Incorporating data-driven methods to simulate the weather and climate

Hristo G. Chipilski

22 March 2024

FSU/DSC Machine Learning Seminar

Outline

1. Traditional numerical weather prediction: Brief introduction to the standard procedure of generating weather forecasts using physics-based models.

2. AI-based weather prediction: Reviewing some of the latest advances in AI to solve the weather forecasting problem.

3. Discussion: Where do we stand today and what is coming next? Highlighting some of my ongoing work in this (new to me) area.

4. Code: Resources for running your own AI weather models and a notebook illustrating the key workflow for one state-of-the-art architecture (FourCastNet).

The governing equations

- Atmospheric evolution described by a set of PDEs.

$$\frac{d\mathbf{v}}{dt} = -\alpha \nabla p - \nabla \phi + \mathbf{F} - 2\boldsymbol{\Omega} \times \mathbf{v} \tag{2.1.19}$$

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot (\rho \mathbf{v}) \tag{2.1.20}$$

$$p\alpha = RT \tag{2.1.21}$$

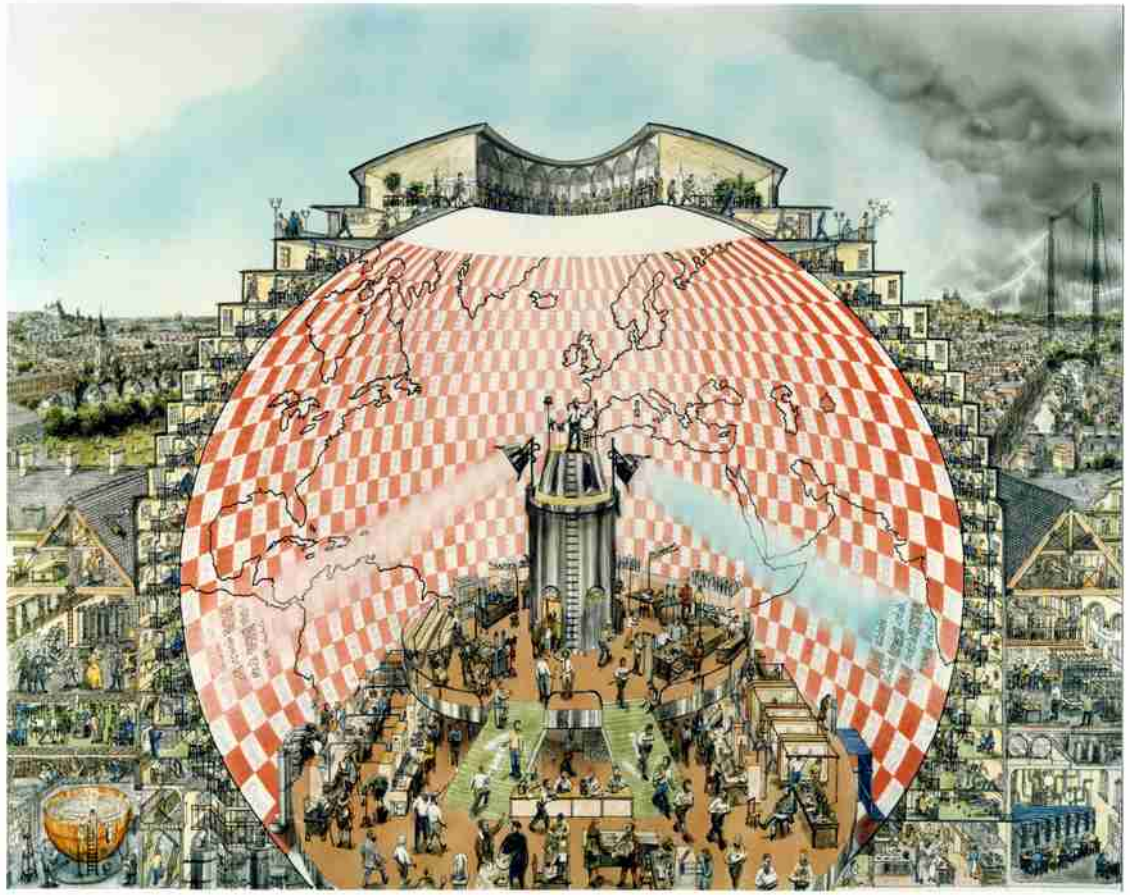
$$Q = C_p \frac{dT}{dt} - \alpha \frac{dp}{dt} \tag{2.1.22}$$

$$\frac{\partial \rho q}{\partial t} = -\nabla \cdot (\rho \mathbf{v} q) + \rho(E - C) \tag{2.1.23}$$

From Kalnay (2003).

- 7 equations with 7 unknowns.

Lewis Richardson's forecast factory



Stephen Conlin's rendition of Richardson's weather forecast factory described in his 1922 book.

- Divide the atmosphere into a grid (mesh) and use the governing equations to predict the weather.
- A large number of humans (64,000) are busy calculating the future weather in a timely fashion.
- Richardson's ideas were quite visionary as they preceded the development of parallel computing architectures by decades.

The problem with Richardson's forecast

- While the weather forecast factory never materialized, Richardson did in fact go about the exercise of hand-computing a weather forecast over NW Europe.
  - A 6-h forecast took him more than 6 weeks to complete!

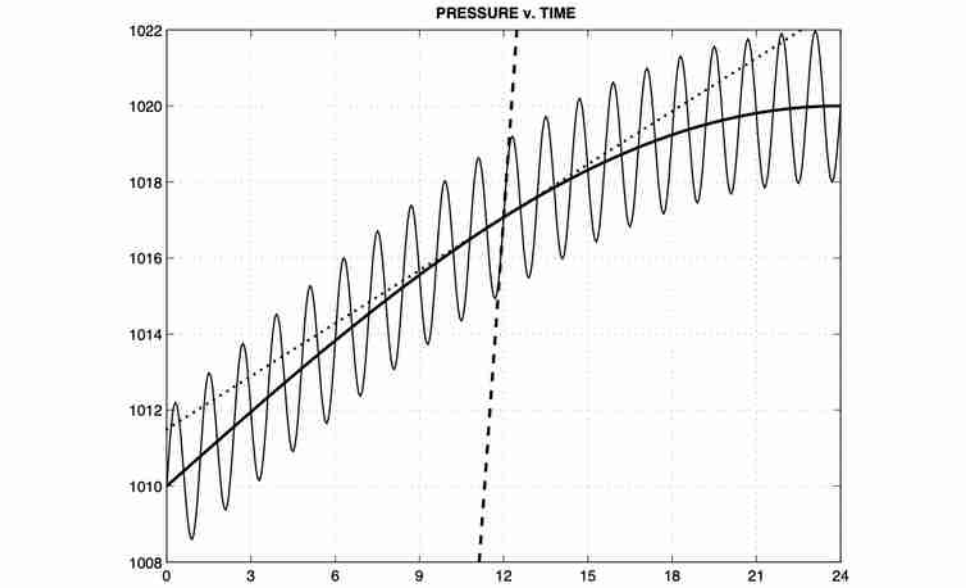


Fig. 1.1. Schematic illustration of pressure variation over a 24 hour period. The thick line is the mean, long-term variation, the thin line is the actual pressure, with high frequency noise. The dotted line shows the rate of change, at 12 hours, of the mean pressure and the dashed line shows the corresponding rate of change of the actual pressure (after Phillips, 1973).

From Lynch (2006).

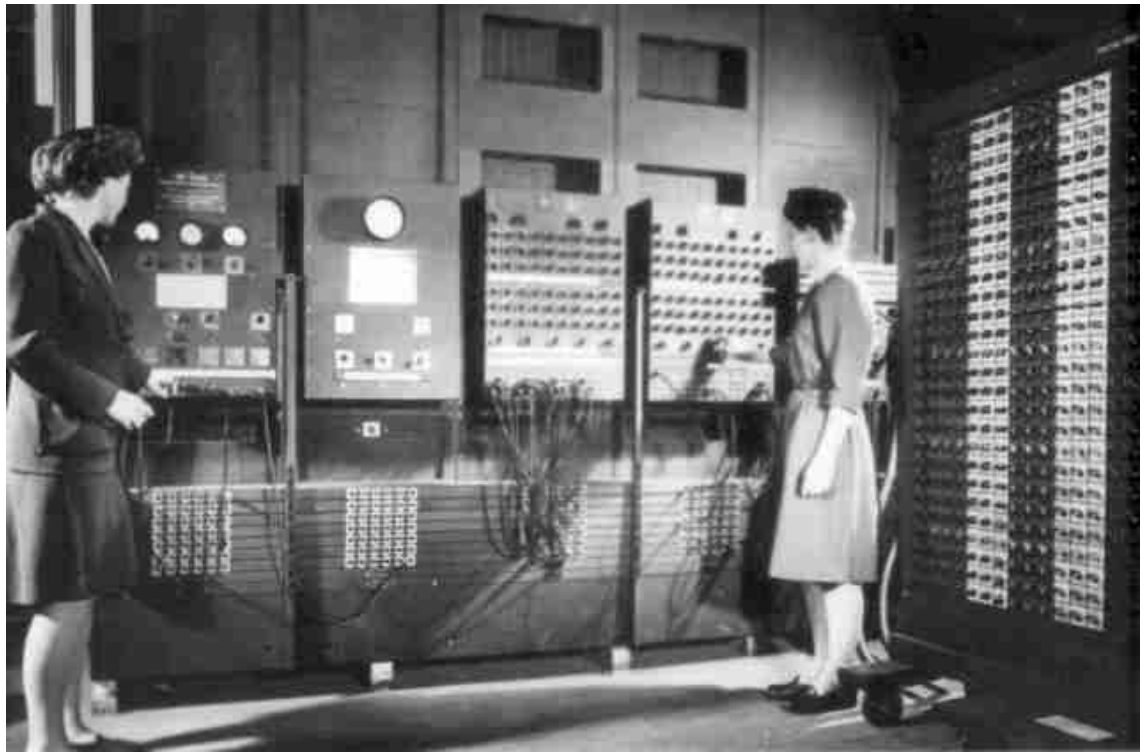
- Unfortunately, Richardson's first forecast wasn't successful – it predicted the atmospheric pressure will increase by 145 hPa for a period of 6h.
- What went wrong? Richardson extrapolated the instantaneous pressure change and assumed it remains constant over a long time period.
  - Neglects the ability of the atmosphere to rapidly adjust to changes.

