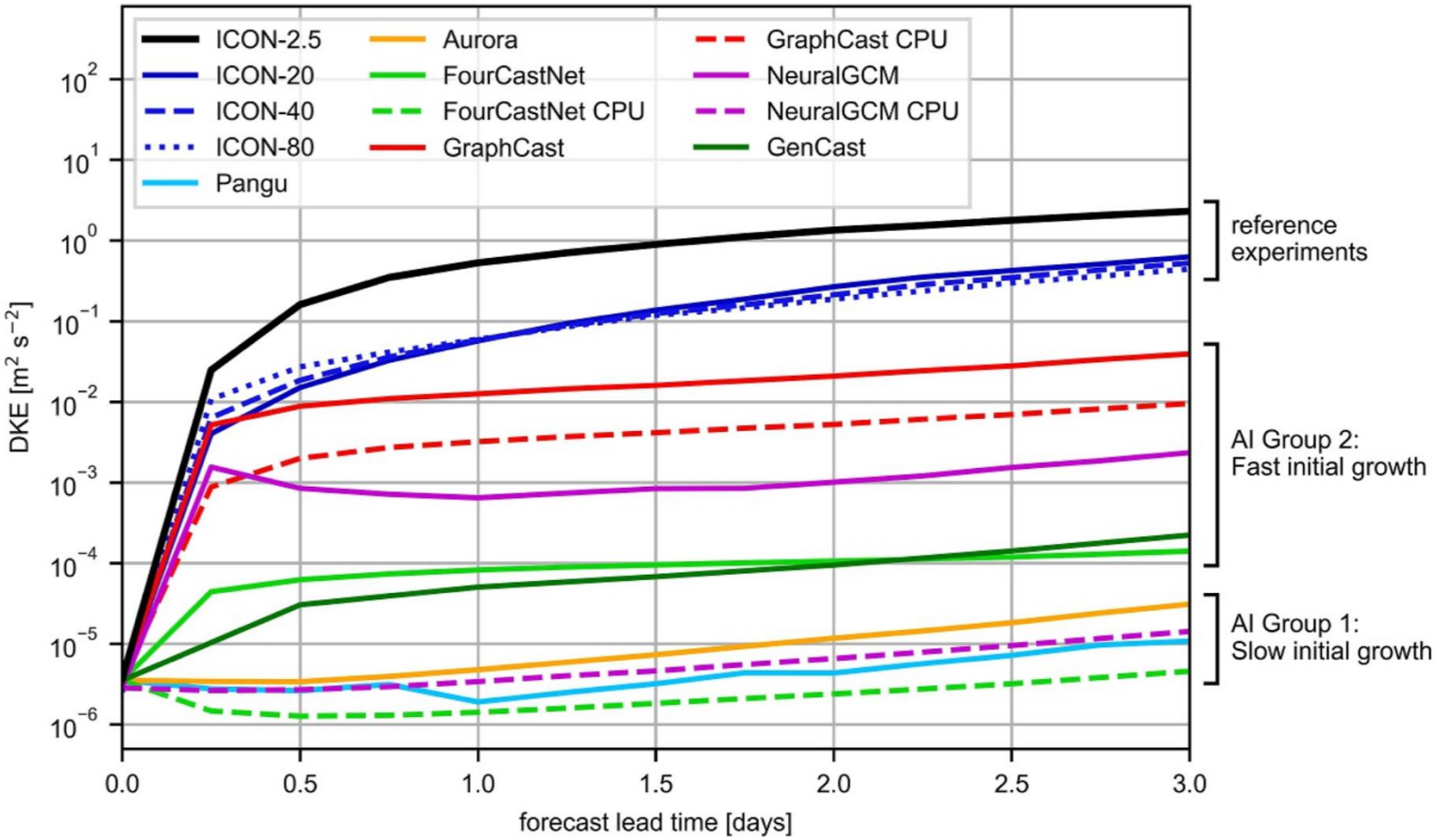
relative forecast improvement of Z500 in week 3 for AIFS reforecasts 2018-2022

Health warning!

Limitations of ML models need to be checked before using them for predictability studies.

