# Is there an end to deep learning for weather and climate?

Martin Schultz, Jülich Supercomputing Center, Germany

# A brief history of AI

- **1943** – McCulloch and Pitts describe artificial neuron
- **1950** – Alan Turing test - if a machine tricks human into thinking it's a human, then it has intelligence.
  First successful numerical weather forecast
- **1956** – John McCarthy introduces the term 'Artificial Intelligence.'
- **1964** – Joseph Weizenbaum - chatbot capable of holding conversations with humans.
  First gigascale computer (1972)
- **1997** – A chess-playing computer Deep Blue defeats chess champion Garry Kasparov.
- **1999** – Sony launches the first robot pet dog AiBO whose personality and skills develop over time.
  First terascale computer (1992)
- **2002** – The first mass-produced robotic vacuum cleaner Roomba learns to navigate homes.
- **2011** – Apple integrates into iPhone Siri - intelligent virtual assistant.
  First petascale computer (2008)
- **2011** – Question answering computer Watson wins 1st place on TV show Jeopardy.
- **2014** – Amazon launches Alexa - intelligent virtual assistant that completes shopping tasks.
  NVIDIA V100 GPU (2017)
  First attempt of a neural network weather model (2018)
- **2020** – GPT-3 tool for automated conversations is introduced
- **2022** – ChatGPT is available to public testing.
  First AI weather forecast surpasses operational model
- **2023** – ECMWF runs AI weather forecasts operationally

Adapted from https://www.infodiagram.com

Review Article | Published: 02 September 2015

# The quiet revolution of numerical weather prediction

Peter Bauer, Alan Thorpe & Gilbert Brunet

*Nature* **525**, 47–55 (2015) | Cite this article

**48k** Accesses | **1239** Citations | **1116** Altmetric | Metrics

1960-2010

Perspective | Published: 22 February 2021

# The digital revolution of Earth-system science

Peter Bauer, Peter D. Dueben, Torsten Hoefler, Tiago Quintino, Thomas C. Schulthess & Nils P. Wedi

*Nature Computational Science* **1**, 104–113 (2021) | Cite this article

**18k** Accesses | **94** Citations | **300** Altmetric | Metrics

2005-2025

FOURCASTNET: A GLOBAL DATA-DRIVEN HIGH-RESOLUTION WEATHER MODEL USING ADAPTIVE FOURIER NEURAL OPERATORS

...g skillful medium-range ...casting

...r: A 3D High-Resolution System ...ccurate Global Weather Forecast

# The AI revolution in weather and climate modeling

Ashesh Chattopadhyay, Rice University, Houston, TX 77005 · Morteza Mardani, NVIDIA Corporation, Santa Clara, CA 95051 · Thorsten Kurth, NVIDIA Corporation, Santa Clara, CA 95051 · David Hall, NVIDIA Corporation, Santa Clara, CA 95051 · Zongyi Li, California Institute of Technology, Pasadena, CA 91125, NVIDIA Corporation, Santa Clara, CA 95051 · Kamyar Azizzadenesheli, Purdue University, West Lafayette, IN 47907 · Pedram Hassanzadeh, Rice University, Houston, TX 77005 · Karthik Kashinath, NVIDIA Corporation, Santa Clara, CA 95051 · Animashree Anandkumar, California Institute of Technology, Pasadena, CA 91125, NVIDIA Corporation, Santa Clara, CA 95051

2022-

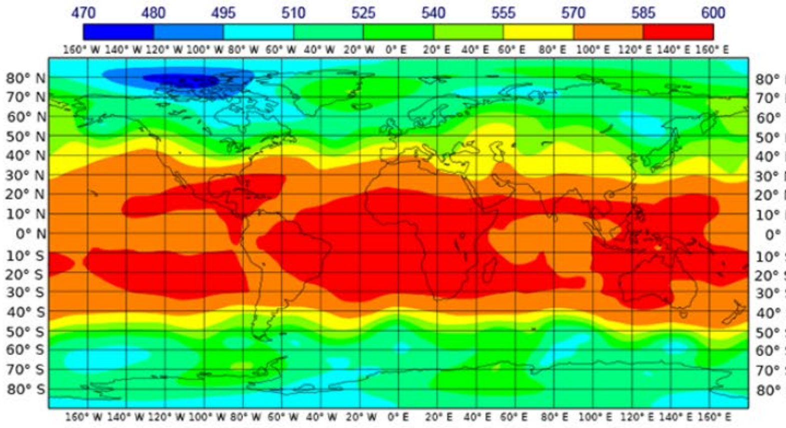
## A few milestones on the path to AI weather forecasts

**2018:** First trial of a pure ML model: FNN with 4 layers

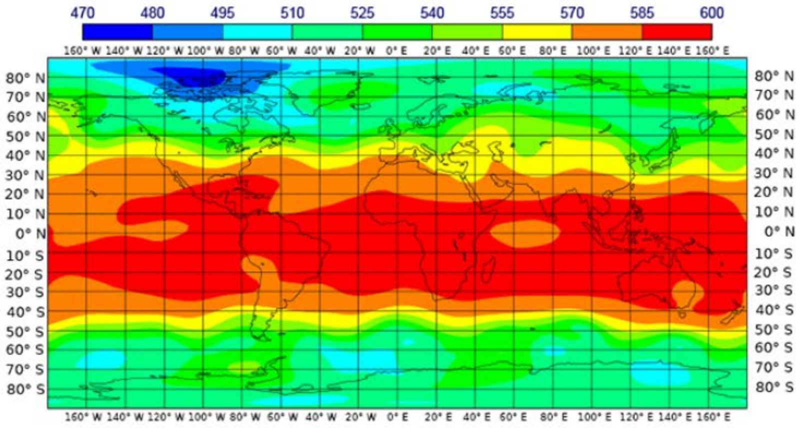
*Input:* hourly Z500, 1 pressure level, 1860 grid points (6° resolution), 67200 snap shots 2010—2017

*Output:* Z500 up to 120 hours ahead (autoregressive rollout)