ENIAC and the first successful NWP forecast

- The first successful NWP forecast had to wait for another 3 decades or so, but eventually happened thanks to the seminar work of Charney, Fjørtoft and von Neumann in 1950 (paper).

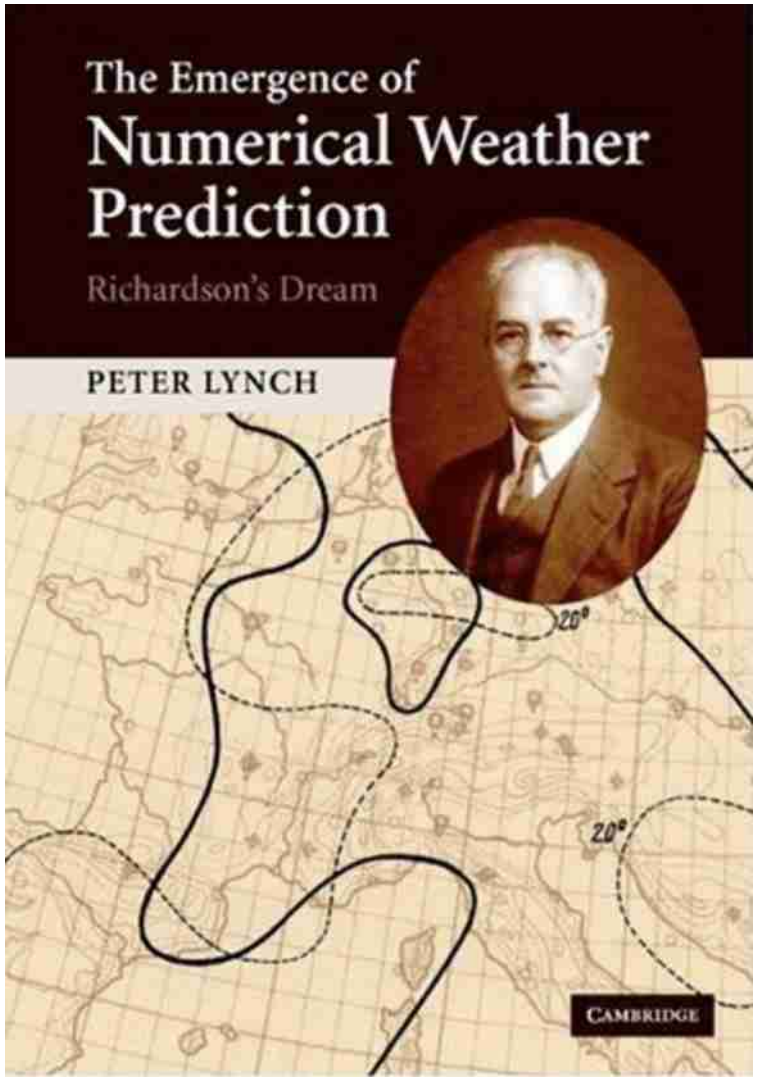
- The simplified (barotropic) dynamics were solved on ENIAC, the first general-purpose electronic computer.
  - 24h forecasts computed in roughly 24h, but most of the time spent on manual operations.

- Although 3 out of the 4 forecasts were worse than the persistence forecasts, this was considered an “enormous scientific advance” that drew a lot of interest.



ENIAC main control panel (Wikipedia).

Textbook recommendation



- Takes us on a journey for how Lewis Richardson's dream of numerical weather prediction was eventually fulfilled.

State-of-the-art NWP nowadays

- At present, physics-based NWP is still the main tool for weather forecasting.

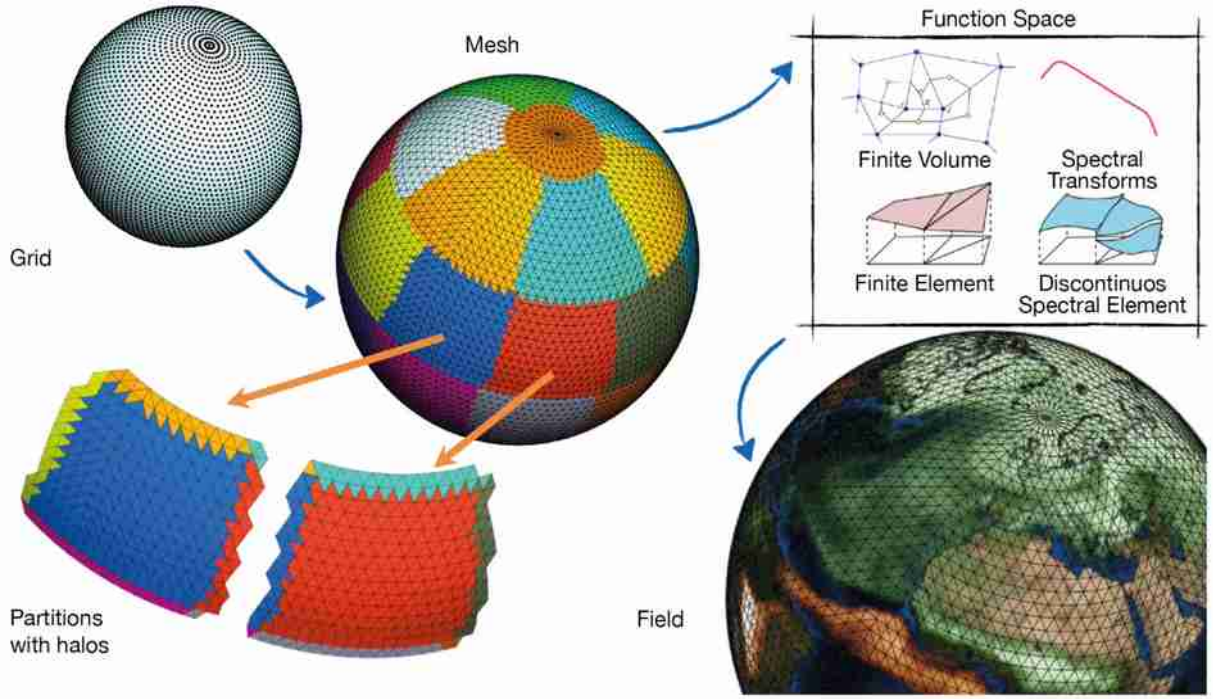
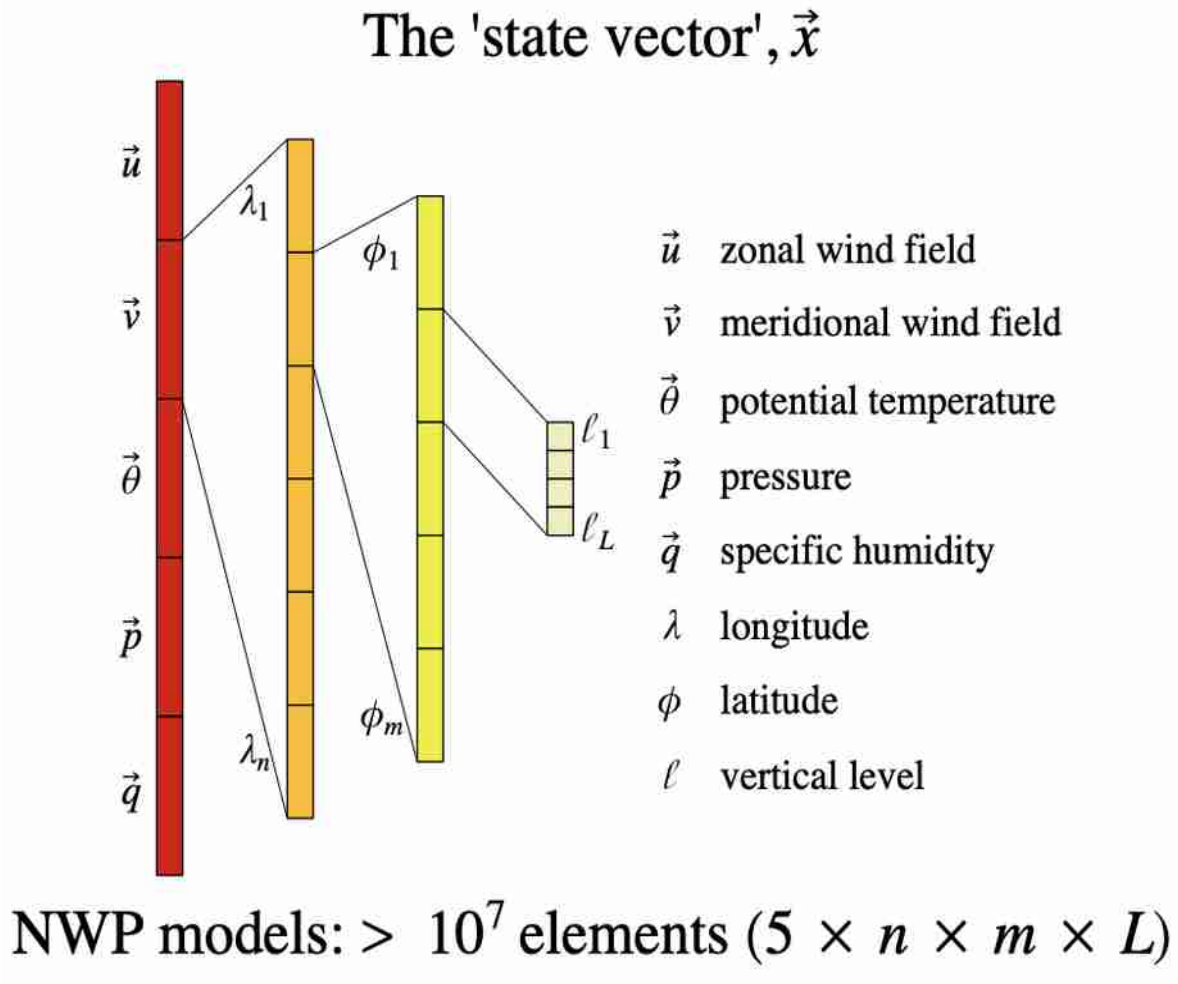


Illustration from recently released Atlas software: OO library for developing flexible models on existing and emerging hardware (including GPU).