Example 1: ML models do not simulate the butterfly effect

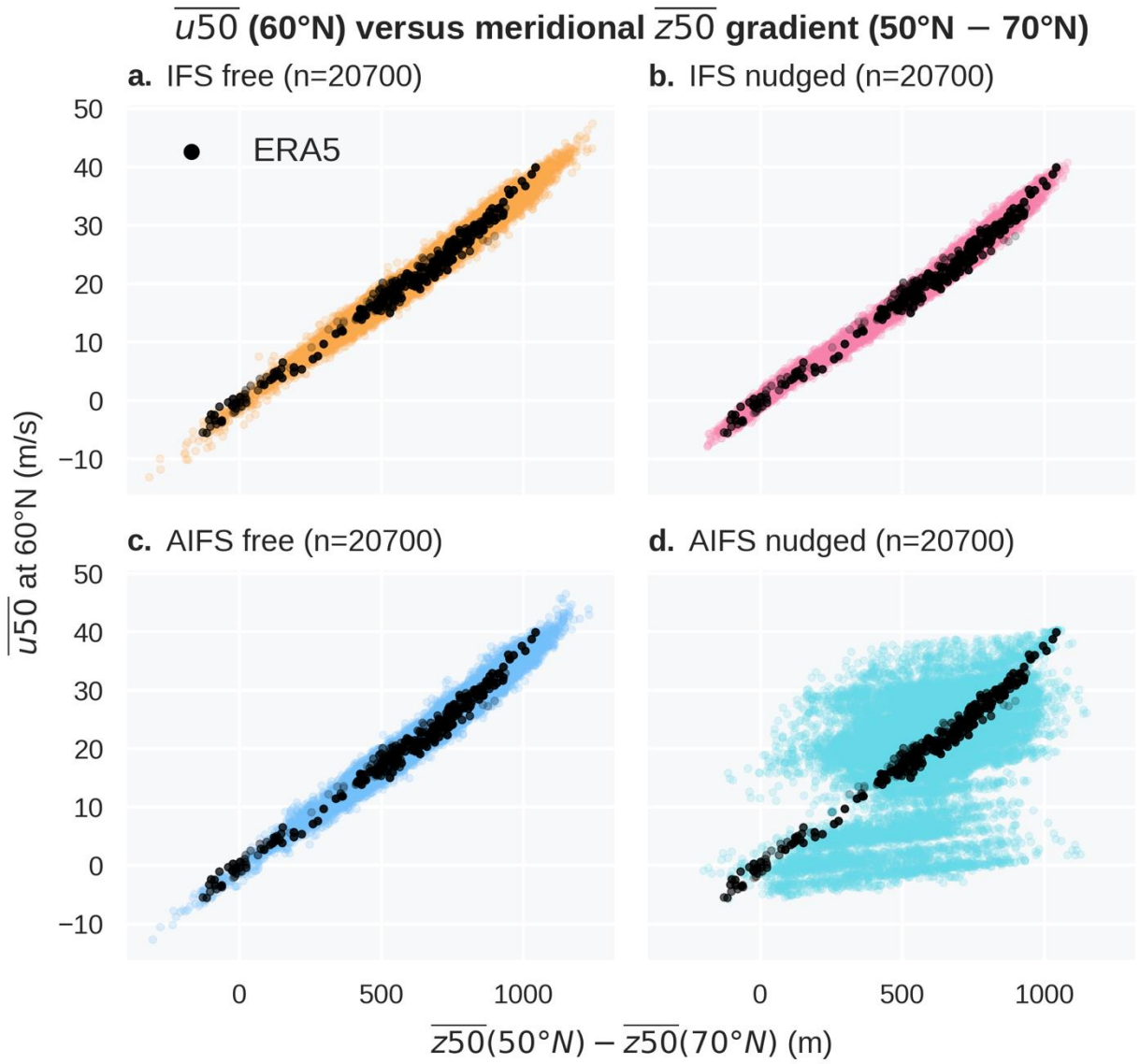
- Inability to simulate error growth correctly for *very small* perturbations
- Not (yet) relevant for NWP, as initial condition uncertainty is still much higher
- But makes models ill-suited to answer questions e.g. on intrinsic limits of predictability



Selz & Craig 2026

Example 2: losing geostrophic balance in relaxation experiments

- Experiment: constrain stratospheric temperature and wind in AIFS-SUBS
- At 50 hPa (and below), winds and geopotential are in geostrophic balance in ERA5, IFS free and nudged, and AIFS – but not in AIFS nudged
- Caveat when using model to study downward propagation of stratospheric signals
- Physical consistency could be improved → feedback for model development



Summary

- In the last ~5 years, ML is increasingly used as a tool for advancing understanding of atmospheric predictability
- Broadly, two different approaches are possible:
  1. modelling sources of predictability directly with small, targeted ML models
  2. Using established ML weather forecast models as a laboratory
- This lecture focussed on the latter – very promising because of 4 properties of these models:
  1. They produce forecasts which are at least as **skilful** as those from physics-based models
  2. They are **cheap** to run and easy to set up
  3. They are **differentiable** by construction
  4. They come with **open** weights – even source code and training configuration is often freely available
- Care is needed in interpreting results: ML forecast models have limitations and are not a drop-in replacement for physical models

Take-away

ML forecast models are cheap, skilful, differentiable and open. That turns classic predictability questions (how fast do errors grow, where does skill come from, what are the limits) into experiments you can run on a single GPU.