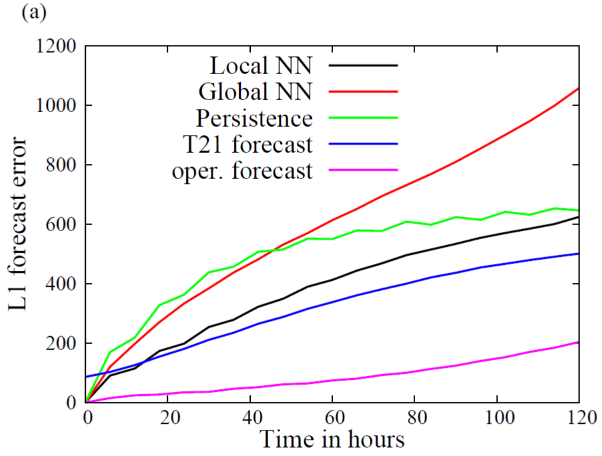
24 h; analysis




24 h; local neural network




Model error



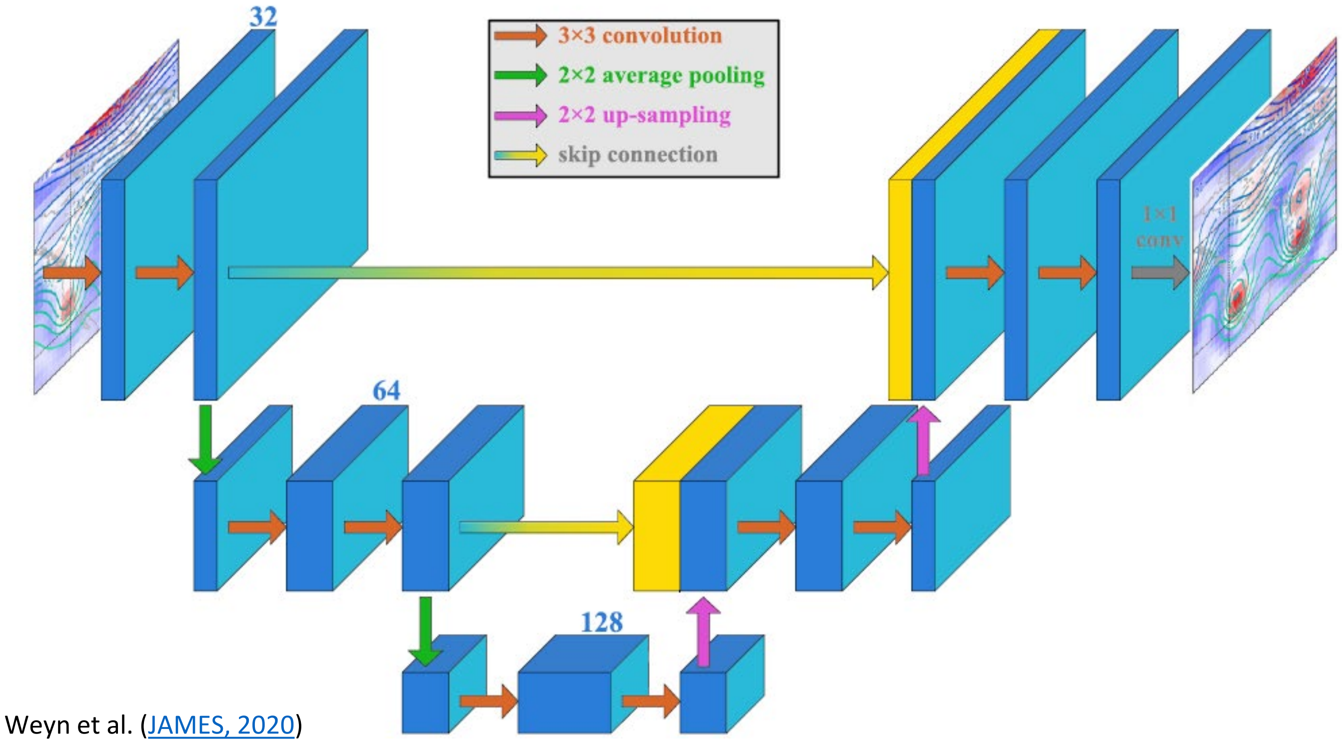

Düben and Bauer (GMD, 2018)

## A few milestones on the path to AI weather forecasts

**2020:** global forecasts up to 14 days with a U-net, 11 Conv2D layers

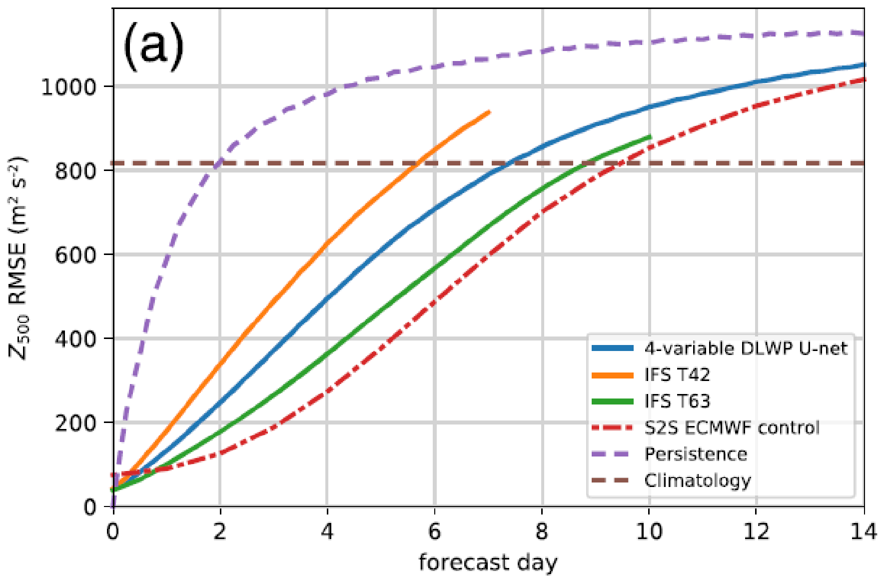
*Input:* 4 variables: $Z_{500}$, $Z_{1000}$, $\tau_{300\text{-}700}$, $T_{2m}$, 6-hourly data, 1917-2012 (100,000 samples), 2° horizontal resolution

*Output:* 4 variables, 6-hourly for t+6 and t+12




Weyn et al. (JAMES, 2020)

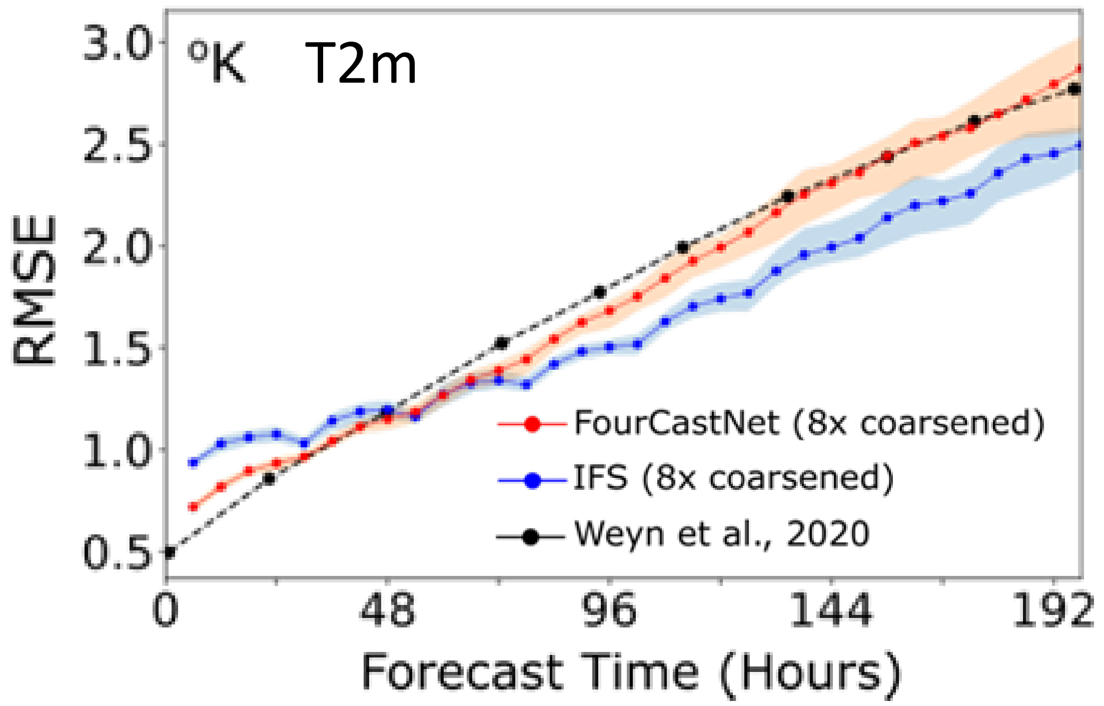
Model error (RMSE)



