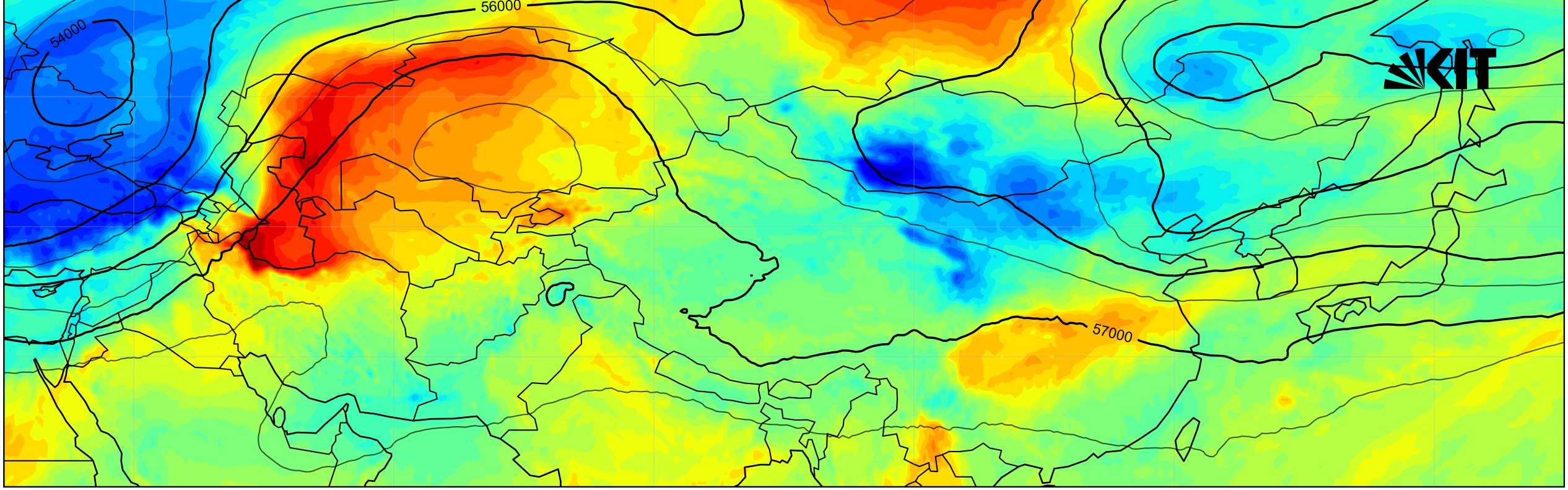


Physics-based models outperform AI weather forecasts of record-breaking extremes

Zhongwei Zhang, Erich Fischer, Jakob Zscheischler, and Sebastian Engelke | May 7, 2026

Literature review: How do AI weather models perform at forecasting extreme events?

- AI models outperform HRES in forecasting cyclone tracks with smaller track errors (Bi *et al.* 2023; Lam *et al.* 2023; Bodnar *et al.* 2025)

Figure source: Figures 2 and 3 in Olivetti & Messori 2024

Literature review: How do AI weather models perform at forecasting extreme events?

- AI models outperform HRES in forecasting cyclone tracks with smaller track errors (Bi *et al.* 2023; Lam *et al.* 2023; Bodnar *et al.* 2025)
- However, the AI forecasts have larger intensity errors than operational physics-based forecasts (DeMaria *et al.* 2025), particularly in the EHem (McGovern *et al.* 2026)

Figure source: Figures 2 and 3 in Olivetti & Messori 2024

Literature review: How do AI weather models perform at forecasting extreme events?

- AI models outperform HRES in forecasting cyclone tracks with smaller track errors (Bi *et al.* 2023; Lam *et al.* 2023; Bodnar *et al.* 2025)
- However, the AI forecasts have larger intensity errors than operational physics-based forecasts (DeMaria *et al.* 2025), particularly in the EHem (McGovern *et al.* 2026)
- The North American winter storm in 2021 is well forecast by AI models, but they underperform HRES at forecasting the 2021 Pacific Northwest heatwave (Pasche *et al.* 2025; McGovern *et al.* 2026)

Figure source: Figures 2 and 3 in Olivetti & Messori 2024

Literature review: How do AI weather models perform at forecasting extreme events?