Use them, and understand what they can (and cannot yet) tell you.

For discussion

Sensitivity

Optimised initial conditions give GraphCast skill beyond 30 days. What exactly has been shown: a limit of the atmosphere, of our analyses, or of the model? What experiment would separate the three?

Ensembles

Could you work with 7,000 members in your research or operational activities? Can you think of new kind of research questions or forecast products that huge ensembles would make possible?

Sources

A network tells you it is confident about week 3 over the North Atlantic. What evidence would convince you that the precursor it points to is causal rather than merely correlated?

Relaxation

You relax the tropics to ERA5 in an ML model and the extratropical response is weaker than in the IFS. Which is wrong: the model's teleconnection, the experiment, or the IFS? How would you find out?

Limitations

Current ML models do not reproduce the butterfly effect yet forecast better than the IFS. Is that a bug or a feature? What would you change in training data or loss to recover it, and what would it cost?

Your research

Take one predictability question from your own work. Which of the four instruments could you apply to it next month, with the tools on the handout? What would stop you?

References (1/2)

[1] **Hakim, G. J. & Masanam, S. (2024).** Dynamical tests of a deep learning weather prediction model. Artif. Intell. Earth Syst. 3, e230090. doi:10.1175/AIES-D-23-0090.1

[2] **Lang, S., Alexe, M., Clare, M. C. A., et al. (2024).** AIFS-CRPS: Ensemble forecasting using a model trained with a loss function based on the Continuous Ranked Probability Score. arXiv:2412.15832. Journal version reported 2026 [verify].

[3] **Lang, S. & Magnusson, L. (2025).** AIFS ENS becomes operational. ECMWF Newsletter 185, 20–24. PDF. See also Alexe et al., Newsletter 181, "Data-driven ensemble forecasting with the AIFS".

[4] **Price, I., Sanchez-Gonzalez, A., Alet, F., et al. (2025).** Probabilistic weather forecasting with machine learning. Nature 637, 84–90. doi:10.1038/s41586-024-08252-9

[5] **Kochkov, D., Yuval, J., Langmore, I., et al. (2024).** Neural general circulation models for weather and climate. Nature 632, 1060–1066. doi:10.1038/s41586-024-07744-y

[6] **Schloer, J., Tietsche, S., Roberts, C. D., Zampieri, L., Lang, S., Mertes, G., Jones, G., Chantry, M. & Vitart, F. (2026).** AIFS-SUBS: Extending data-driven forecasting to sub-seasonal timescales. arXiv:2607.05100

[7] **Chen, L., Zhong, X., Li, H., et al. (2024).** A machine learning model that outperforms conventional global subseasonal forecast models. Nat. Commun. 15, 6425. doi:10.1038/s41467-024-50714-1

[8] **Kent, C., Scaife, A. A., Dunstone, N. J., Smith, D., et al. (2025).** Skilful global seasonal predictions from a machine learning weather model trained on reanalysis data. npj Clim. Atmos. Sci. doi:10.1038/s41612-025-01198-3 [verify DOI]

[9] **Mahesh, A., Collins, W., Bonev, B., et al. (2025).** Huge ensembles Part I: design of ensemble weather forecasts using spherical Fourier neural operators; Part II: properties of a huge ensemble of hindcasts. Geosci. Model Dev. 18, 5575–5603 and 5605–5633. doi:10.5194/gmd-18-5575-2025, doi:10.5194/gmd-18-5605-2025

[10] **Baño-Medina, J., Sengupta, A., Doyle, J. D., Reynolds, C. A., Watson-Parris, D. & Delle Monache, L. (2025).** Are AI weather models learning atmospheric physics? A sensitivity analysis of cyclone Xynthia. npj Clim. Atmos. Sci. doi:10.1038/s41612-025-00949-6