## A few milestones on the path to AI weather forecasts

**2022:** Global forecasts at 0.25 ° resolution, 20 variables at 5 pressure levels, Transformer + Fourier Neural Operators

Model error

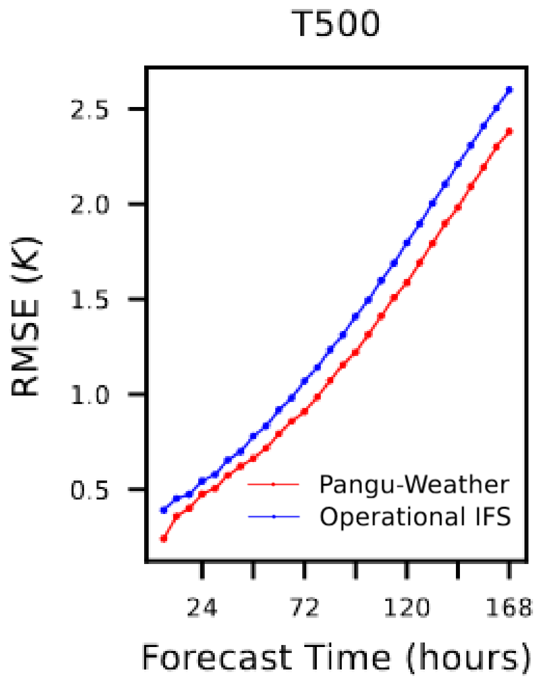



Pathak et al. (2022): Fourcastnet

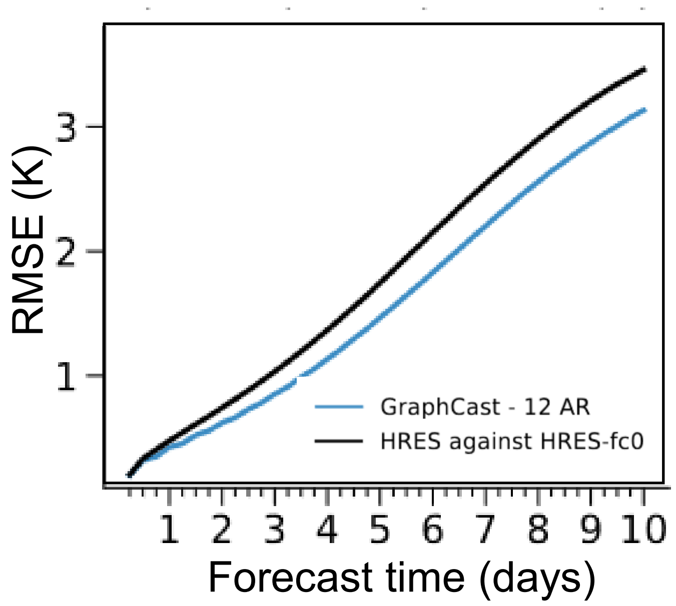
## A few milestones on the path to AI weather forecasts