- Operational NWP centres such as ECWMF produce and disseminate weather forecasts daily at a resolution < 10km on a global scale.
- Highly sophisticated numerical schemes.
  - ECMWF' IFS uses 2-time-level, semi-implicit and semi-Lagrangian time integrator with a reduced Gaussian grid in the horizontal and a finite element scheme in the vertical. Equations integrated both in physical and spectral space.

Big data problem

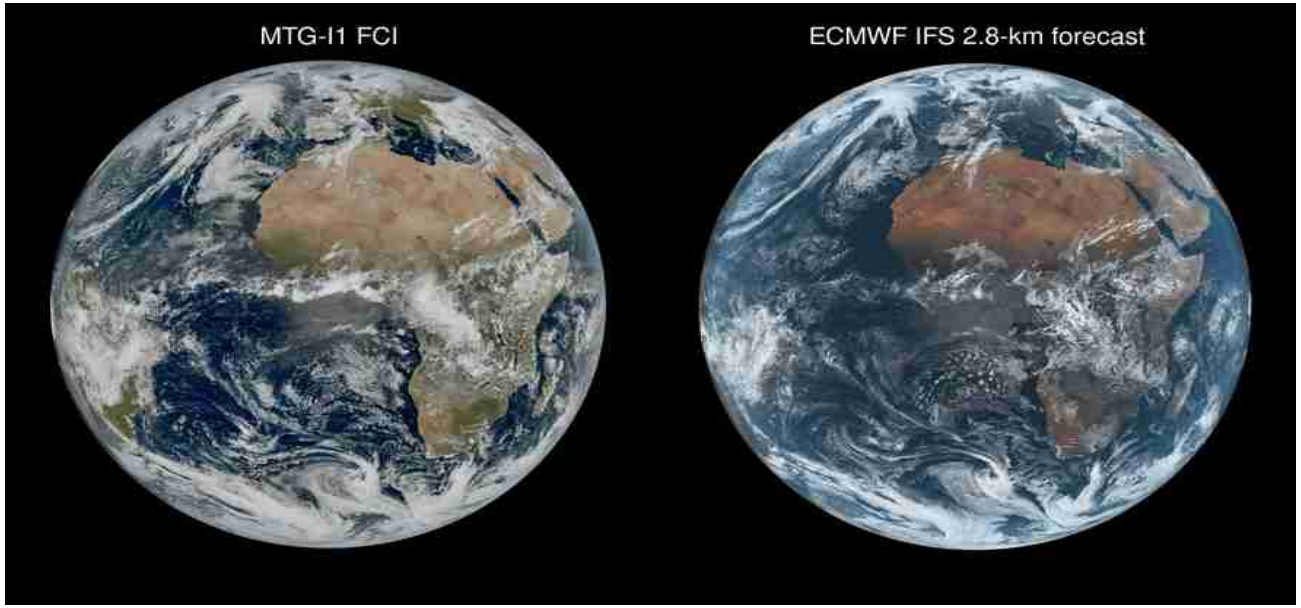


Courtesy of Ross Bannister (2006).

- While the underlying physics are well understood, it is the size of the NWP problem which makes it extremely challenging.
- The degrees of freedom in a typical operational NWP model can reach as high as 10^9 (1 billion!).
- Similarly, the number of observations can exceed 10^6 (1 million).

Digital twins of the Earth

- Simulations have become so realistic that major efforts are underway to build so-called digital twins of the atmospheric dynamics.

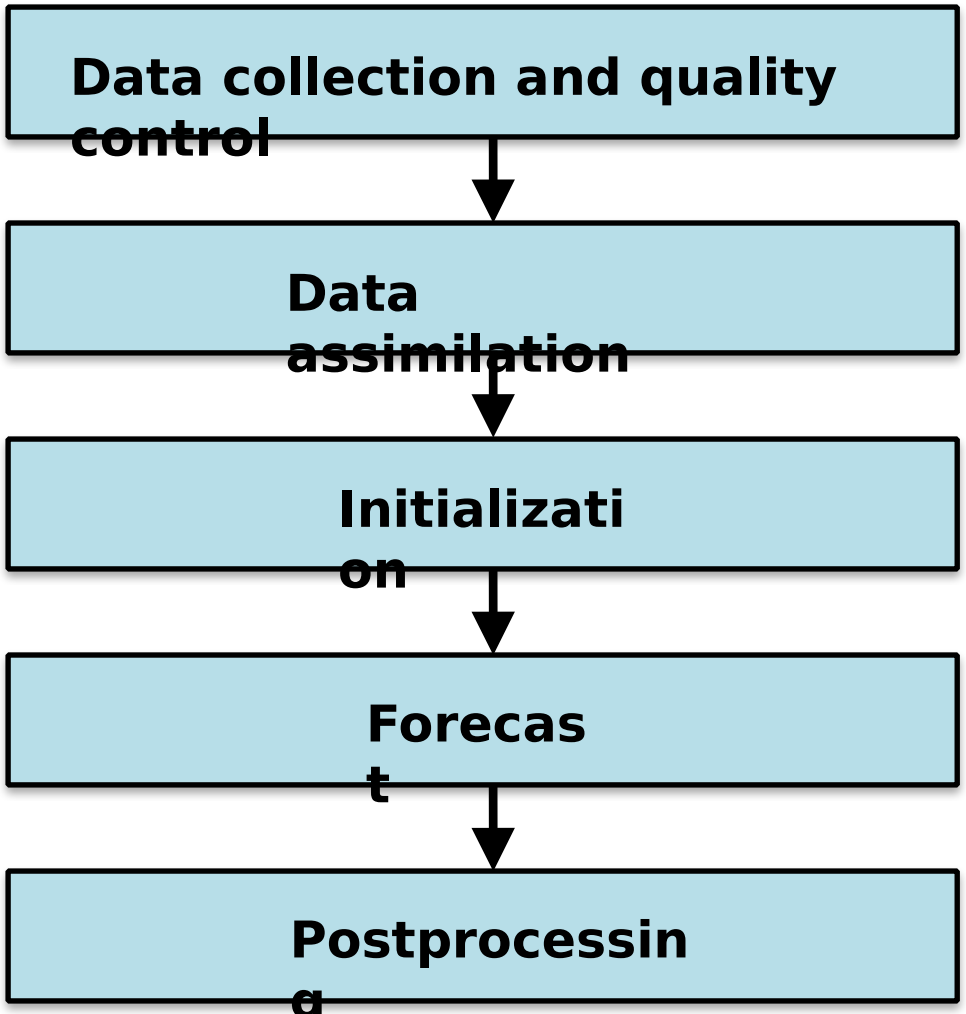


- Broader definition of digital twins: virtual representations of products, people, processes and environments.

- Digital twins of the Earth use a fusion of numerical simulations and observations to create a virtual replica that is indistinguishable from reality.

Left: MeteoSat 3rd Generation Imager (MTF-I1). Right: IFS simulation.

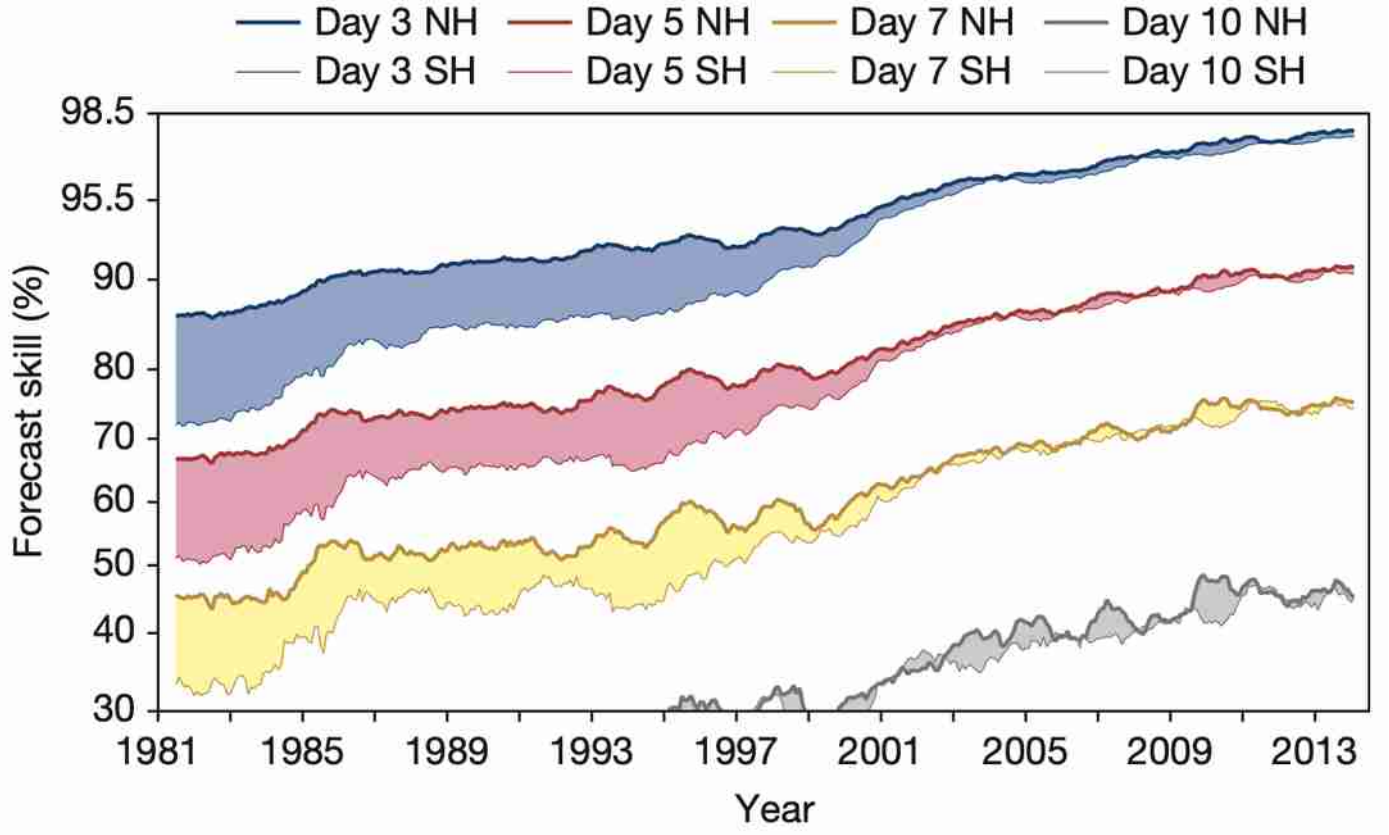
Standard workflow



- Each forecast cycle takes 6h (NCEP in the US) or 12h (ECMWF) to complete.