- AI models outperform HRES in forecasting cyclone tracks with smaller track errors (Bi *et al.* 2023; Lam *et al.* 2023; Bodnar *et al.* 2025)
- However, the AI forecasts have larger intensity errors than operational physics-based forecasts (DeMaria *et al.* 2025), particularly in the EHem (McGovern *et al.* 2026)
- The North American winter storm in 2021 is well forecast by AI models, but they underperform HRES at forecasting the 2021 Pacific Northwest heatwave (Pasche *et al.* 2025; McGovern *et al.* 2026)

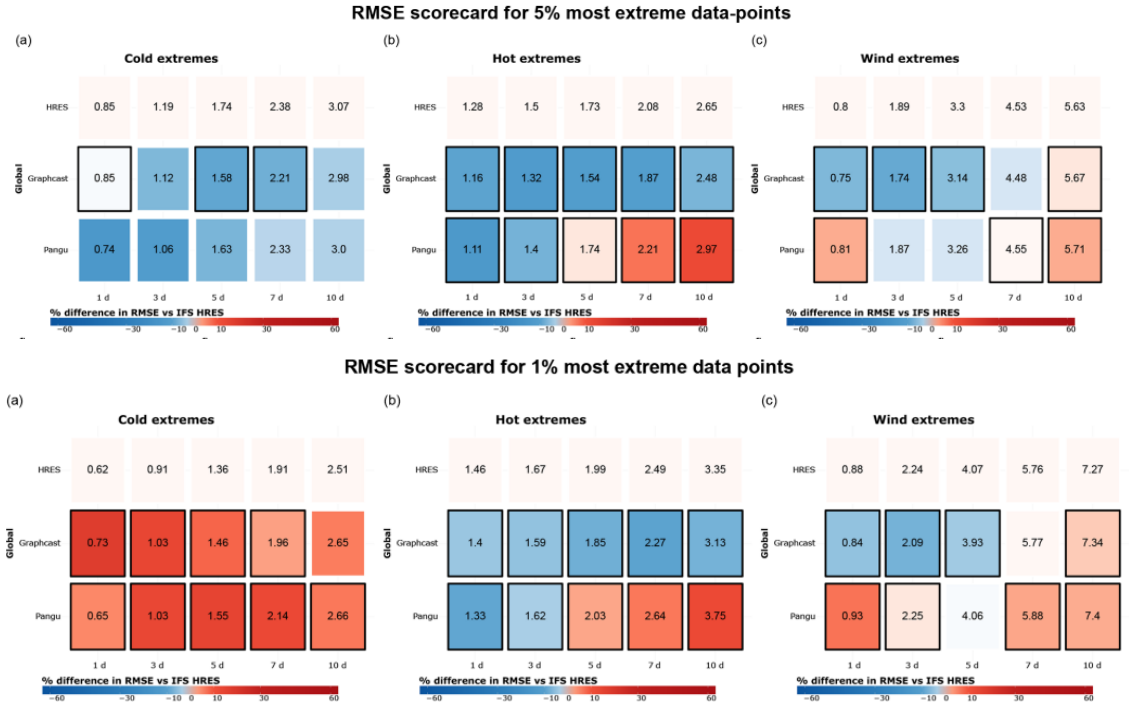
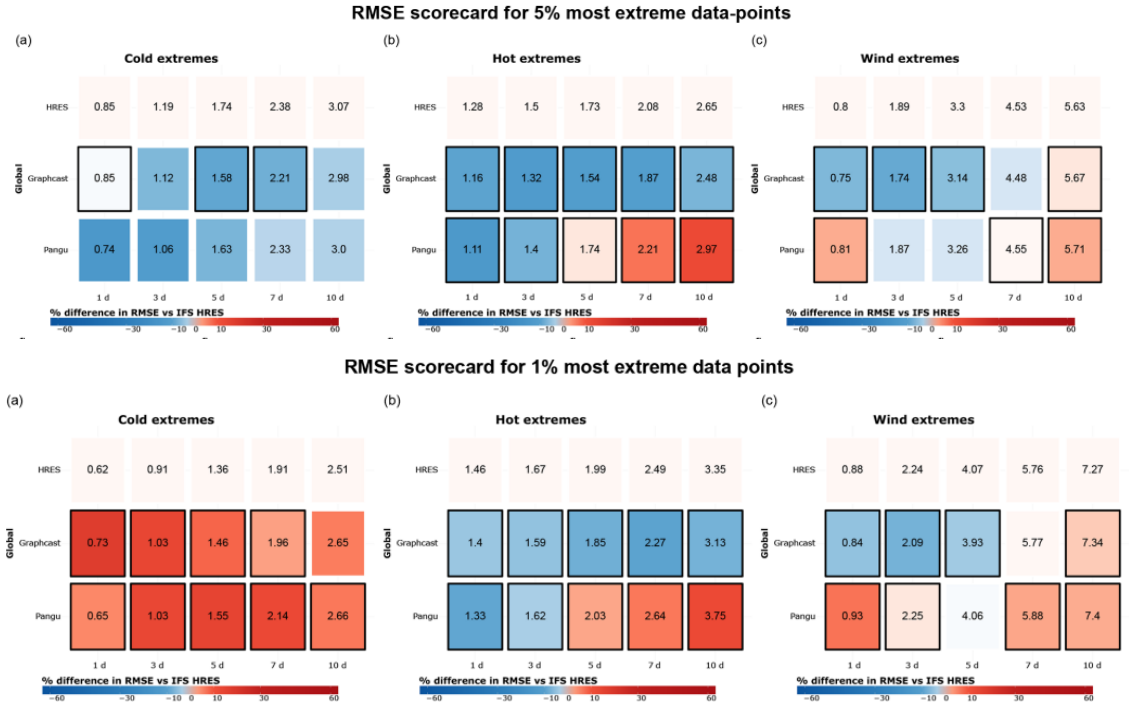