[11] **Vonich, P. T. & Hakim, G. J. (2024).** Predictability limit of the 2021 Pacific Northwest heatwave from deep-learning sensitivity analysis. Geophys. Res. Lett. 51, e2024GL110651. doi:10.1029/2024GL110651

[12] **Vonich, P. T. & Hakim, G. J. (2025/26).** Atmospheric predictability beyond 30 days with machine learning. arXiv:2504.20238; Artif. Intell. Earth Syst., doi:10.1175/AIES-D-26-0009.1 [verify]

[13] **Winkler, J. & Denhard, M. (2025).** Arnoldi singular vector perturbations for machine learning weather prediction. arXiv:2506.22450

[14] **Alet, F., Price, I., El-Kadi, A., et al. (2025).** Skillful joint probabilistic weather forecasting from marginals (FGN). arXiv:2506.10772

[15] **Bülte, C., Horat, N., Quinting, J. & Lerch, S. (2024/25).** Uncertainty quantification for data-driven weather models. arXiv:2403.13458

[16] **Almeida, R., Otero, N., Fernández-Torres, M.-Á. & Ma, J. (2025/26).** On the predictive skill of AI-based weather models for extreme events using uncertainty quantification. Artif. Intell. Earth Syst. doi:10.1175/AIES-D-25-0113.1; arXiv:2511.17176

[17] **Watt-Meyer, O., Henn, B., McGibbon, J., et al. (2025).** ACE2: accurately learning subseasonal to decadal atmospheric variability and forced responses. npj Clim. Atmos. Sci. 8, 205. doi:10.1038/s41612-025-01090-0

[18] **Dheeshjith, S., Subel, A., Adcroft, A., et al. (2025).** Samudra: an AI global ocean emulator for climate. Geophys. Res. Lett. 52, e2024GL114318. doi:10.1029/2024GL114318

[19] **Duncan, J. P. C., et al. (2025/26).** SamudrACE: fast and accurate coupled climate modeling with 3D ocean and atmosphere emulators. arXiv:2509.12490; reported in GRL 2026 [verify]

[20] **Andersson, T. R., Hosking, J. S., Pérez-Ortiz, M., et al. (2021).** Seasonal Arctic sea ice forecasting with probabilistic deep learning. Nat. Commun. 12, 5124. doi:10.1038/s41467-021-25257-4

[21] **Mayer, K. J. & Barnes, E. A. (2021).** Subseasonal forecasts of opportunity identified by an explainable neural network. Geophys. Res. Lett. 48, e2020GL092092. doi:10.1029/2020GL092092

[22] **Gordon, E. M. & Barnes, E. A. (2022).** Incorporating uncertainty into a regression neural network enables identification of decadal state-dependent predictability in CESM2. Geophys. Res. Lett. 49, e2022GL098635 [verify article no.]

[23] **Shin, N.-Y., Kim, D., Kang, D., Kim, H. & Kug, J.-S. (2024).** Deep learning reveals moisture as the primary predictability source of MJO. npj Clim. Atmos. Sci. 7, 11. doi:10.1038/s41612-023-00561-6

[24] **Toms, B. A., Barnes, E. A. & Hurrell, J. W. (2021).** Assessing decadal predictability in an Earth-system model using explainable neural networks. Geophys. Res. Lett. 48, e2021GL093842. doi:10.1029/2021GL093842

[25] **Labe, Z. M. & Barnes, E. A. (2022).** Predicting slowdowns in decadal climate warming trends with explainable neural networks. Geophys. Res. Lett. 49, e2022GL098173. doi:10.1029/2022GL098173

[26] **Mamalakis, A., Barnes, E. A. & Ebert-Uphoff, I. (2022).** Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. Artif. Intell. Earth Syst. 1(4). doi:10.1175/AIES-D-22-0012.1

[27] **Selz, T. & Craig, G. C. (2023).** Can artificial intelligence-based weather prediction models simulate the butterfly effect? Geophys. Res. Lett. 50, e2023GL105747. doi:10.1029/2023GL105747

[28] **Selz, T. & Craig, G. C. (2026).** Can AI-based weather prediction models simulate the butterfly effect? The role of architecture and implementation. JGR Machine Learning and Computation 3, e2025JH001180. doi:10.1029/2025JH001180

[29] **Hassanzadeh, P., Li, W., Sun, Y. Q., Wang, J., Wikner, A., Finkel, J. & Weare, J. Q. (2026).** Missing the butterfly and predicting the past: features or bugs of accurate AI weather models? arXiv:2608.25835