**2022/23:** The breakthrough – DL models outperform IFS HRES

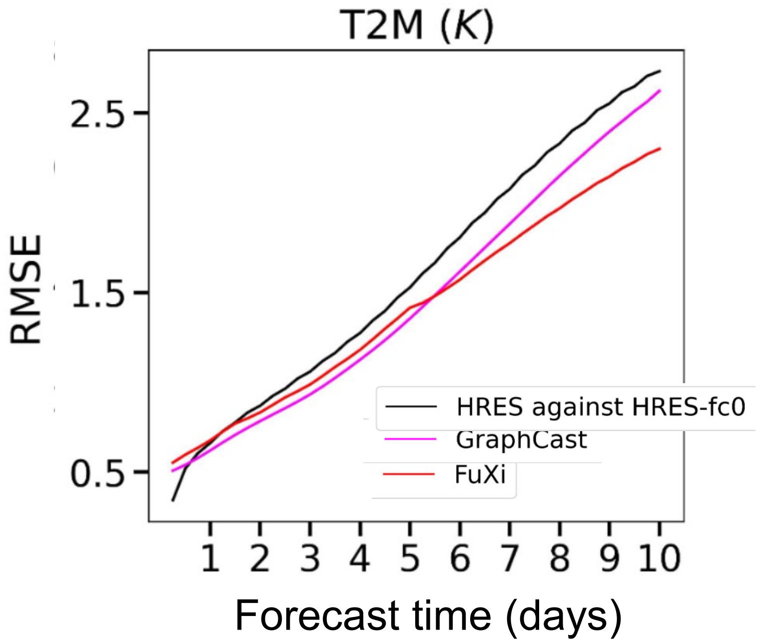
Model error




Bi et al. (2022): Pangu-Weather




Lam et al. (2022): GraphCast

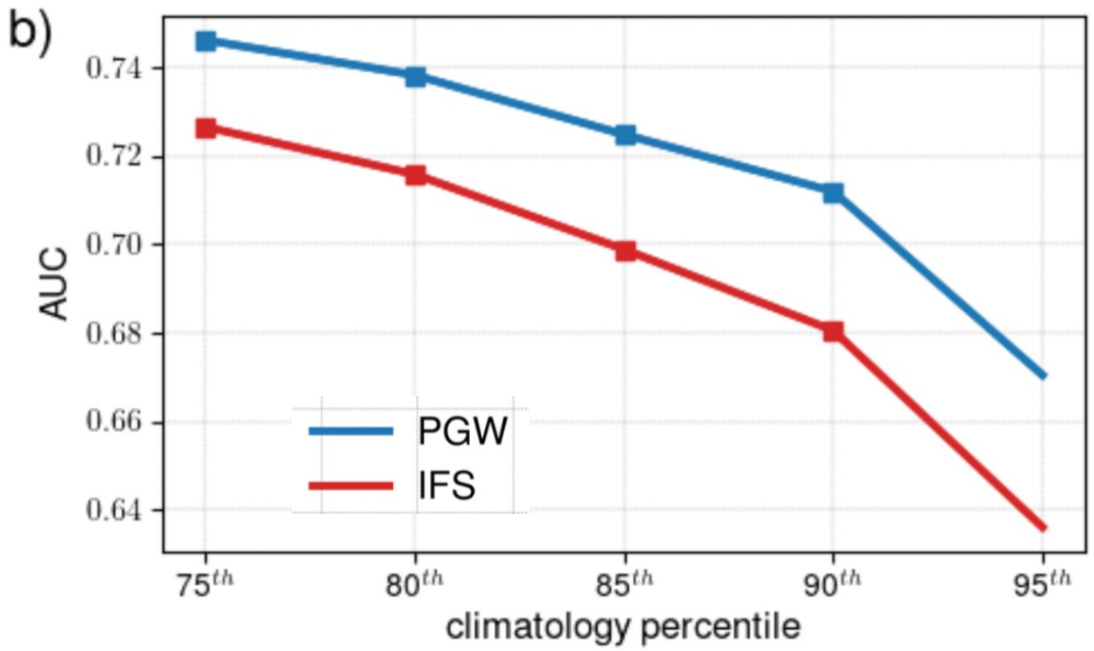



Chen et al. (2023): FuXi

## Limitations of DL weather models

**2022/23:** The breakthrough – DL models outperform IFS HRES; **but are they really better?**

**T2m extremes Europe, summer 2022**

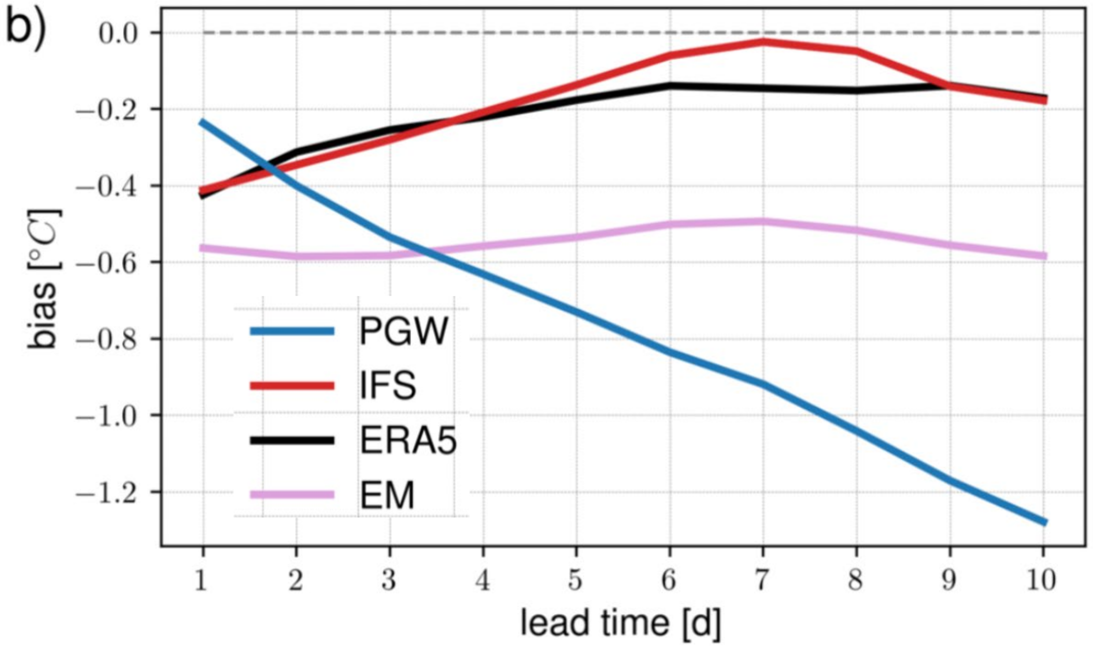



Ben-Bouallegue et al. (Arxiv, 2023)

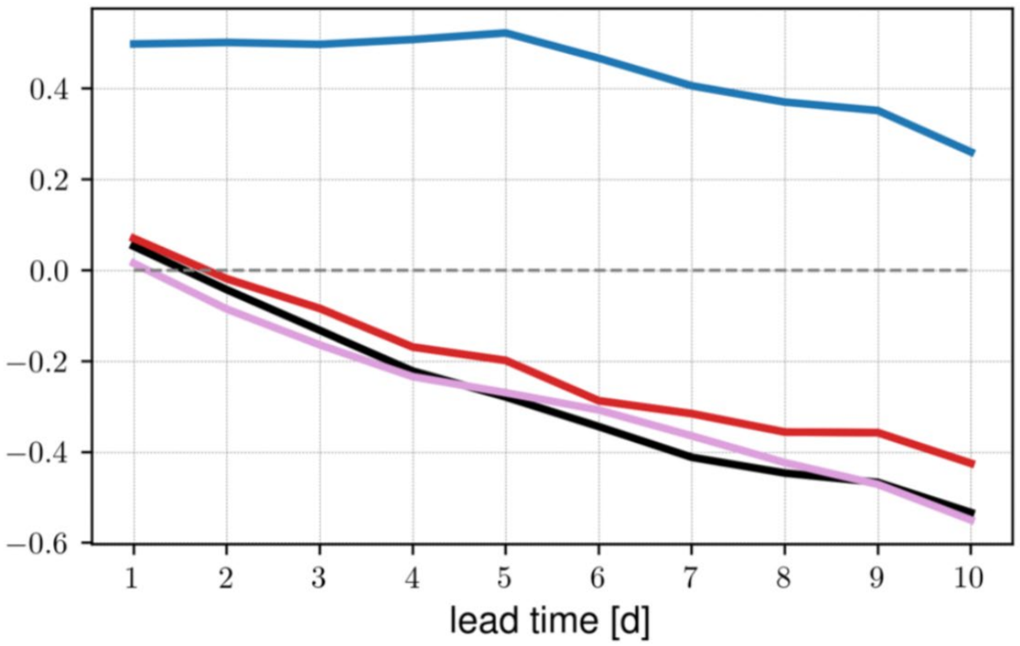
## Limitations of DL weather models

**2022/23:** The breakthrough – DL models outperform IFS HRES; **but are they really better?**

**T2m Europe, summer 2022**




**T2m Europe, winter 2022/23**




Ben-Bouallegue et al. (Arxiv, 2023)

## Limitations of DL weather models

**IFS-HRES** | **ERA5** | **PanguWeather**

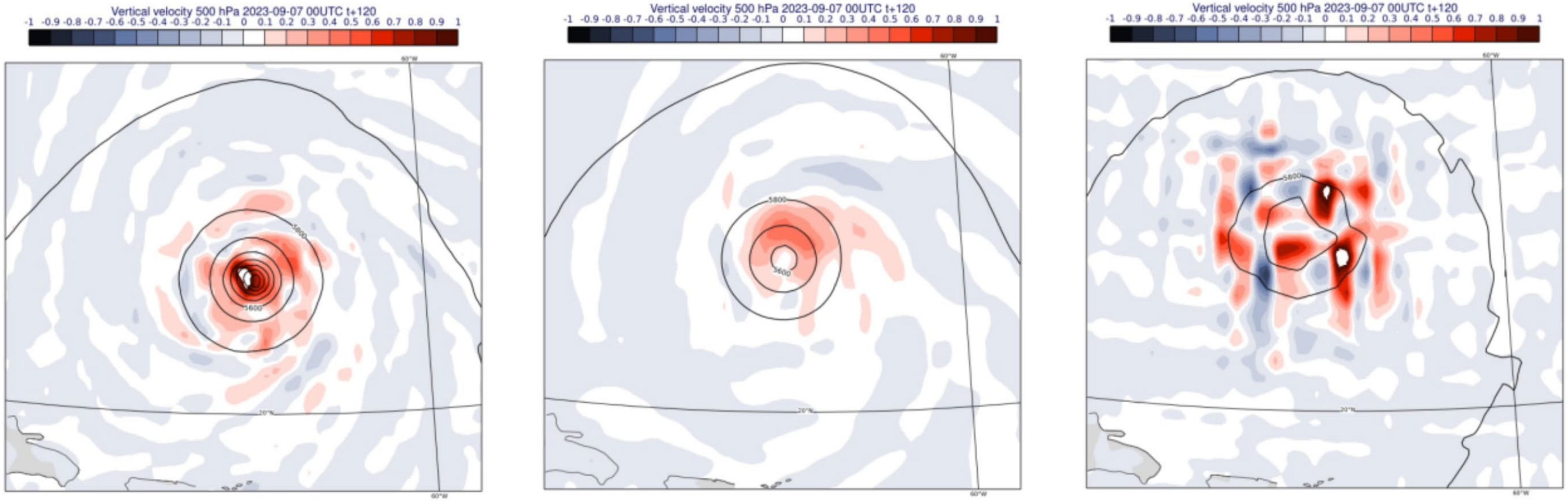



Figure 10: *Plots of t+120 hours forecast vertical velocities (shaded, units: m/s) at 500 hPa from:*