Figure source: Figures 2 and 3 in Olivetti & Messori 2024

Literature review: How do AI weather models perform at forecasting extreme events?

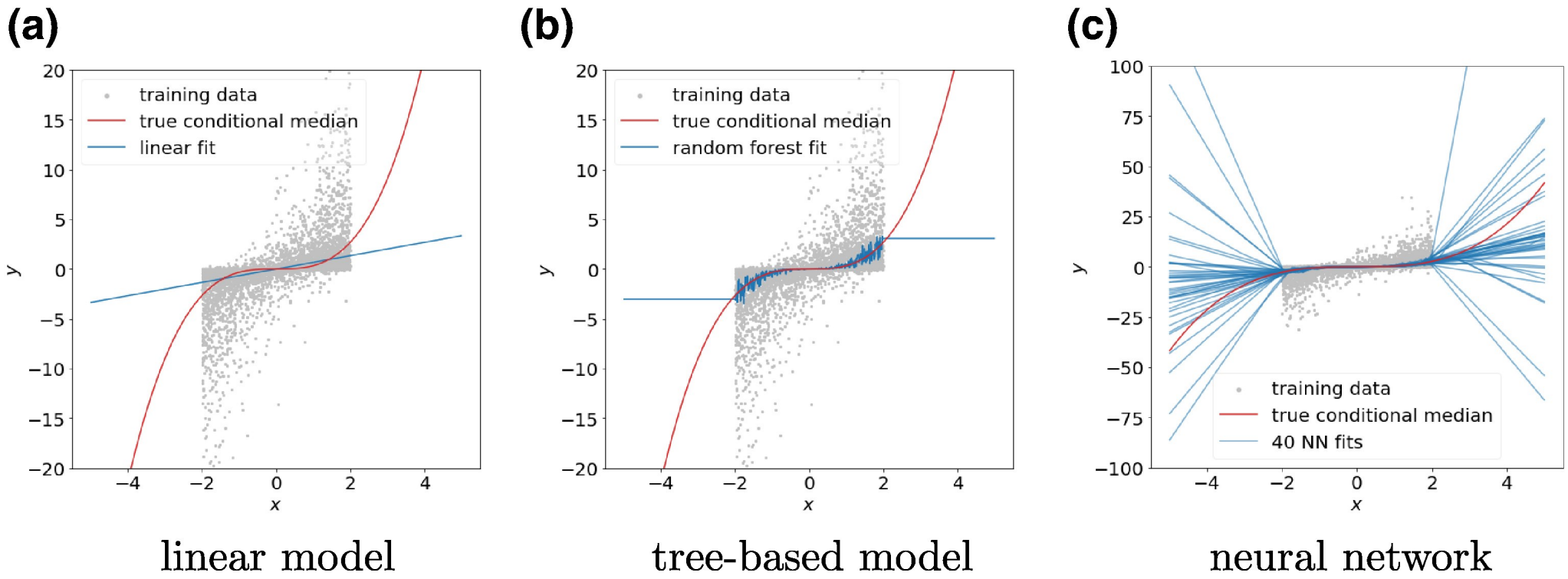
- AI models outperform HRES in forecasting cyclone tracks with smaller track errors (Bi *et al.* 2023; Lam *et al.* 2023; Bodnar *et al.* 2025)
- However, the AI forecasts have larger intensity errors than operational physics-based forecasts (DeMaria *et al.* 2025), particularly in the EHem (McGovern *et al.* 2026)
- The North American winter storm in 2021 is well forecast by AI models, but they underperform HRES at forecasting the 2021 Pacific Northwest heatwave (Pasche *et al.* 2025; McGovern *et al.* 2026)



Conclusion: no consistent results, verification results often depend on **which type of extreme events** and **which specific events** are considered; often no clear **interpretation** of the results

Figure source: Figures 2 and 3 in Olivetti & Messori 2024

Extrapolation problem: How do neural networks underlying AI models extrapolate?