Graphic: World Meteorological Organization.

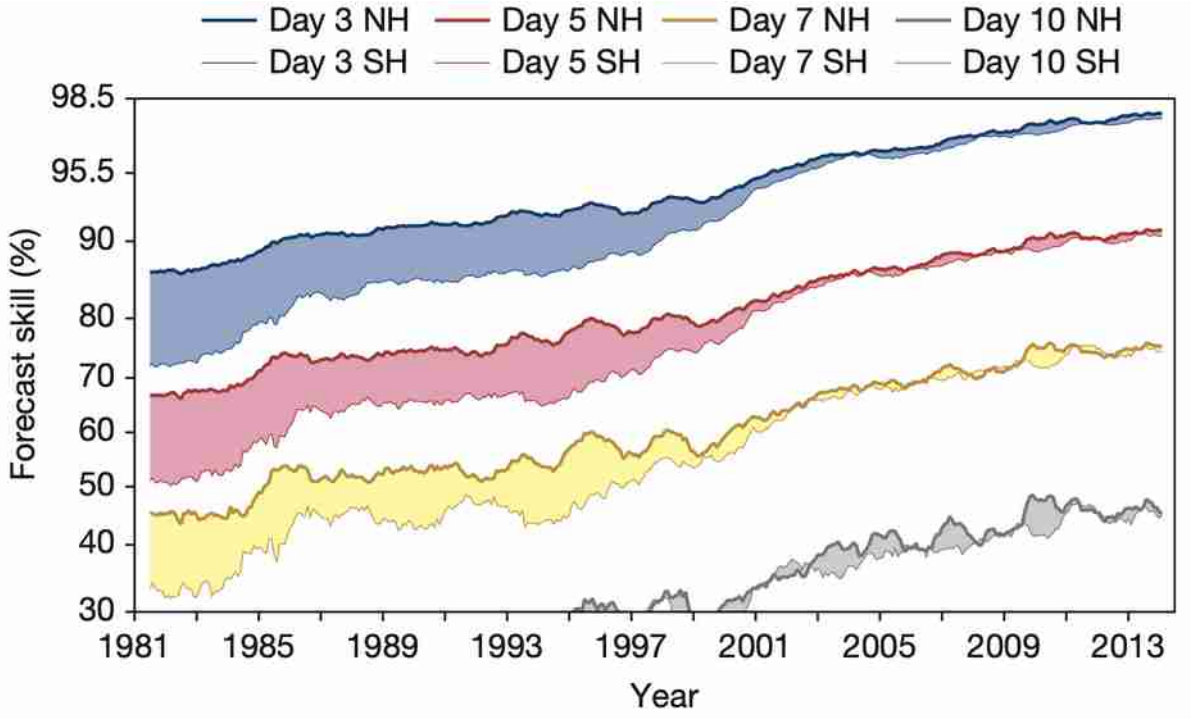
Quiet (first) revolution of NWP

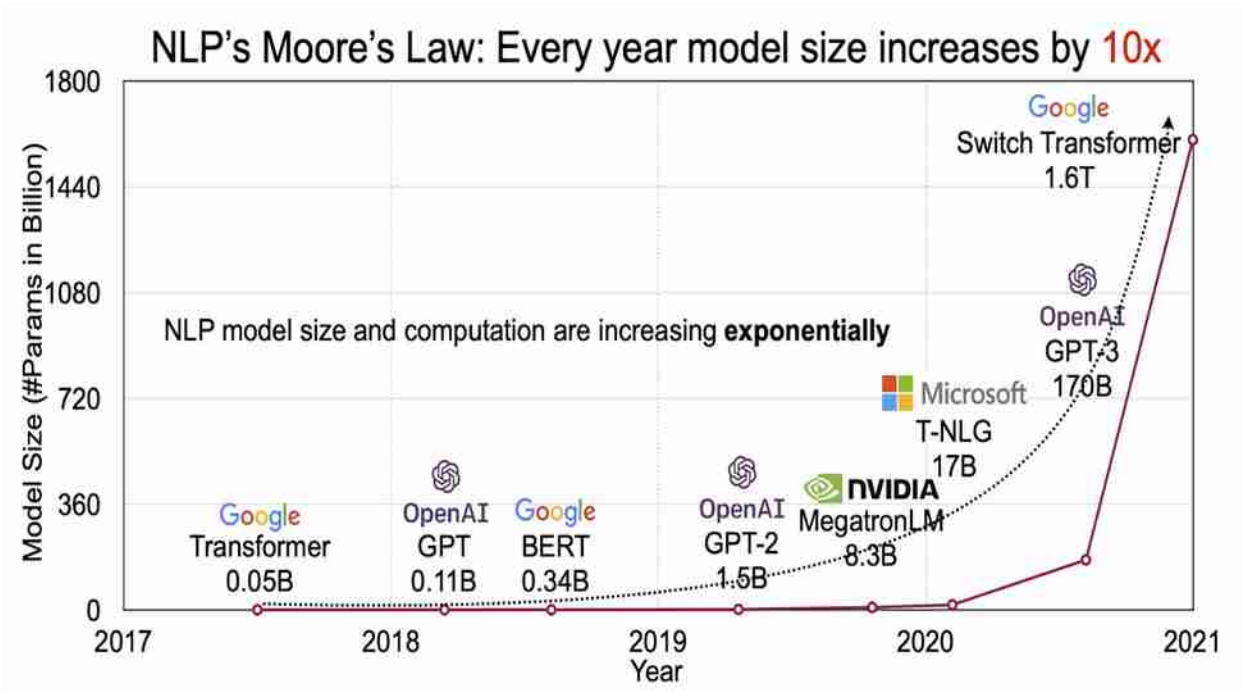


From Bauer et al. (2015).

- NWP advances not the result of a single or a few big discoveries as in other disciplines, but through steady improvements in areas such as:
  - physical parameterizations
  - data assimilation
  - ensemble forecasting
- Note closing of the "skill gap" between NH and SH: mostly due to better exploitation of satellite data in the SH.

Contrasting with rapid AI advances

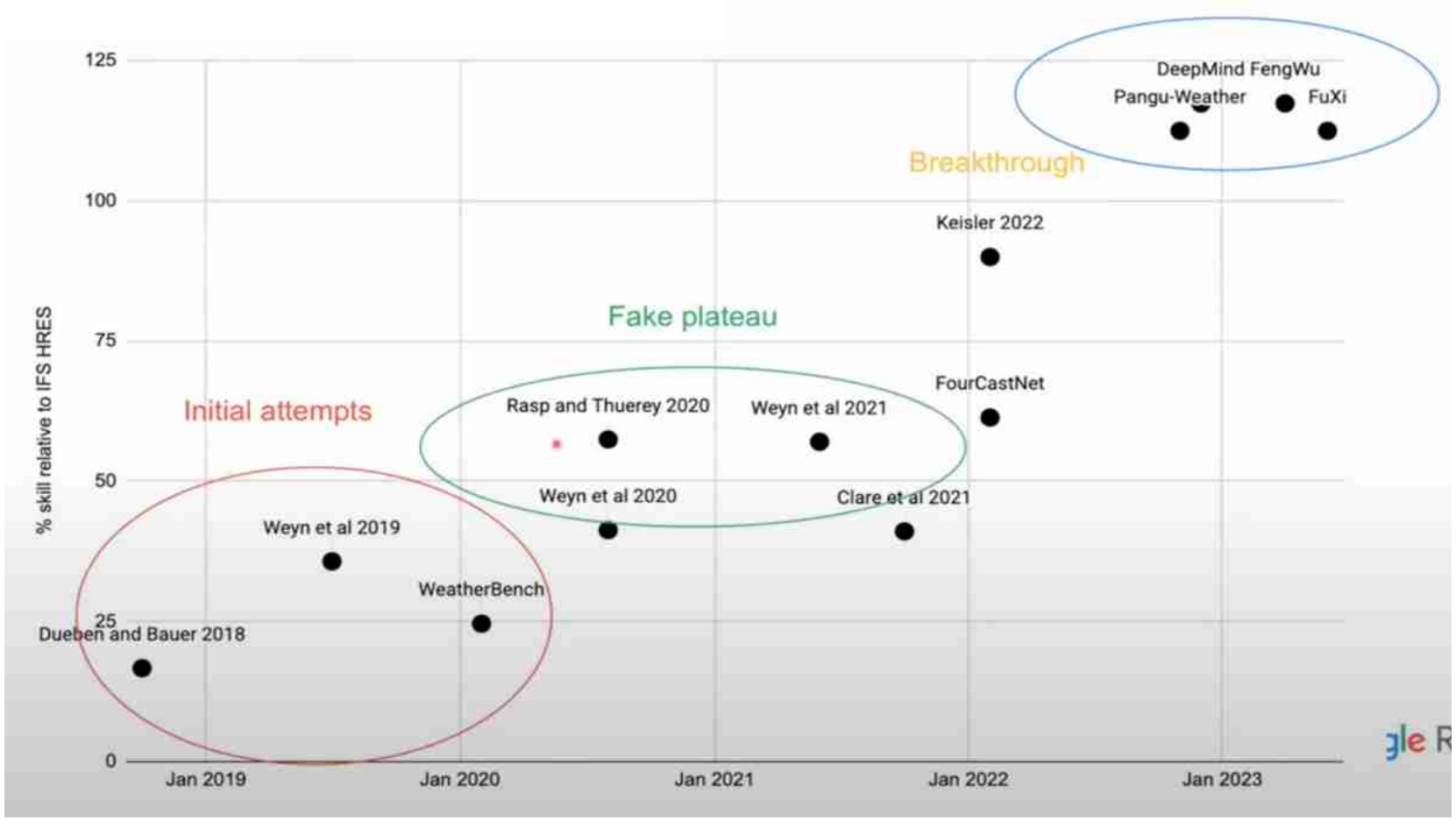




From Rasp (2024)'s ISDA-online talk.