Bonavita, [2024]

**DNN stability and robustness**

„However, for the local networks that use the special treatment of the area around the pole, as discussed in the previous section, the forecast error diverges for the 7×7 and the 9×9 configuration. […likely that this can be fixed…]“ (Düben and Bauer, 2018)

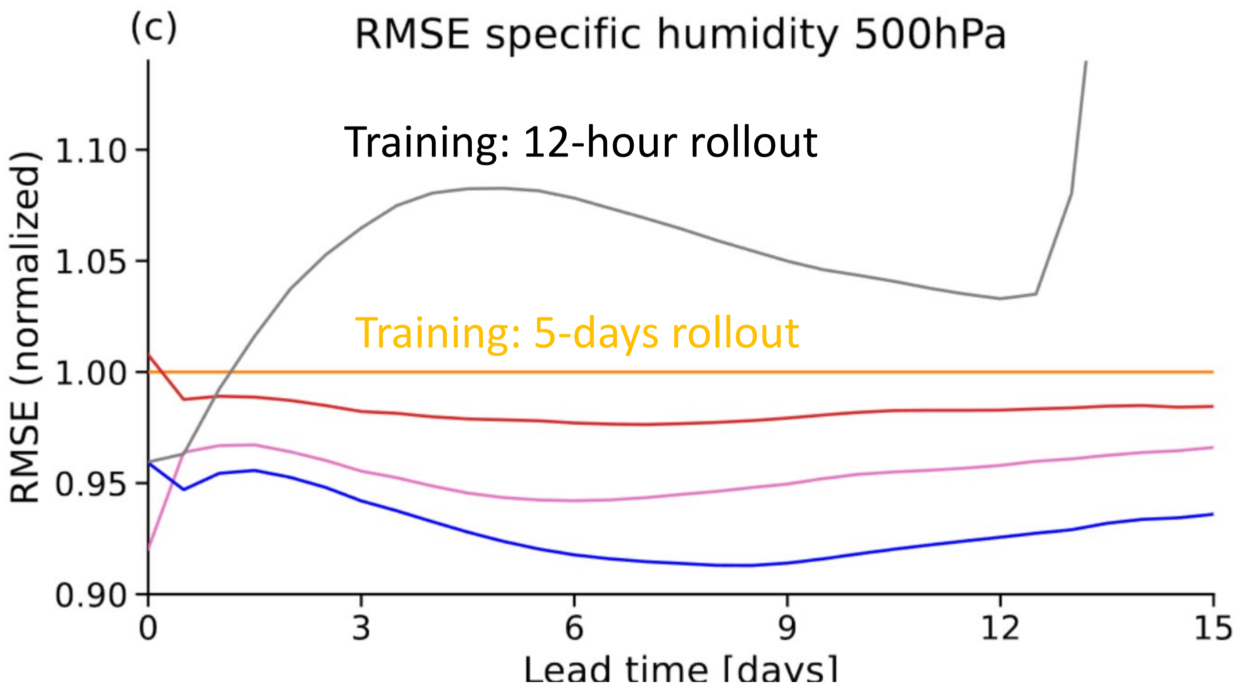
„Every one of the 4-week forecasts initialized twice weekly in the 2-year test set (210 total forecasts) was free from instabilities and the amplification of spurious perturbations.“ (Weyn et al., 2020)

„Moreover, the similarities in error growth of a data-driven forecast and a standard NWP forecast indicate similar sensitivities to chaos between ML-based and physically-based models.“ (Ben Bouallegue et al., 2023)

„More generally, the discussion above and the results presented here highlight one of the main challenges for the next generation of data-driven ML prediction models, namely, how to produce forecasts that are skilful and at the same time dynamically and physically consistent at all relevant spatial scales.“ (Bonavita, 2024)

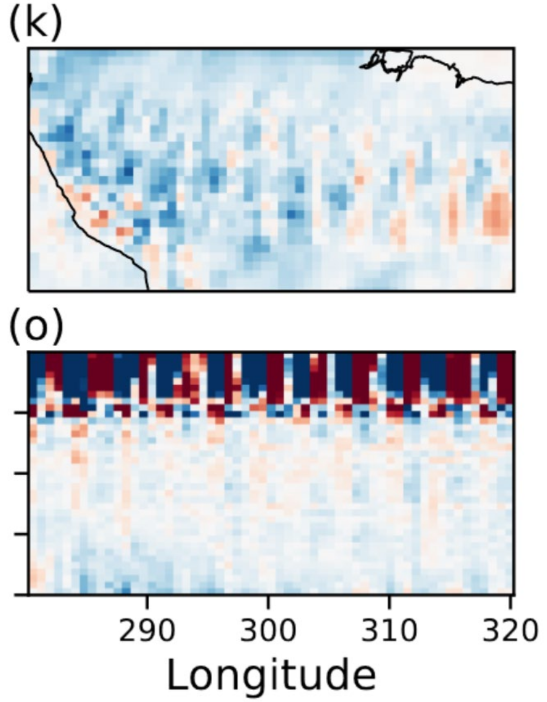
## Long-term rollout

Weyn et al. (2020): up to 1 year
Watt-Meyer et al. (2023) [ACE]: up to 100 years
Kochkov et al. (2024) [NeuralGCM]: up to 40 years (22 out of 37 runs stable)




Kochkov et al. (Arxiv, 2024)