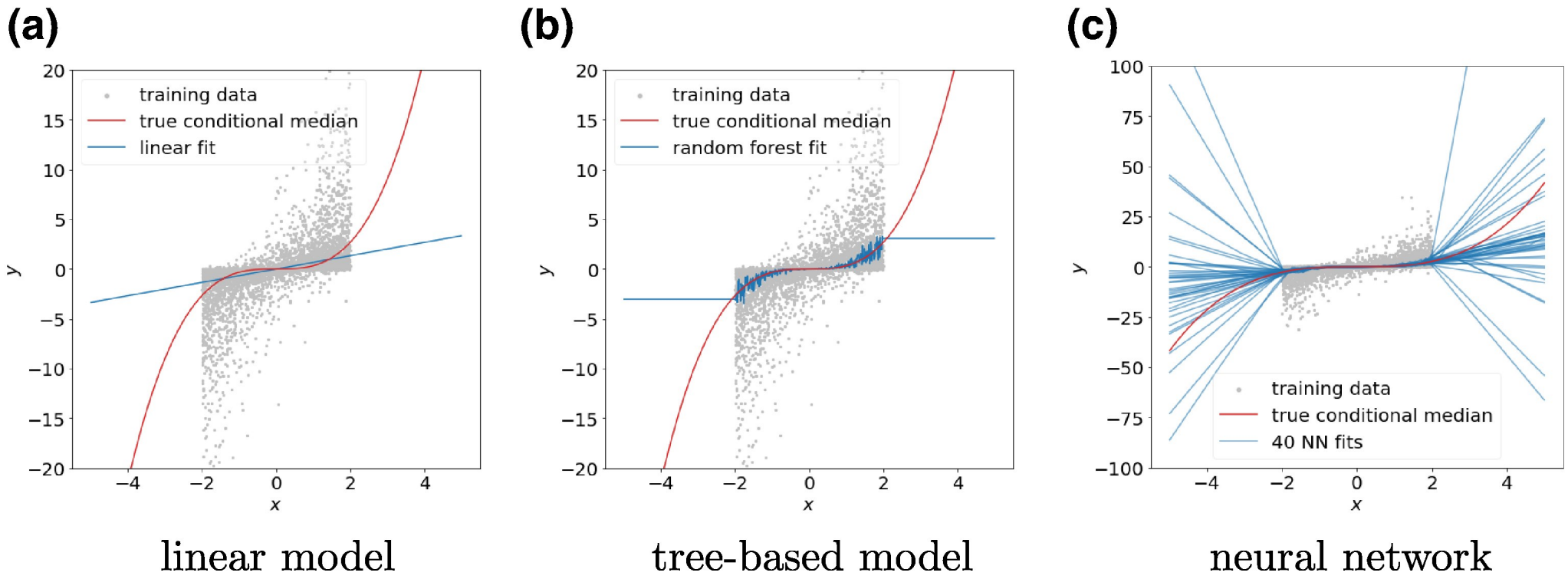


linear model | tree-based model | neural network

- All neural networks fit almost perfectly the training data

Figure source: Figures 1 in Shen & Meinshausen 2025

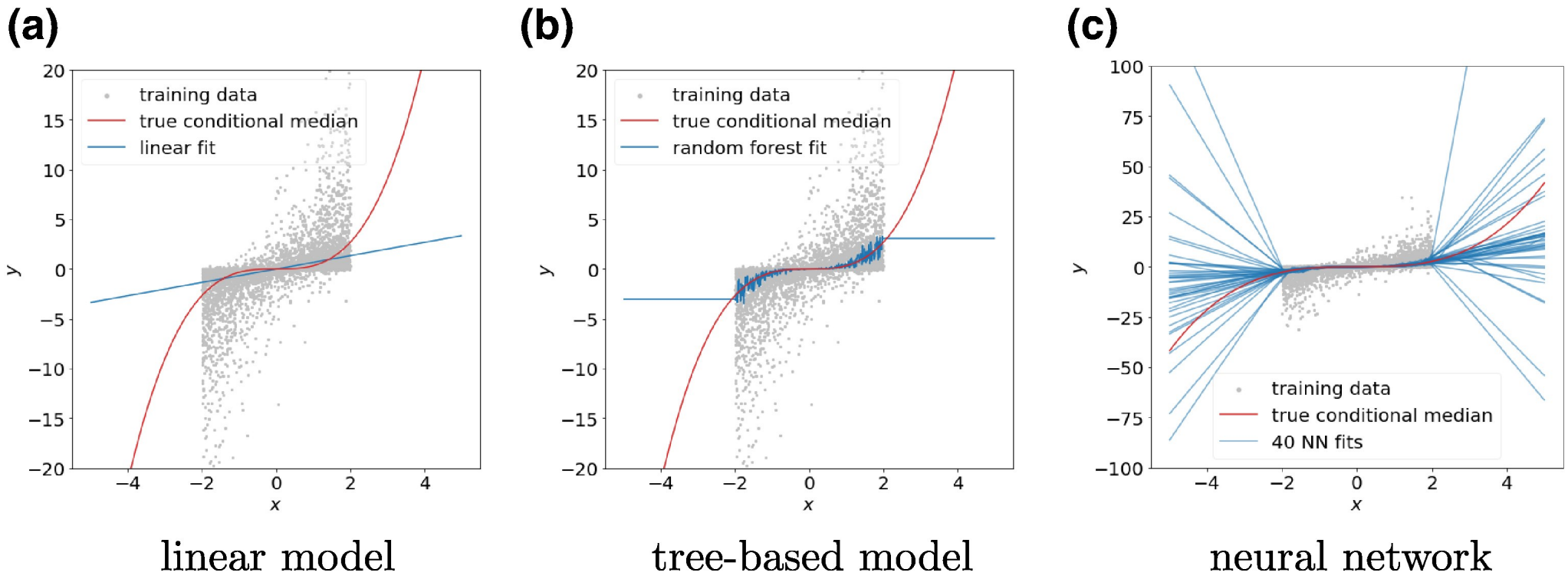
Extrapolation problem: How do neural networks underlying AI models extrapolate?