- Size of transformer-based ML models is increasing exponentially: explosive growth compared with the more gradual NWP developments.
- Has the weather and climate enterprise taken advantage of these rapid developments?

AI and the second revolution in weather prediction

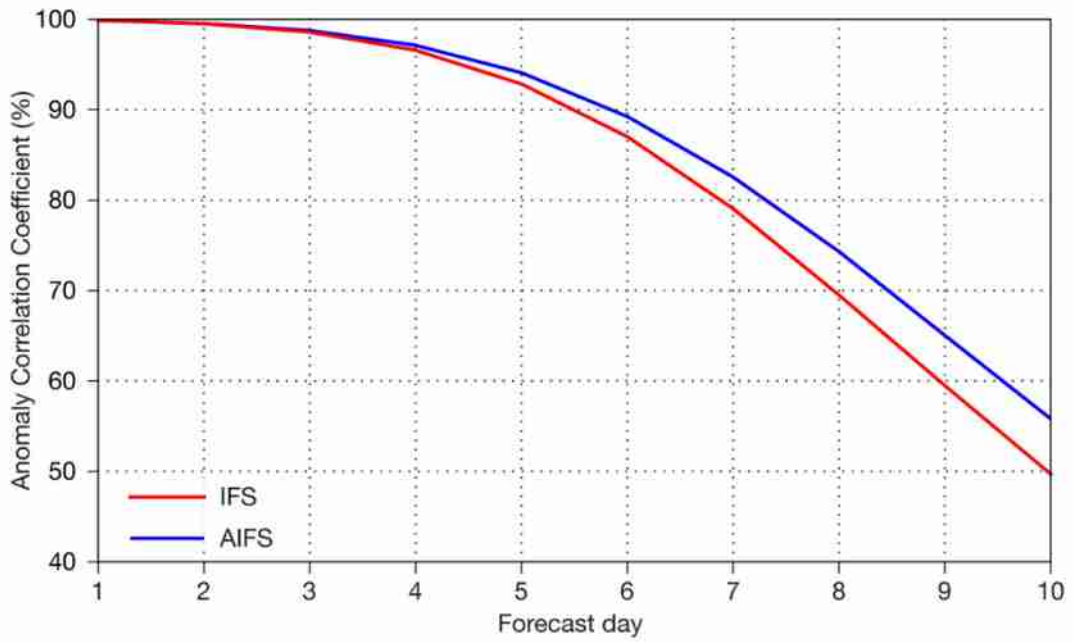


From Rasp (2024)'s ISDA-online talk.

- Pace of progress in data-driven AI much faster than what NWP community is used to.
  - Reflects rapid AI advances.

Towards operational implementation

- ECMWF is now running an experimental AIFS model which can produce 10-day forecasts with 6-hour time steps in ~1min!

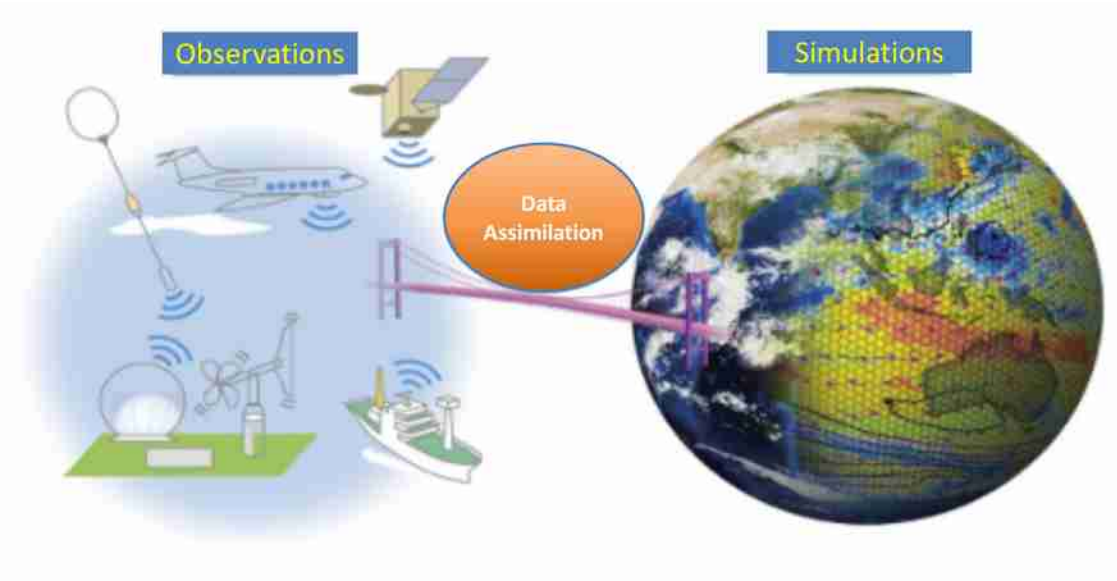


AIFS forecast skill. We show the northern hemisphere Anomaly Correlation Coefficient (ACC) for geopotential height at 500 hPa of IFS forecasts (red, dashed) and AIFS forecasts (blue) for 2022. Higher values indicate better skill. Link to article.

- Experimental system mostly based on graph neural networks and inspired by work from Ryan Keisler and Google DeepMind's GraphCast.
- So far, results are very promising – experimental system outperforms operational model.
  - Resolution of 1° (~110km) still coarse => improvements limited to large-scale parameters.

How are DWP models trained?

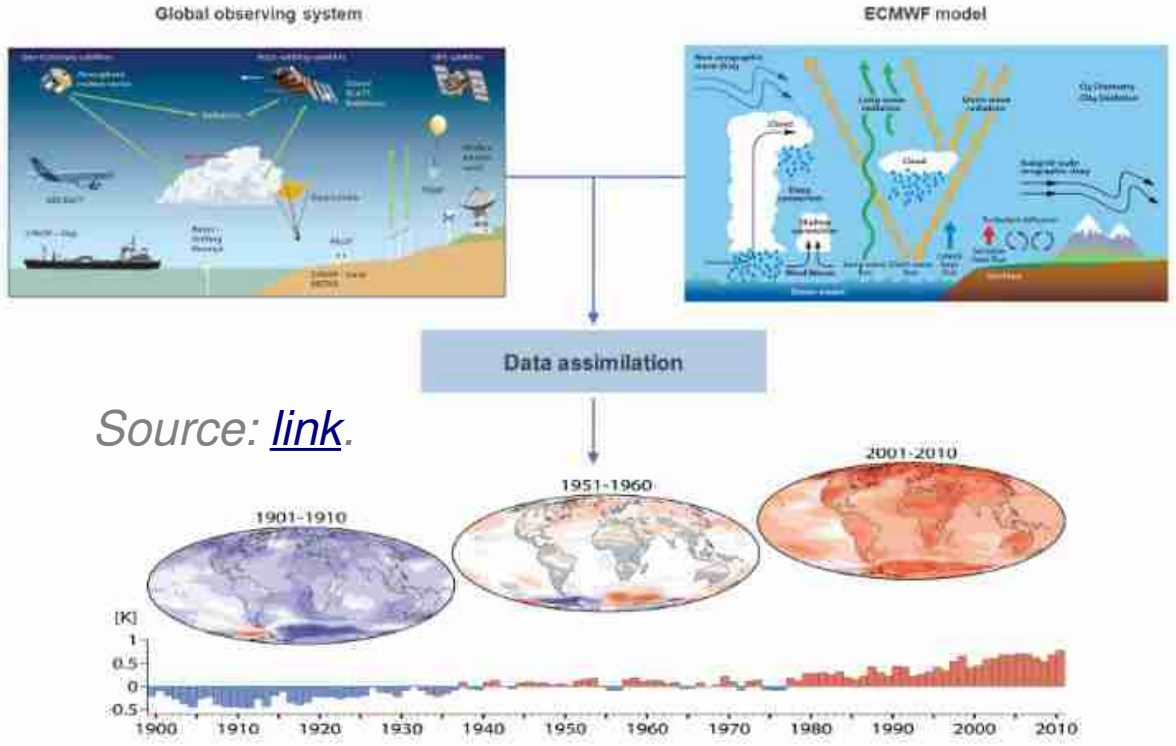
- Before describing more details about the different types of Data-driven Weather Prediction (DWP) models, it it is important to understand how they are trained.
- Training datasets come from the so-called reanalysis products.
- What is reanalysis? The process of reconstructing the Earth's past climate by integrating historical observations and models in a process called data assimilation.
  - The subject of *ISC 4933/5935: Computational Aspects of Data Assimilation.*



Taken from Zhaoxia Pu's website.

ECMWF's ERA5 reanalysis dataset

- Almost all big DWP models use the reanalysis products offered by ECMWF.
- ERA5 is a 5th generation atmospheric reanalysis of the global climate for the period Jan 1940 to present.
  - Replaced the previous ERA-Interim reanalysis dataset.
- Download instructions can be found here.



Source: link.