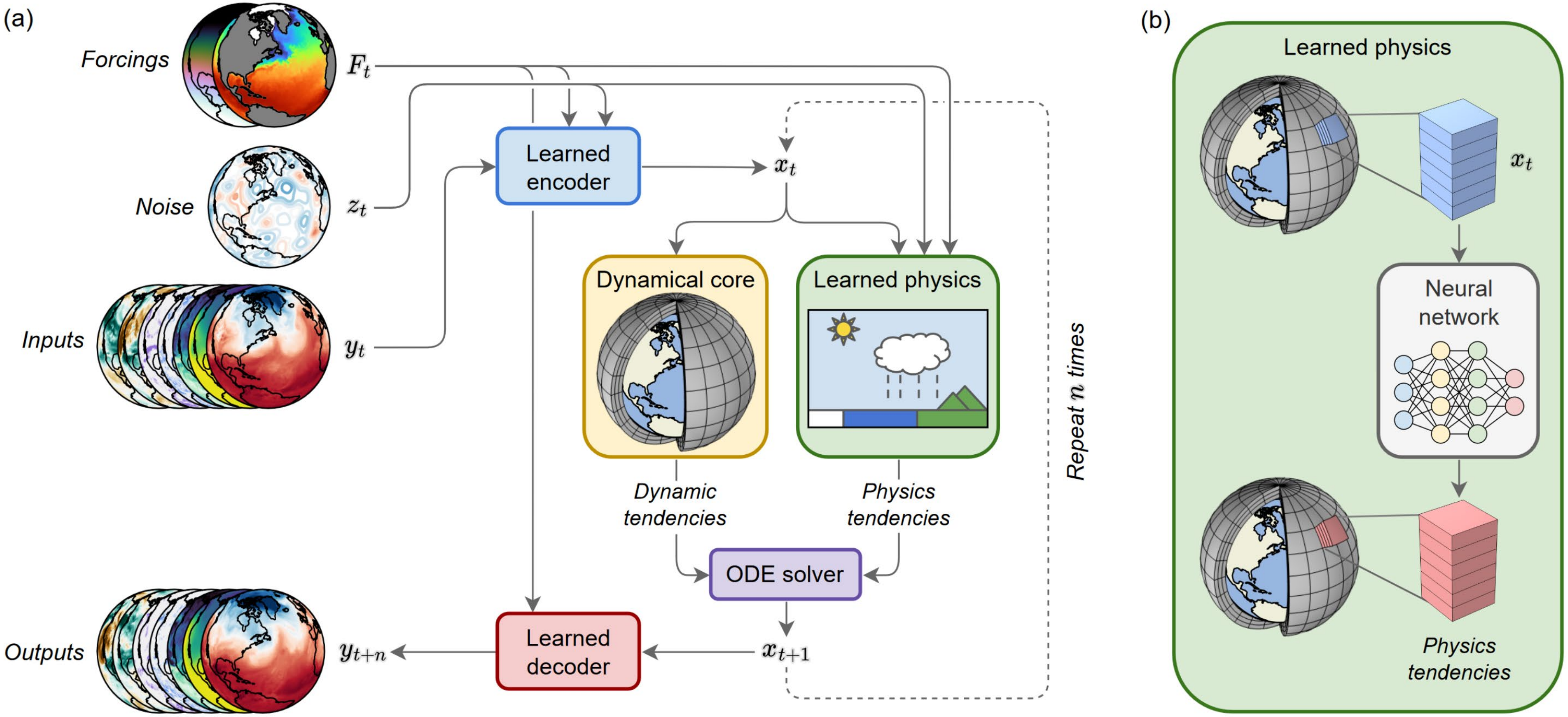
Case study on temperature instability NeuralGCM; after 139 days




Kochkov et al. (Arxiv, 2024)

**Climate time scales; NeuralGCM**

High quality deterministic forecasts up to 10 days, probabilistic forecasts up to 15 days, and stable „weather" on century time scales

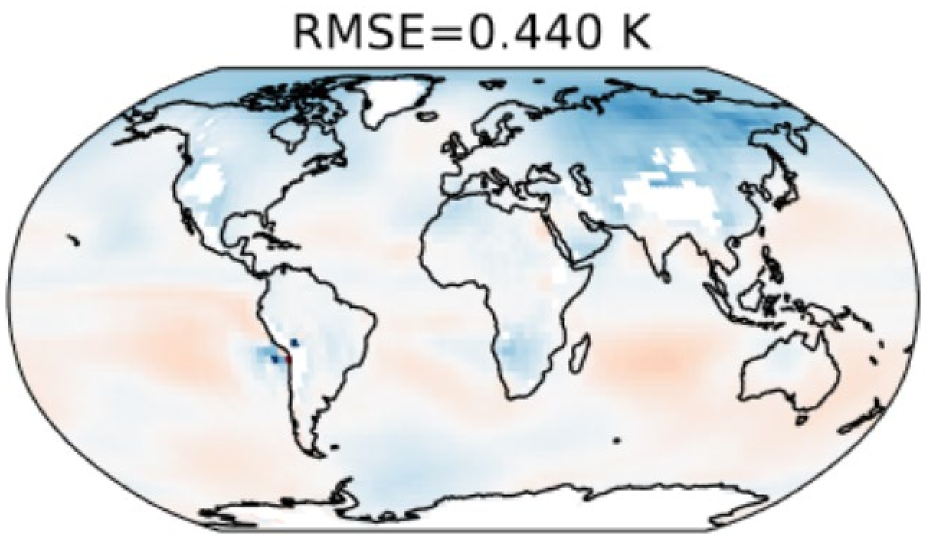



Kochkov et al. (Arxiv, 2024)

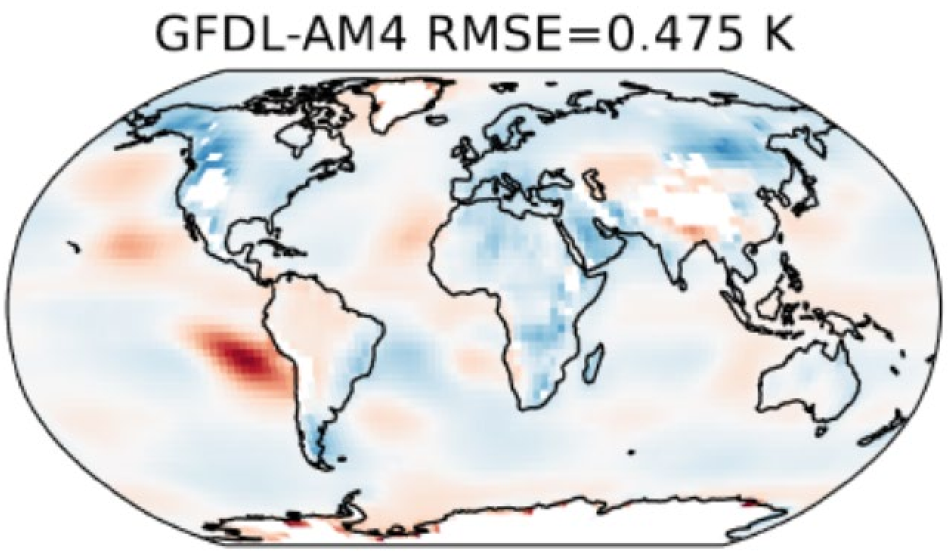
## Climate time scales: Neural GCM achieves reduced bias in climate predictions

### 850 hPa temperature bias averaged 1981-2014

**Worst** of 22 Neural GCM simulations




**Best** of 22 CMIP6 GCM simulations



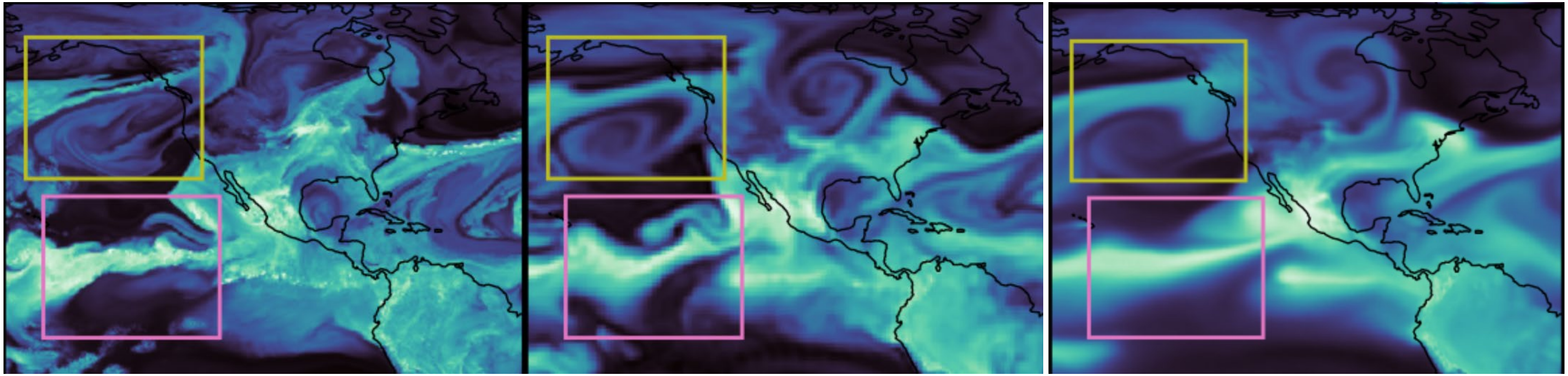

Kochkov et al. (Arxiv, 2024)