- All neural networks fit almost perfectly the training data
- Yet they exhibit **uncontrollable and arbitrary behavior** outside the training domain $[-2, 2]$

Figure source: Figures 1 in Shen & Meinshausen 2025

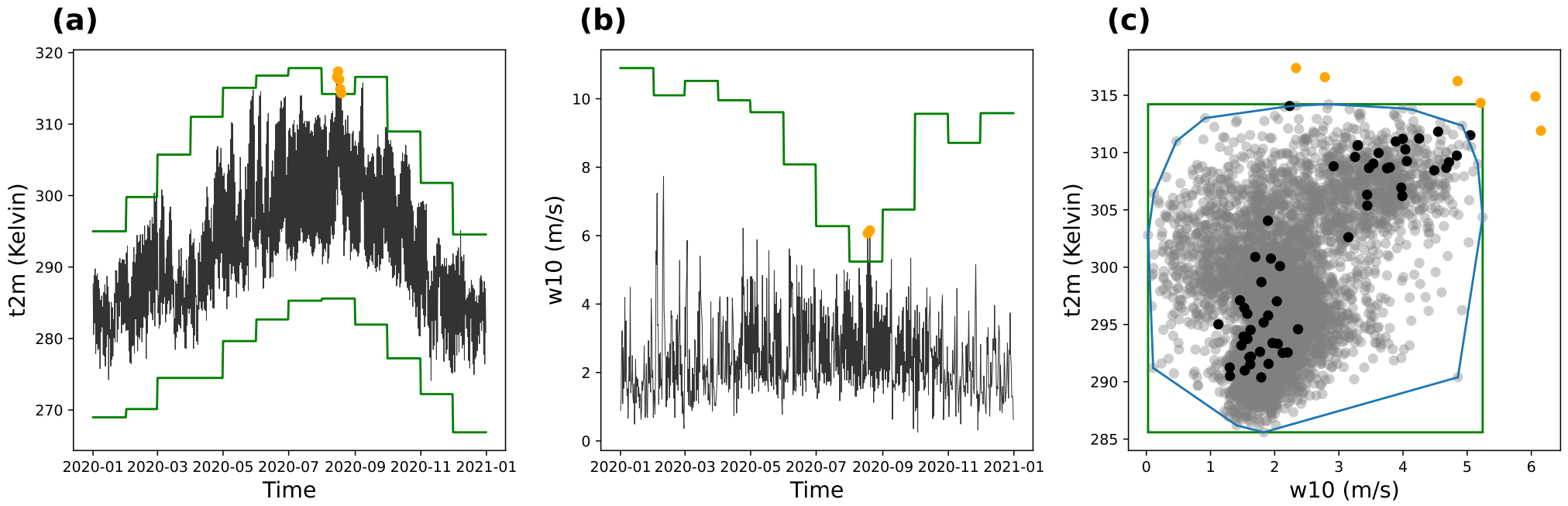
Extrapolation problem: How do neural networks underlying AI models extrapolate?