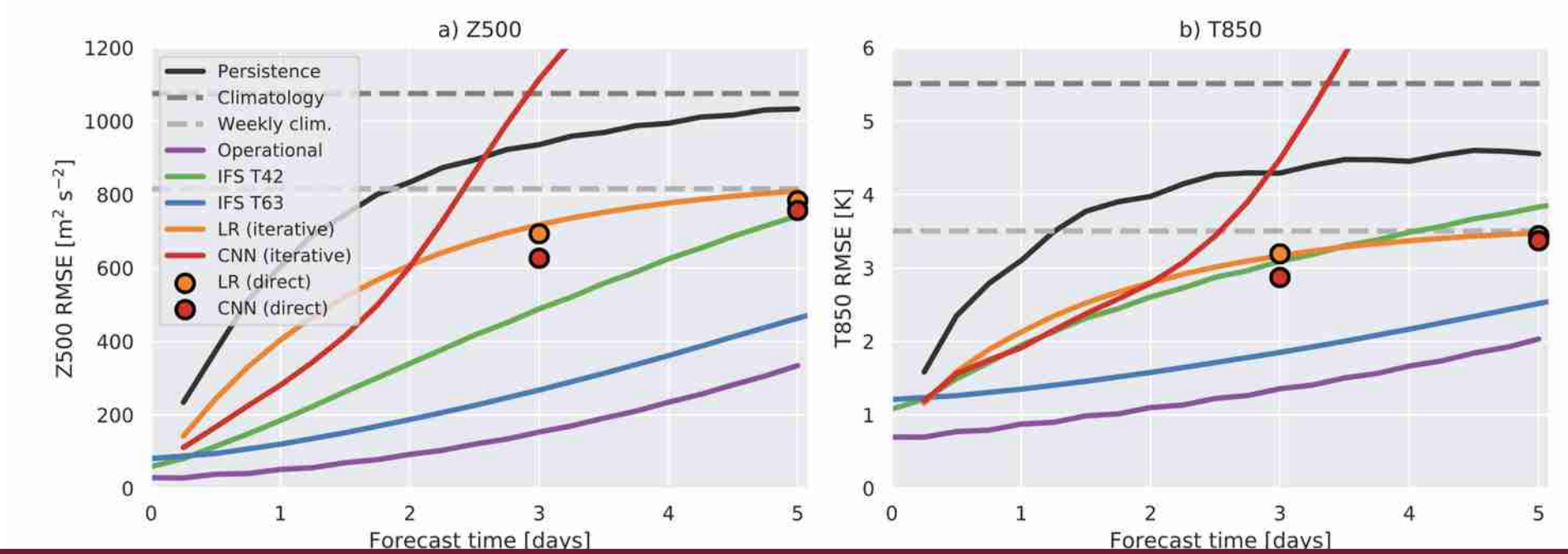
Details

- 1h output frequency
- 31-km spatial resolution
- 137 vertical levels (surface to 80 km)
- uncertainty about all variables (at reduced resolution)

WeatherBench for verification of DWP models

- **Motivation**: The lack of a common dataset and evaluation metrics make inter-comparison between studies difficult.

- **WeatherBench** is a benchmark dataset for evaluating data-driven medium-range weather forecasting.
  - Derived from ERA5 that has been processed to facilitate the use of ML models.



Rasp et al. (2020)

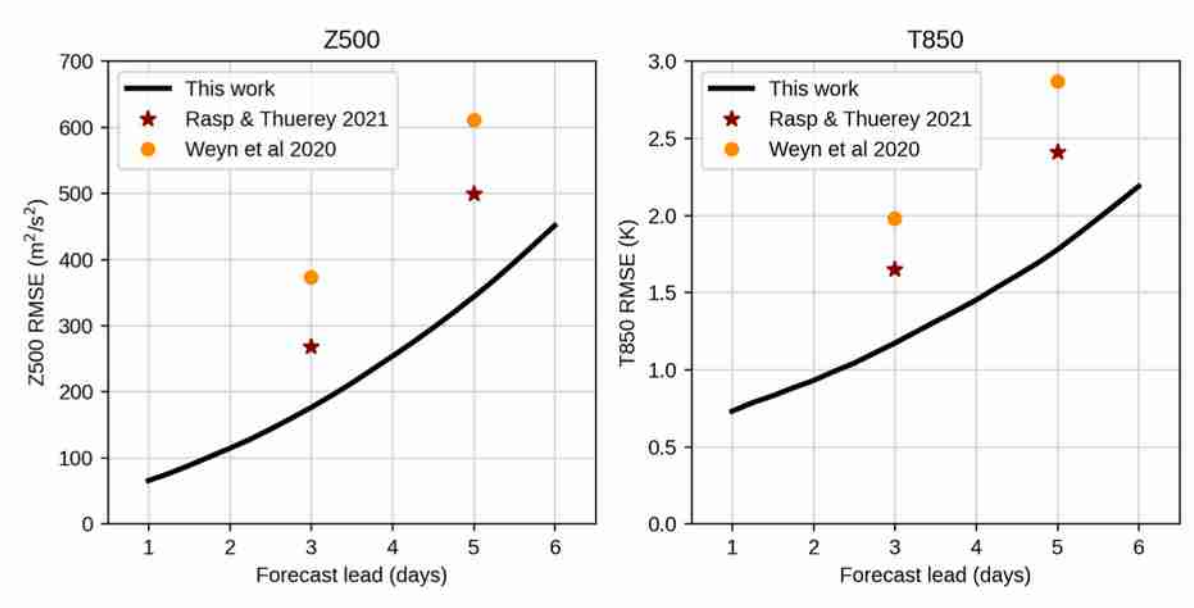
Example: Training and verification datasets

Data availability

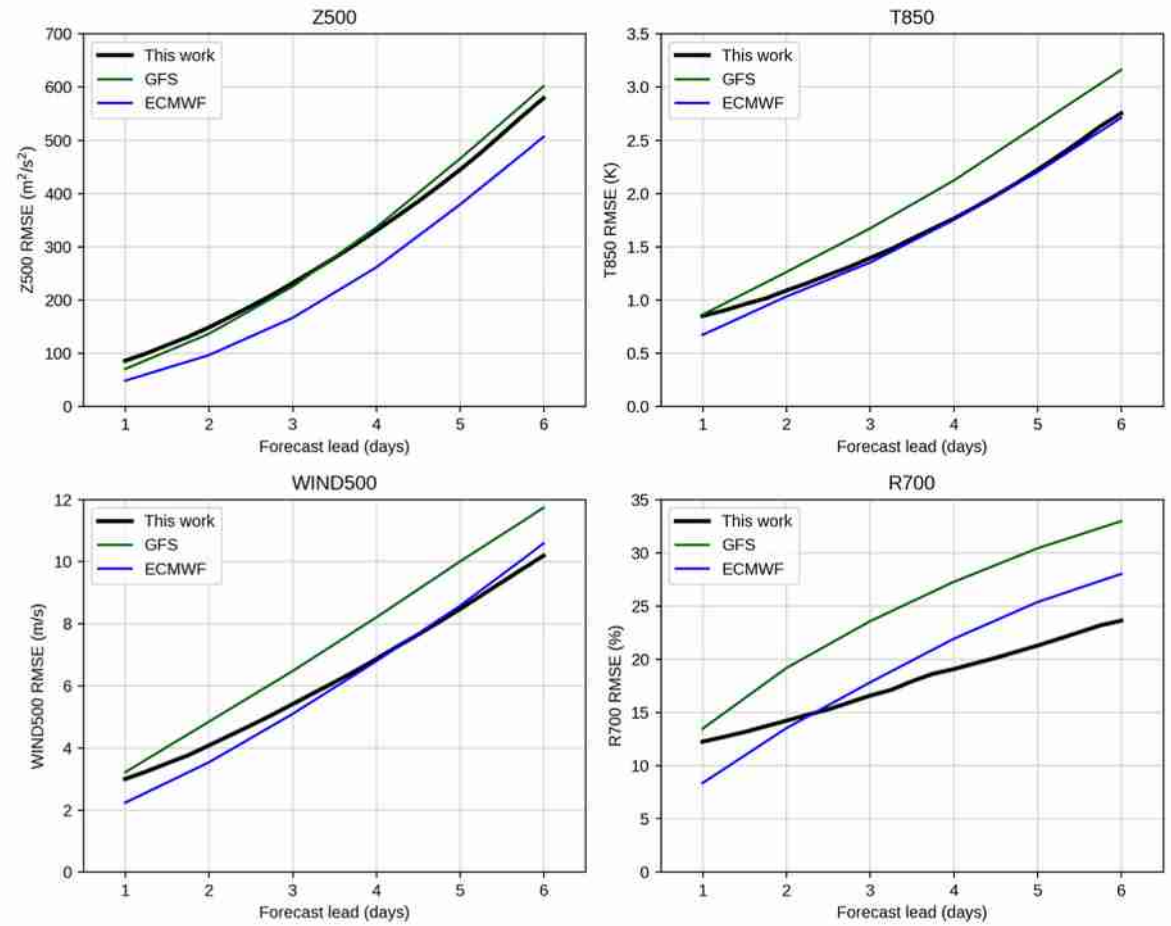
For training and evaluating NeuralGCM models we used publicly available ERA5 dataset [14], originally downloaded from https://cds.climate.copernicus.eu/ and available via Google Cloud Storage in Zarr format at `gs://gcp-public-data-arco-era5/ar/full_37-1h-0p25deg-chunk-1.zarr-v3`. To compare NeuralGCM to operational and data-driven weather models we used forecast datasets distributed as part Weatherbench2 [10] at https://weatherbench2.readthedocs.io/en/latest/data-guide.html, to which we have added NeuralGCM forecasts for 2020. To compare NeuralGCM to atmospheric models in climate settings we used CMIP6 data available at https://catalog.pangeo.io/browse/master/climate/, as well as X-SHiELD [24] outputs available on Google Cloud storage in a "requester pays" bucket at `gs://ai2cm-public-requester-pays/C3072-to-C384-res-diagnostics`. The Radiosonde Observation Correction using Reanalyses (RAOBCORE) V1.9 that was used as reference tropical temperature trends was downloaded from https://webdata.wolke.img.univie.ac.at/haimberger/v1.9/.

Ryan Keisler's GNN model (1)

- Improves upon previous DWP models.
- First AI model to achieve comparable performance to operational, full-resolution NWP models from GFS and ECWMF on a 1 deg scale.