## Bluriness

„The data-driven forecast appears smoother than the operational IFS forecast but the level of smoothness does not seem to increase with the forecast lead time, as we might expect when training toward RMSE.“ (Ben Bouallegue et al., 2023)

**Specific humidity, 700 hPa, t+10 days**

IFS HRES (0.1°) NeuralGCM (0.7°) Graphcast (0.25°)



Kochkov et al. ([Arxiv, 2024](#))

## Ensemble modelling

Classical approach: perturb initial conditions and parametrisations (also adopted by FuXi)
Deep learning: Use generative models

**Example:** AtmoRep (Lessig et al., 2023) → talk by Ilaria Luise tomorrow

Inherent probabilistic formulation (and probabilistic loss)

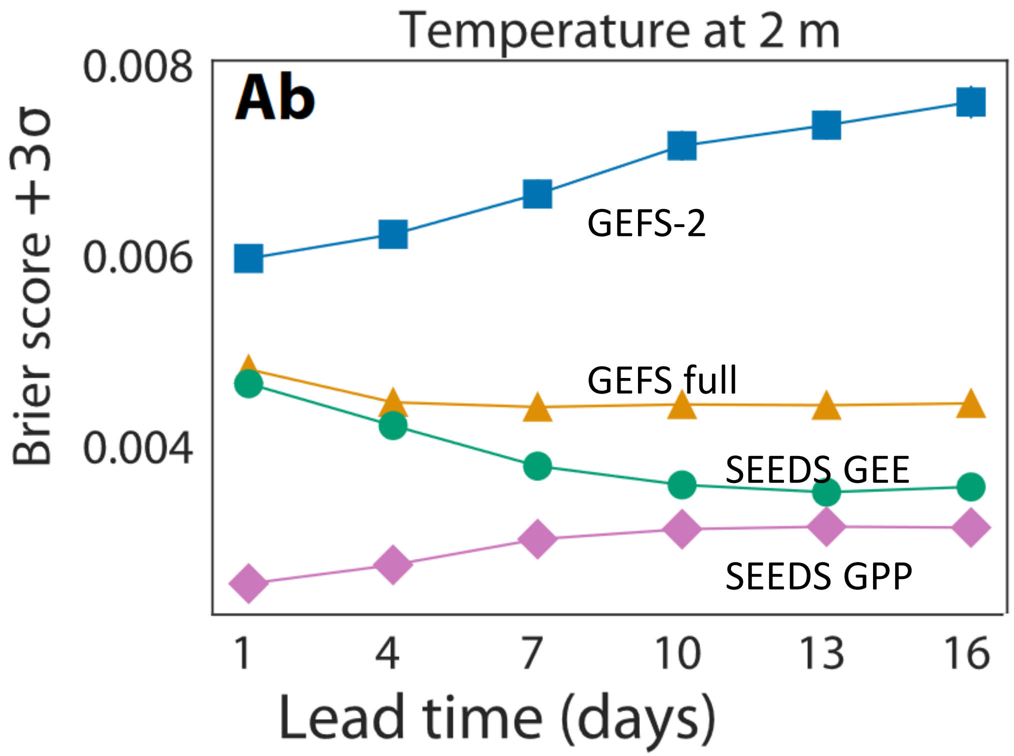
$$p_\theta(y|x,\alpha)$$

approx. initial condition
$x = (\zeta, \mu, T, z, \cdots)$

auxiliary information
(e.g. global time)

## Ensemble modelling

Google‘s Scalable Ensemble Envelope Diffusion Sampler (SEEDS)




Carver et al. (Science, 2024)

See also:

J. Leinonen, U. Hamann, D. Nerini, U. Germann, G. Franch, Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification. arXiv:2304.12891 [physics.ao-ph] (25 April 2023).
Z. Gao, X. Shi, B. Han, H. Wang, X. Jin, D. Maddix, Y. Zhu, M. Li, Y. Wang, PreDiff: Precipitation nowcasting with latent diffusion models. arXiv:2307.10422 [cs.LG] (28 December 2023).
H. Addison, E. Kendon, S. Ravuri, L. Aitchison, P. Watson, Machine learning emulation of a local-scale UK climate model, in *NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning* (Climate Change AI, 2022); www.climatechange.ai/papers/neurips2022/21/paper.pdf.
S. Bassetti, B. Hutchinson, C. Tebaldi, B. Kravitz, DiffESM: Conditional emulation of Earth system models with diffusion models. arXiv:2304.11699 [physics.ao-ph] (23 April 2023).

## Challenges ahead

- Direct use of observations (work in progress)
- Climate scenarios
- Multi-scale models (has been demonstrated → SEEDS)
- Upper atmosphere and tracer transport (work initiated)
- Earth system modeling: ocean, sea ice, land, biogeochemical cycles, atmospheric chemistry