- All neural networks fit almost perfectly the training data
- Yet they exhibit **uncontrollable and arbitrary behavior** outside the training domain $[-2, 2]$
- **Question**: Do AI weather models suffer from the extrapolation problem?

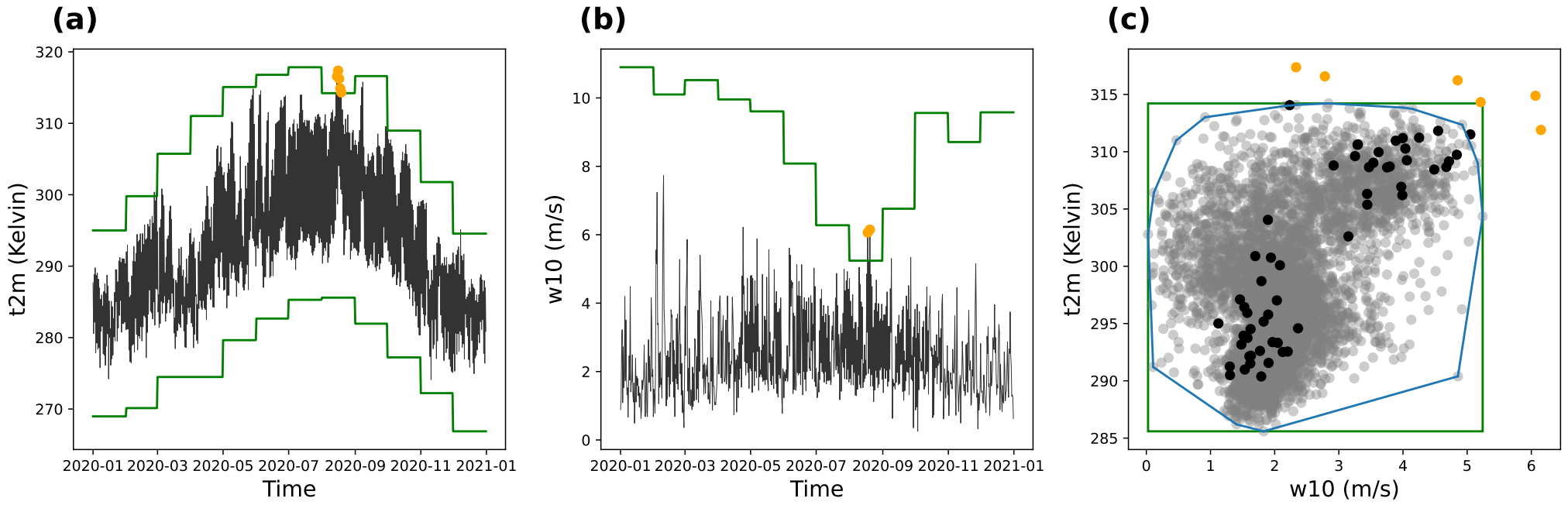
Figure source: Figures 1 in Shen & Meinshausen 2025

How to define the training domain of AI weather models?