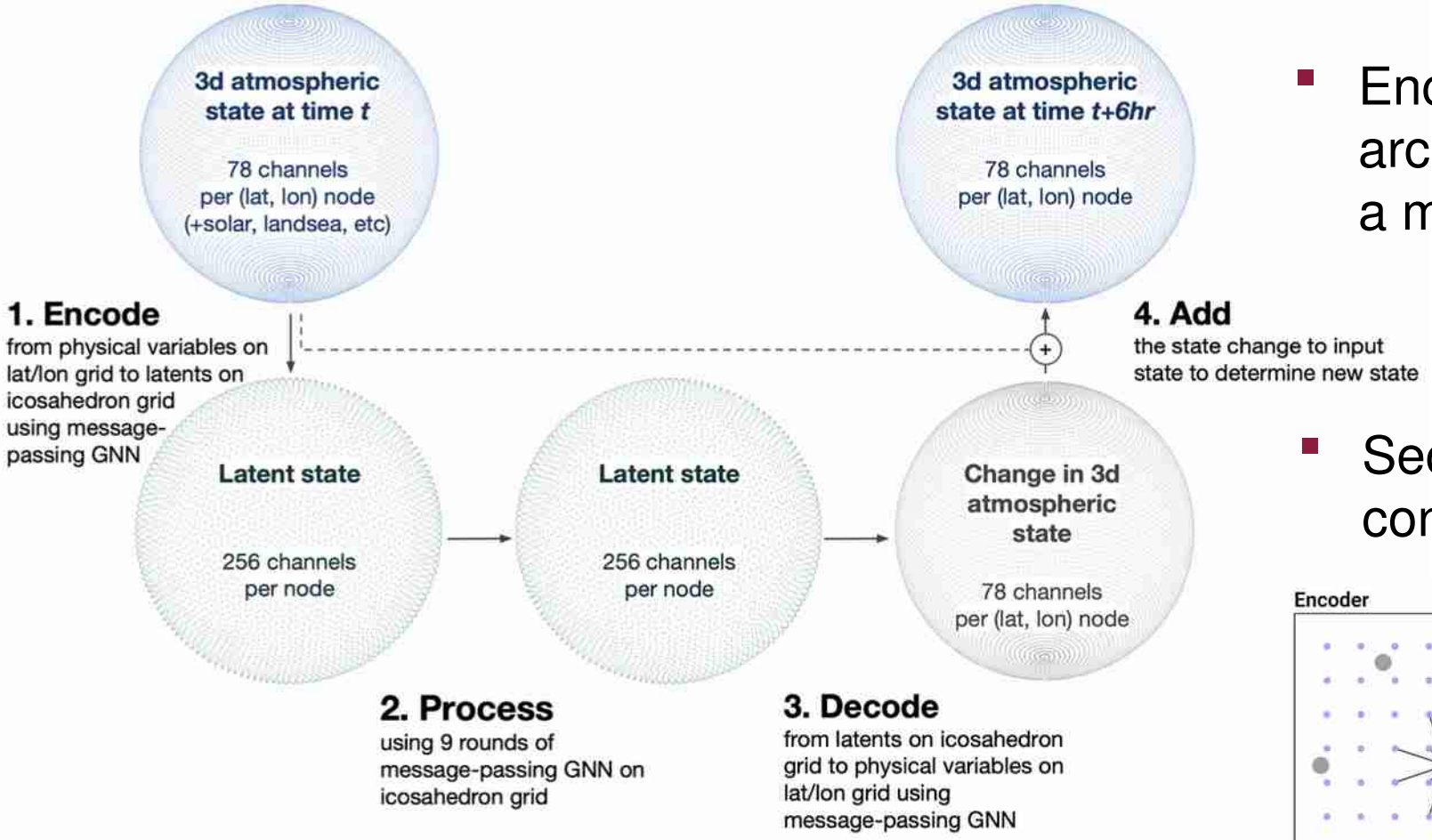


Keisler (2022)



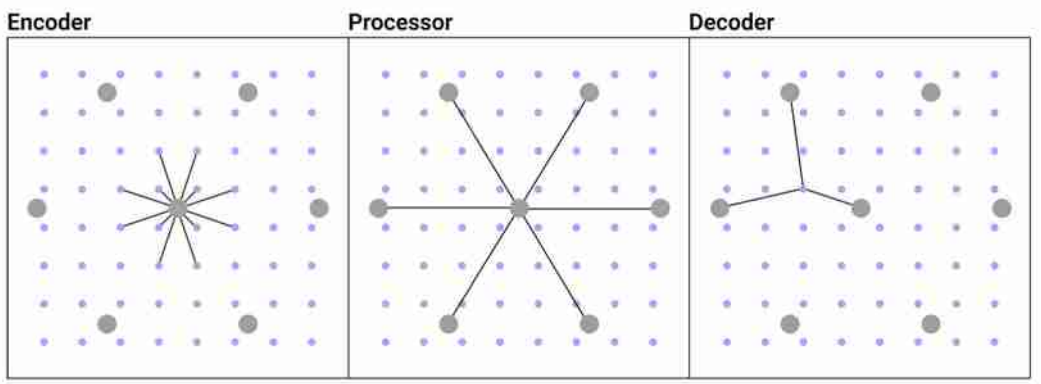
Keisler (2022)

Ryan Keisler's GNN model (2)

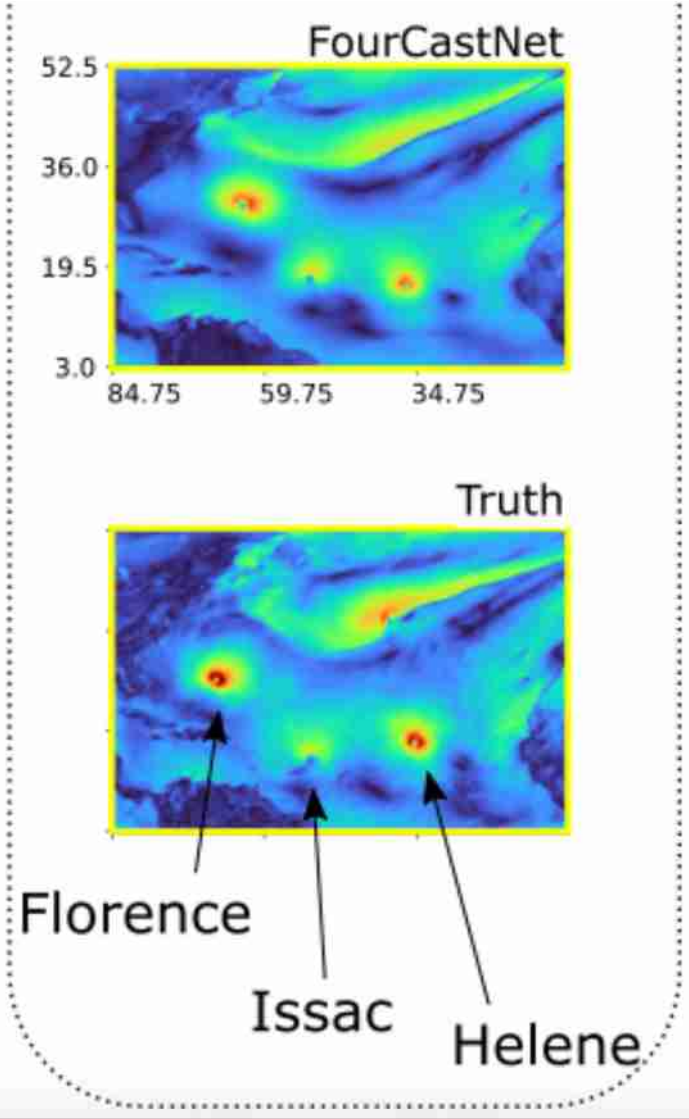


Keisler (2022)

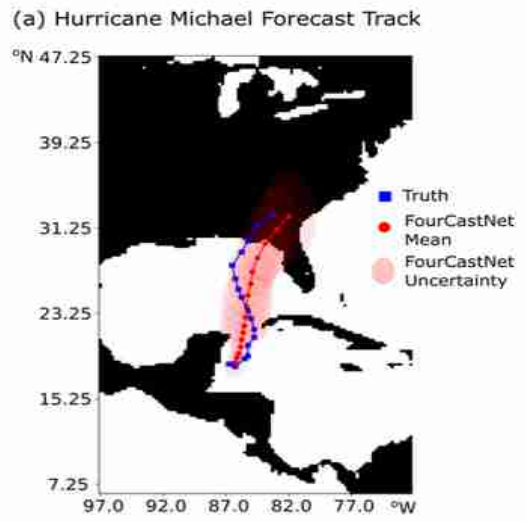
- Encode-process-decode architecture, each implemented via a message-passing GNN.
- Second figure shows local graph connectivity for each component.



FourCastNet (1)