## Upper atmosphere more difficult? Or simply overlooked?

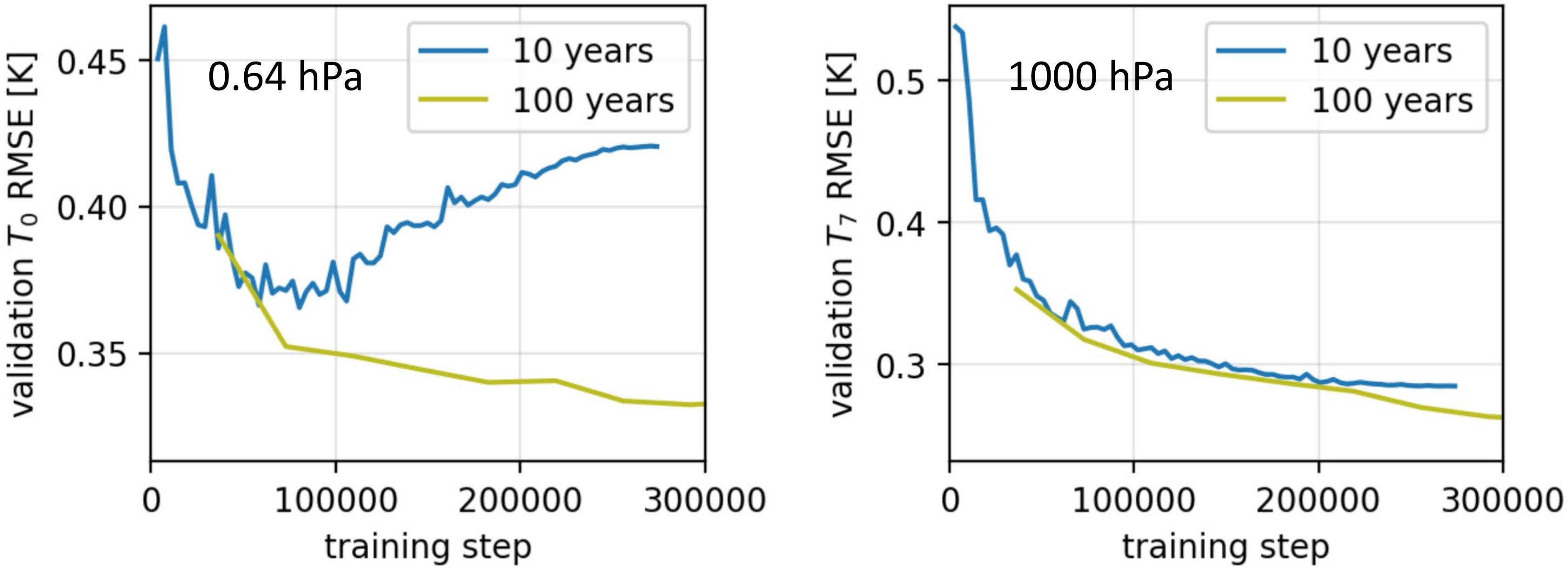


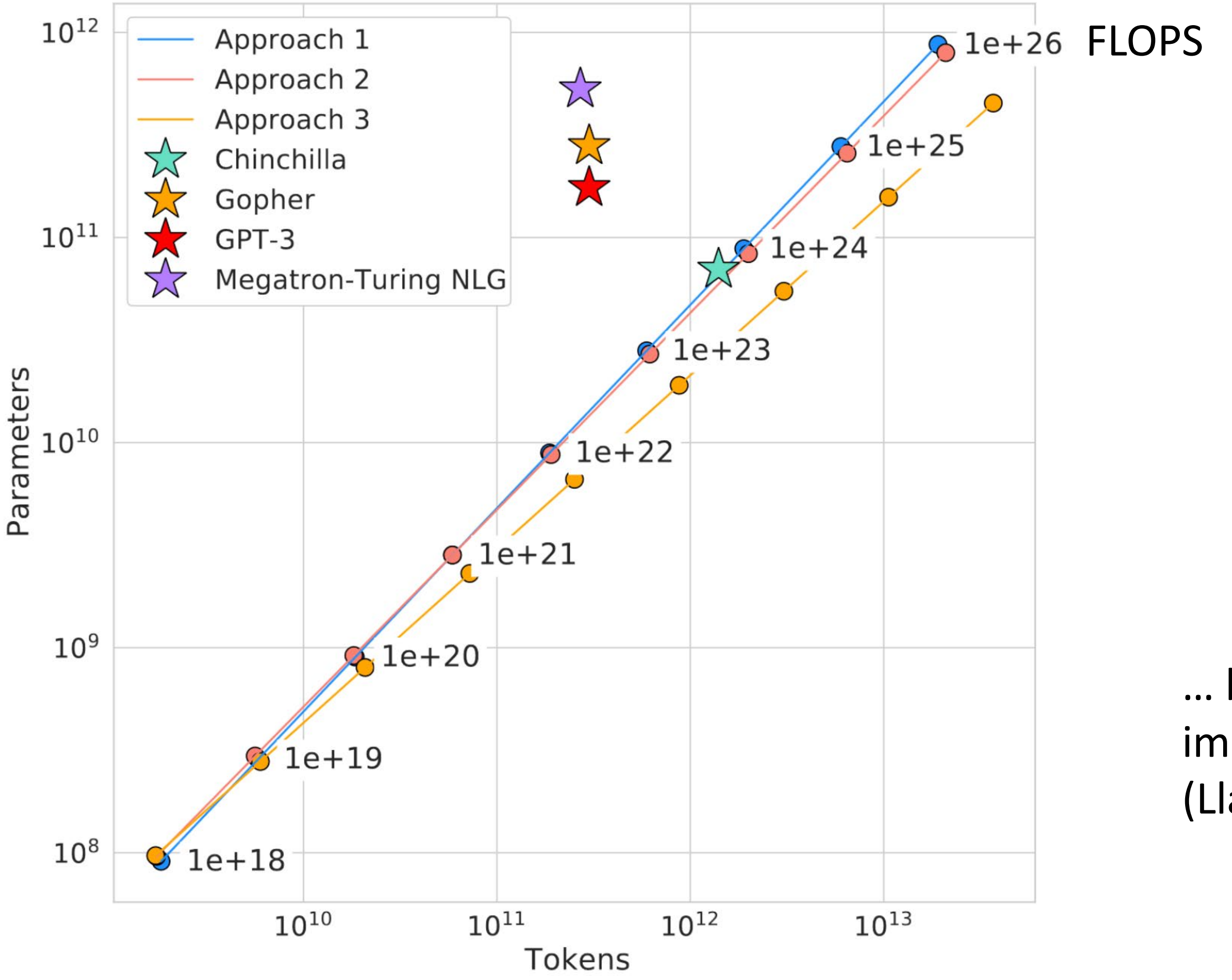

Figure 15: The validation RMSE of (left) $T_0$ and (right) $T_7$ for models trained on (blue) a 10-year dataset and (olive) a 100-year dataset.

Watt-Meyer et al. (Arxiv, 2023)

## Computational costs

„The computational cost of the model is negligible; it has a throughput of 256 ensemble members (at 2° resolution) per 3 min on Google Cloud TPUv3-32 instances and can easily scale to higher throughput by deploying more accelerators. […] Training [of the 114 mio parameter model] takes slightly less than 18 hours on a 2 × 2 × 4 TPUv4 cluster. “ (Carver et al., 2024)

## Scaling laws?




„Chinchilla" optimal training; Hoffmann et al. (Arxiv, 2021)

… but: additional training improves performance! (Llama3; Karpathy posts on X)

# Training strategies (here: large language models)

Deciding on a model architecture

Deciding on a model parallelism strategy

Deciding on the model size

Scaling laws

Trade-off of large language model sizes

Issues and questions related to tensor precision

What to chose between fp32, fp16, bf16

Mixed-precisions for optimizers, weights, specifics modules

How to finetune and integrate a model trained in a precision in another precision

Selecting training hyper-parameters and model initializations

Learning rate and learning rate schedules

Questions on batch size

Maximizing throughput

Avoiding, recovering from and understanding instabilities

Detecting instabilities early

Training tips to reduce instabilities

Issues with data and data processing

Debugging software and hardware failures

Tips on what metrics to follow during the training

From https://github.com/huggingface/large_language_model_training_playbook

## Conclusions

- AI models will become the standard tools for weather forecasting at all scales
- AI models can produce excellent deterministic forecasts and quantify uncertainties well
- Some models (all?) exhibit some physical inconsistencies; these can likely be healed
- Larger models exhibit good robustness and (limited?) capabilities for extrapolation
- Tendency towards very large foundation models less clear than in language area (scaling laws?)
- Incorporating Earth system feedbacks on all time scales is probably the largest challenge ahead


Registration open until 31 July (student rate until 15 June)

**Supported by**

CESOC CENTER FOR EARTH SYSTEM OBSERVATION AND COMPUTATIONAL ANALYSIS · ECMWF · UNIVERSITÄT BONN · TRA MODELLING · JÜLICH Forschungszentrum · UNIVERSITY OF COLOGNE