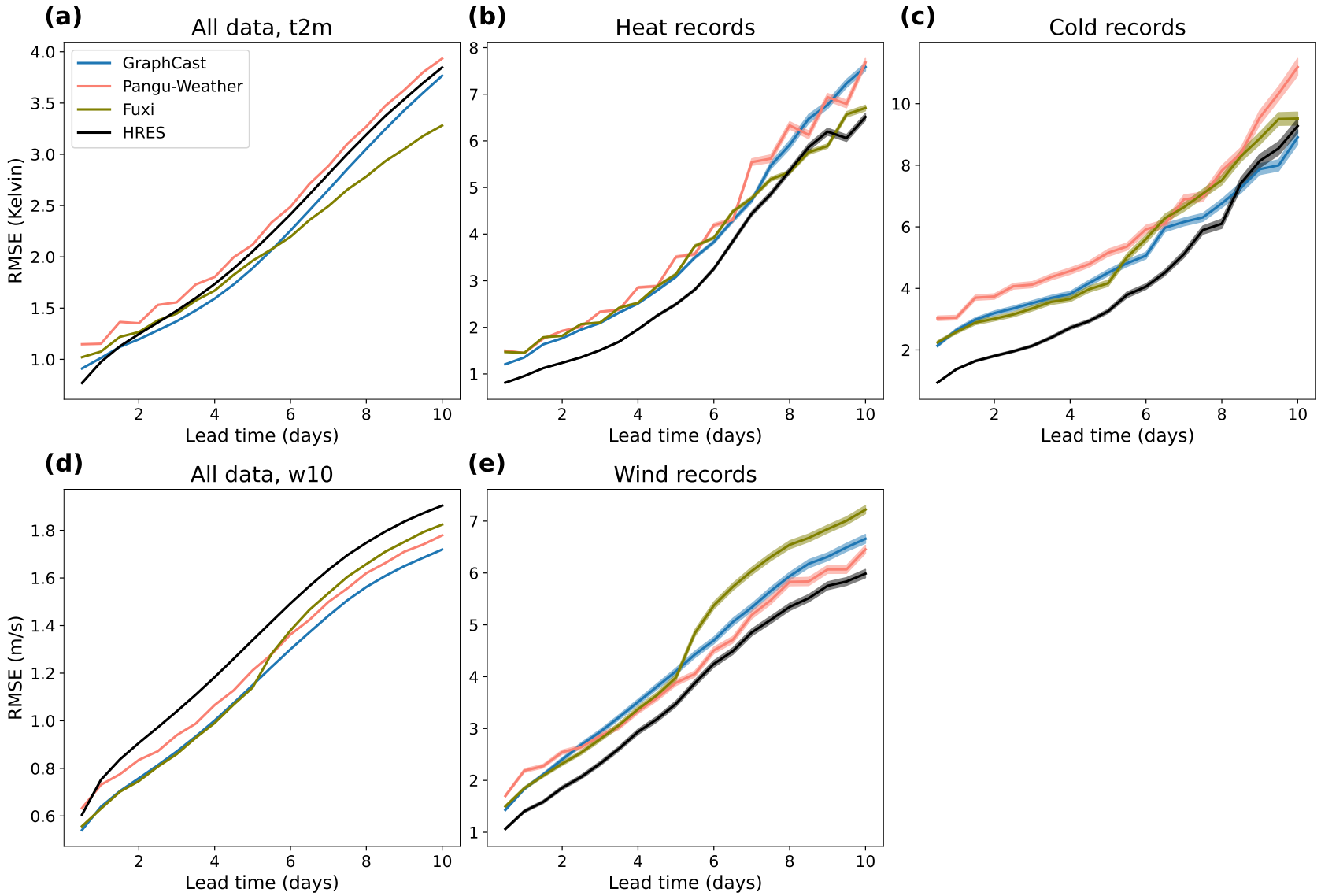
- Define records for each variable, each grid cell, and each month based on the training data of AI models (ERA5 from 1979 – 2017)

How to define the training domain of AI weather models?