[30] **Chen, J., Oskarsson, J., Driscoll, S. & Schemm, S. (2026).** Butterfly effect and the kinetic energy cascade in probabilistic machine learning weather prediction models. arXiv:2609.18489

[31] **Bonavita, M. (2024).** On some limitations of current machine learning weather prediction models. Geophys. Res. Lett. 51, e2023GL107377. doi:10.1029/2023GL107377

[32] **Sun, Y. Q., Hassanzadeh, P., Zand, M., Chattopadhyay, A., Weare, J. & Abbot, D. S. (2025).** Can AI weather models predict out-of-distribution gray swan tropical cyclones? PNAS 122, e2420914122. doi:10.1073/pnas.2420914122

[33] **Rackow, T., Koldunov, N., Lessig, C., et al. (2024).** Robustness of AI-based weather forecasts in a changing climate. arXiv:2409.18529

References (2/2)

[34] **Jung, T., Miller, M. J. & Palmer, T. N. (2010).** Diagnosing the origin of extended-range forecast errors. Mon. Wea. Rev. 138, 2434–2446. Accessible summary: Jung et al., "Diagnosing forecast error using relaxation experiments", ECMWF Newsletter 2008. See also Jung (2011) QJRMS 137, 598–606 (relaxation vs 4D-Var).

[35] **Jung, T., Kasper, M. A., Semmler, T. & Serrar, S. (2014).** Arctic influence on subseasonal midlatitude prediction. Geophys. Res. Lett. 41, 3676–3680.

[36] **Magnusson, L. (2017).** Diagnostic methods for understanding the origin of forecast errors. Q. J. R. Meteorol. Soc. 143, 2129–2142.

[37] **Hoskins, B. J., Fonseca, R., Blackburn, M. & Jung, T. (2012).** Relaxing the tropics to an "observed" state: analysis using a simple baroclinic model. Q. J. R. Meteorol. Soc. 138, 1618–1626.

[38] **Li, S., Dias, J., Moore, B. & Quinting, J. (2026).** Impacts of tropical forecast errors on two extreme precipitation events: insights from relaxation experiments using machine-learning weather prediction models. Weather Clim. Dynam. 7, 787–803. doi:10.5194/wcd-7-787-2026

[39] **Li, S., Namdev, P. & Quinting, J. (2026).** Impact of systematic relaxation experiments on subseasonal forecast skill in machine learning weather models. EGU General Assembly, EGU26-4190; precursor Li & Quinting, EMS2025-483 (recorded talk).

[40] **Dai, G. (2024).** Exploring the predictability of Eurasian extreme events: a comparative study of dynamic and AI models. ECMWF informal seminar, 29 Oct 2024. events.ecmwf.int/event/443. Perspective: Mu, M., Qin, J. & Dai, G. (2025). Adv. Atmos. Sci. 42, 1–8, doi:10.1007/s00376-024-4372-7; classical precursor Dai & Mu (2020) Adv. Atmos. Sci. 37, 307–317. [no paper yet; AI models not named on seminar page]

[41] **Perkan, U., Zaplotnik, Ž. & Skok, G. (2025).** Forecast error diagnostics in neural weather models. arXiv:2506.11987 (small convolutional model; tropical, boundary-layer and stratospheric relaxation cross-checked against autodiff sensitivity).

[42] **Duan, S., Zhang, J., Bonfils, C. & Pallotta, G. (2025).** Testing NeuralGCM's capability to simulate future heatwaves based on the 2021 Pacific Northwest heatwave event. npj Clim. Atmos. Sci. doi:10.1038/s41612-025-01137-2 (wind nudging, $\tau = 3$ h, in NeuralGCM; initialisation shock in the stochastic version).

[43] **Huang, et al. (2026).** Diagnosing optimal forcing of the western North Pacific anomalous anticyclone in a deep-learning model. Geophys. Res. Lett. doi:10.1029/2026GL123658 [author list and figures not checked]

[44] **Peings, Y., et al. (2026).** Subseasonal forecasting and MJO teleconnections in ML weather prediction models. J. Geophys. Res. Atmos. doi:10.1029/2025JD044910 [author list not checked]

[45] **Baño-Medina, J., et al. (2025).** Harnessing AI data-driven global weather models for climate attribution: an analysis of the 2017 Oroville Dam extreme atmospheric river. Artif. Intell. Earth Syst. doi:10.1175/AIES-D-24-0090.1. Related: AI-driven weather forecasts to accelerate climate change attribution of heatwaves, Earth's Future (2025), doi:10.1029/2025EF006453.

[46] **Polichtchouk, I., Clare, M., Chantry, M., Gascón, E., Maier-Gerber, M., Vannière, B. & Lang, S. (2026).** Hybrid weather prediction using spectral nudging toward machine-learning forecasts. arXiv:2604.22522; summary in ECMWF Newsletter 184. Earlier: Husain, Separovic et al. (2024) GEM nudged toward GraphCast, arXiv:2407.06100.

[47] **Wu, et al. (2026).** GraphCast and the downward propagation of sudden stratospheric warming anomalies. J. Geophys. Res. Atmos. doi:10.1029/2025JD044852 [title paraphrased; verify]

[48] **Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., et al. (2019).** Inferring causation from time series in Earth system sciences. Nat. Commun. 10, 2553. doi:10.1038/s41467-019-10105-3

[49] **Runge, J., Nowack, P., Kretschmer, M., Flaxman, S. & Sejdinovic, D. (2019).** Detecting and quantifying causal associations in large nonlinear time series datasets. Sci. Adv. 5, eaau4996. doi:10.1126/sciadv.aau4996. Software: tigramite.

[50] **Kretschmer, M., Coumou, D., Donges, J. F. & Runge, J. (2016).** Using causal effect networks to analyze different Arctic drivers of midlatitude winter circulation. J. Climate 29, 4069–4081. doi:10.1175/JCLI-D-15-0654.1

[51] **Kretschmer, M., Runge, J. & Coumou, D. (2017).** Early prediction of extreme stratospheric polar vortex states based on causal precursors. Geophys. Res. Lett. 44, 8592–8600. doi:10.1002/2017GL074696

[52] **Di Capua, G., Kretschmer, M., Runge, J., et al. (2019).** Long-lead statistical forecasts of the Indian summer monsoon rainfall based on causal precursors. Wea. Forecasting 34, 1377–1394. doi:10.1175/WAF-D-19-0002.1

[53] **Saggioro, E. & Shepherd, T. G. (2019).** Quantifying the timescale and strength of Southern Hemisphere intraseasonal stratosphere–troposphere coupling. Geophys. Res. Lett. 46, 13479–13487. doi:10.1029/2019GL084763

[54] **Nowack, P., Runge, J., Eyring, V. & Haigh, J. D. (2020).** Causal networks for climate model evaluation and constrained projections. Nat. Commun. 11, 1415. doi:10.1038/s41467-020-15195-y

[55] **Kretschmer, M., Adams, S. V., Arribas, A., Prudden, R., Robinson, N., Saggioro, E. & Shepherd, T. G. (2021).** Quantifying causal pathways of teleconnections. Bull. Amer. Meteor. Soc. 102, E2247–E2263. doi:10.1175/BAMS-D-20-0117.1

[56] **Runge, J., Gerhardus, A., Varando, G., Eyring, V. & Camps-Valls, G. (2023).** Causal inference for time series. Nat. Rev. Earth Environ. 4, 487–505. doi:10.1038/s43017-023-00431-y. See also Camps-Valls et al. (2023) Phys. Rep. 1044, 1–68, doi:10.1016/j.physrep.2023.10.005.

[57] **Spuler, F., Kretschmer, M., Balmaseda, M. A., Gudoshava, M. & Shepherd, T. G. (2026).** Disentangling regional impacts of joint teleconnections using causal representation learning. arXiv:2603.02879

[58] **Pérez-Carrasquilla, J. S., Nichol, J., Robledo, J., Bull, D., Dagon, K., Evans, J. & Molina, M. J. (2026).** Tracing the space-time causal origins of Earth system extremes. arXiv:2607.10033 [author list not checked]

[59] **Tian, et al. (2026).** Evaluating ML weather models for data assimilation: fundamental limitations in tangent-linear and adjoint properties. Geophys. Res. Lett. doi:10.1029/2025GL119402 [author list not checked]

[60] **Miloshevich, G., Cozian, B., Abry, P., Borgnat, P. & Bouchet, F. (2023).** Probabilistic forecasts of extreme heatwaves using convolutional neural networks in a regime of lack of data. Phys. Rev. Fluids 8, 040501. doi:10.1103/PhysRevFluids.8.040501

[61] **Lopez-Gomez, I., McGovern, A., Agrawal, S. & Hickey, J. (2023).** Global extreme heat forecasting using neural weather models. Artif. Intell. Earth Syst. 2, e220035. doi:10.1175/AIES-D-22-0035.1