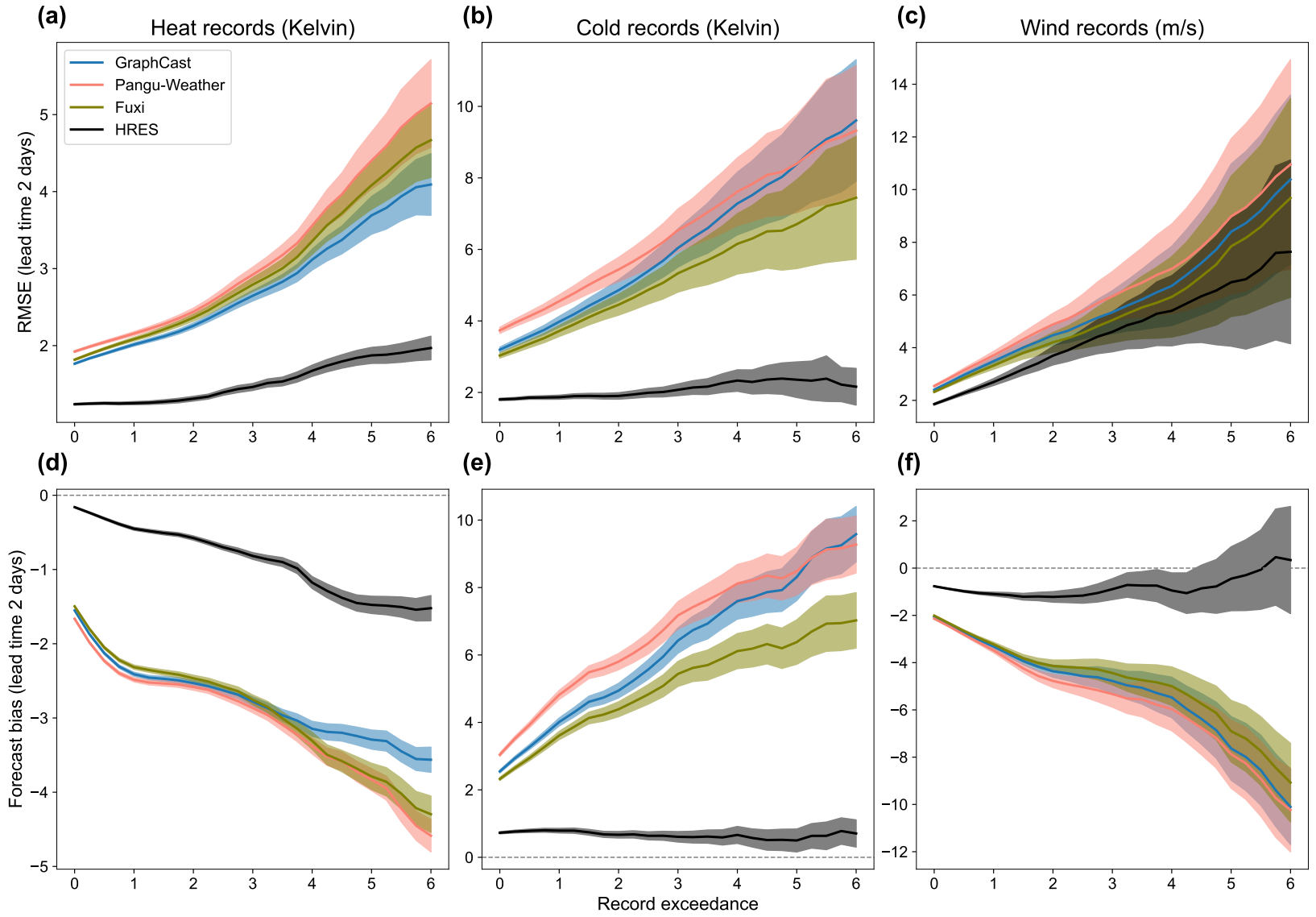


- Define records for each variable, each grid cell, and each month based on the training data of AI models (ERA5 from 1979 – 2017)
- Systematically evaluate AI forecasts of record-breaking heat, cold, and wind extremes, and compare to physics-based HRES forecast

Model comparison on records' intensity



Forecast errors and bias against record exceedance



Summary

- Through systematic forecast evaluation, we show that for record-breaking weather extremes, the leading physics-based numerical model still consistently outperforms state-of-the-art deterministic AI models such as GraphCast, Pangu-Weather and Fuxi for heat, cold, and wind across nearly all lead times
- We further find that the examined AI models tend to underestimate both the frequency and intensity of record-breaking extreme events
- Here we focused on deterministic AI weather models. Evaluation of probabilistic AI weather forecasts is ongoing
- Interpretation is equally important as verification
- Open-source pre-trained model weights and forecast data are crucial for independent evaluation to build trust in the AI models

Thank you for your attention!

Contact: zhongwei.zhang@kit.edu

References I

1. Bi, K. *et al.* Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619,** 533–538 (2023).
2. Bodnar, C. *et al.* A foundation model for the Earth system. *Nature* **641,** 1180–1187 (2025).
3. DeMaria, M. *et al.* An Operations-Based Evaluation of Tropical Cyclone Track and Intensity Forecasts from Artificial Intelligence Weather Prediction Models. *Artificial Intelligence for the Earth Systems* **4** (2025).
4. Lam, R. *et al.* Learning skillful medium-range global weather forecasting. *Science* **382,** 1416–1421 (2023).
5. McGovern, A. *et al. Extreme Weather Bench: A framework and benchmark for evaluation of high-impact weather.* Preprint at arXiv:2605.01126. 2026.
6. Olivetti, L. & Messori, G. Do data-driven models beat numerical models in forecasting weather extremes? A comparison of IFS HRES, Pangu-Weather, and GraphCast. *Geoscientific Model Development* **17,** 7915–7962 (2024).
7. Pasche, O. C., Wider, J., Zhang, Z., Zscheischler, J. & Engelke, S. Validating Deep Learning Weather Forecast Models on Recent High-Impact Extreme Events. *Artificial Intelligence for the Earth Systems* **4,** e240033 (2025).

References II

8. Shen, X. & Meinshausen, N. Engression: extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **87,** 653–677 (2025).