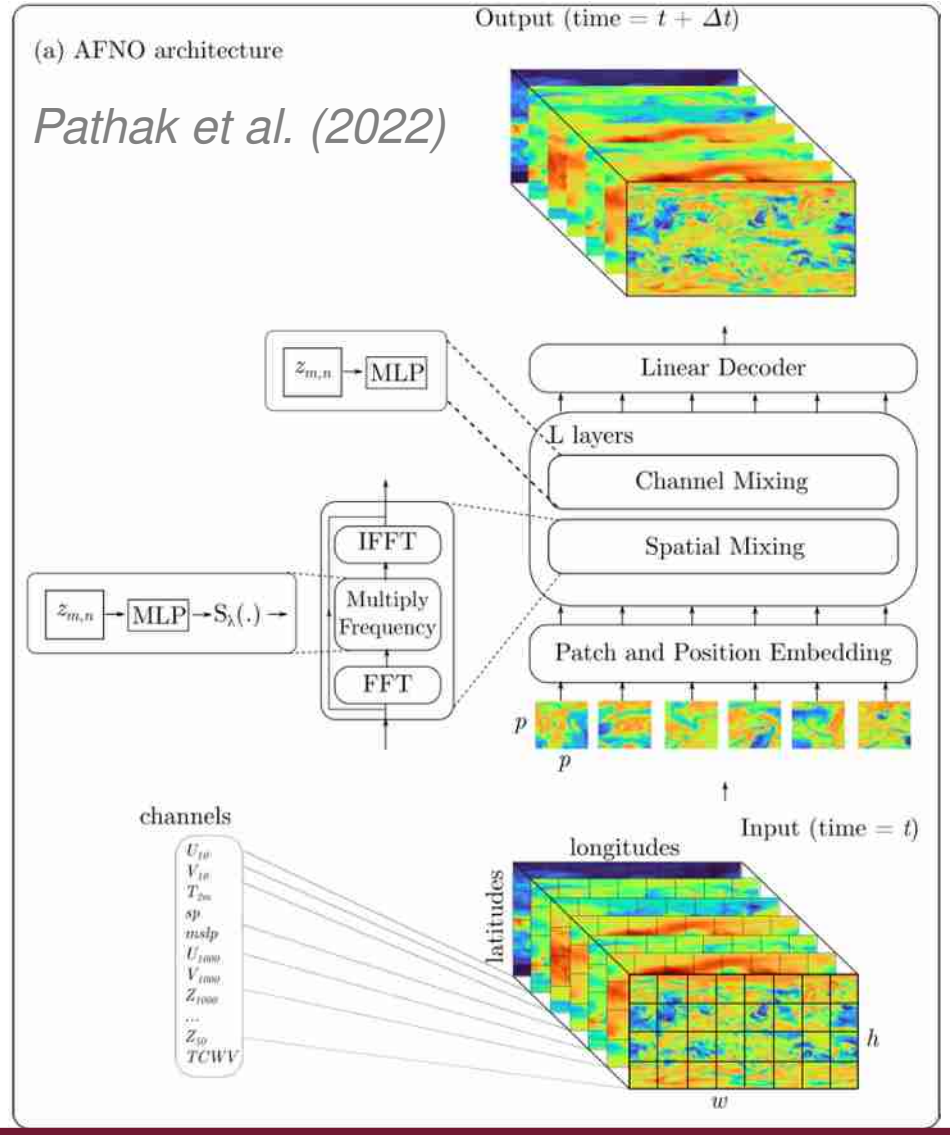


- Released in Feb 2022 by NVIDIA.
- Perhaps the most impressive feature of this model its high resolution (0.25 deg or just less than 30 km) enabled by the use of operator learning architectures.
  - 8 times greater resolution than SOTA DWP models.
- Impressive speedups: 45,000 times on a node-hour basis.
- One of the first architectures to emphasize the importance of probabilistic forecasting (more on that later).



Both images taken from Pathak et al. (2022)

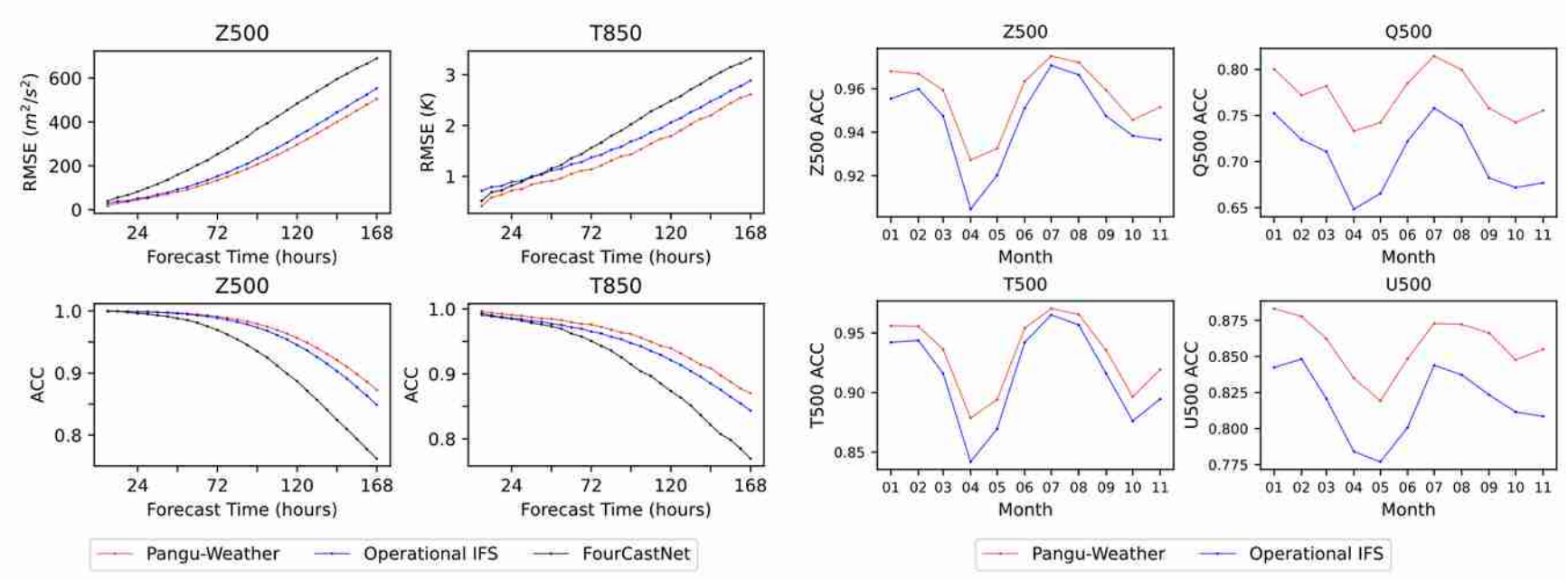
FourCastNet (2)

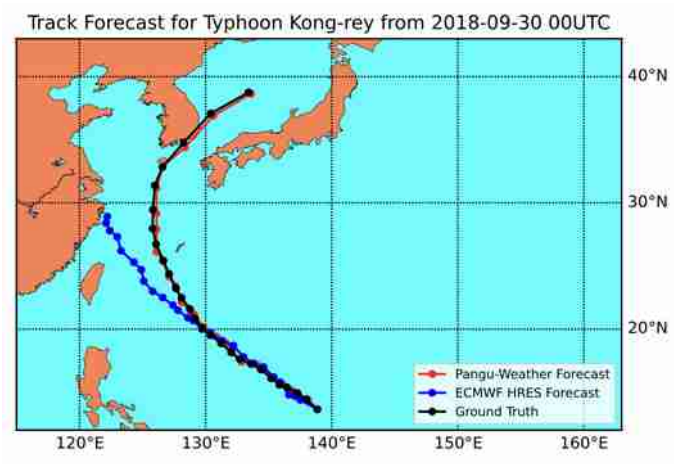


- **Architecture**: Adaptive Fourier Neural Operator (AFNO) using a Fourier transform-based token-mixing scheme with a vision transformer.
  - Input frame divided into patches.
  - Each patch embedded in a higher dimensional space with a large number of latent channels.
  - Position embedding carried out to form a sequence of tokens.
  - Tokens mixed spatially in the Fourier domain.
  - For each token, the latent channels are mixed.
  - Process repeated in L layers.
  - Linear decoder applied to get state at next time step.
- Additional fine-tuning done but not shown here.

Pangu-Weather (1)

- Released in Nov 2022 by Huawei.
- Comparable resolution to FourCastNet (0.25 deg). Training multiple models with different lead times.
- **Major achievement**: First time AI-based method to outperform traditional NWP for all variables and time ranges (1 hour to 1 week).





Bi et al. (2022)

Pangu-Weather (2)

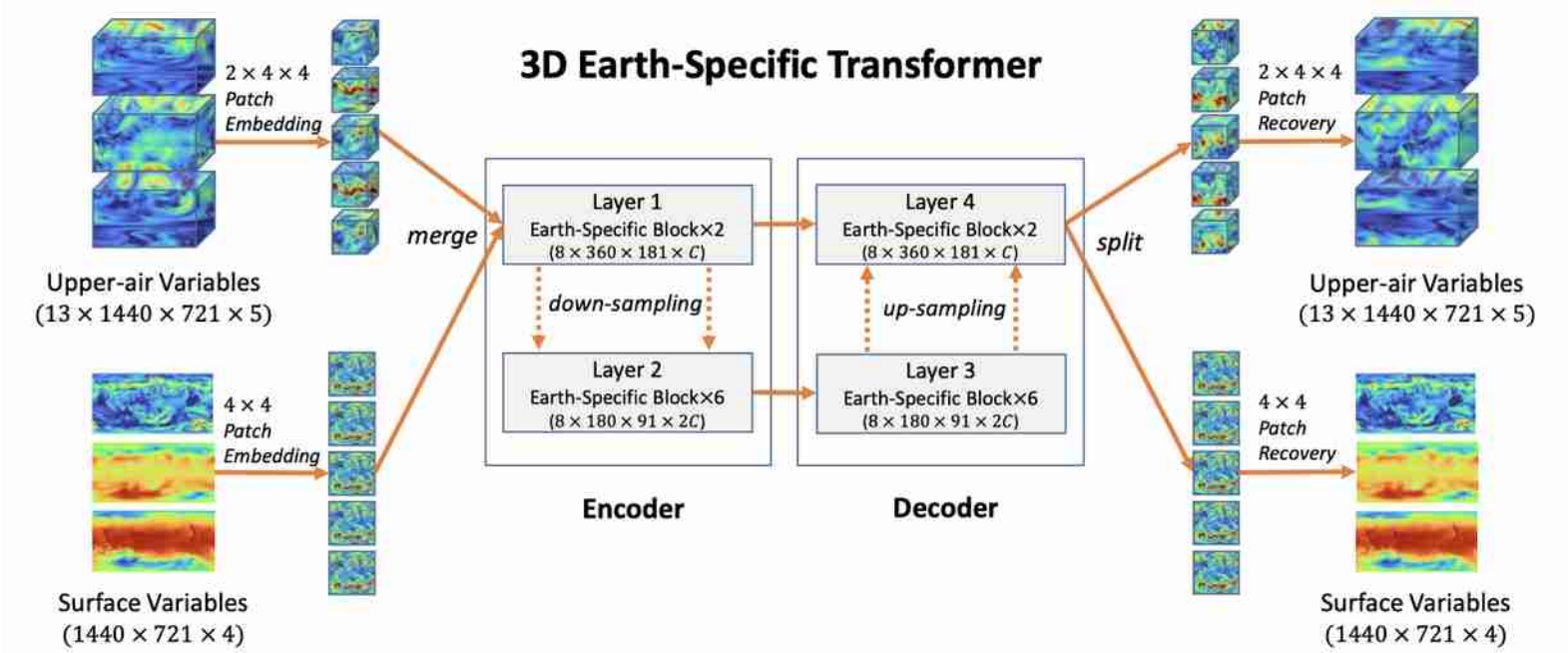


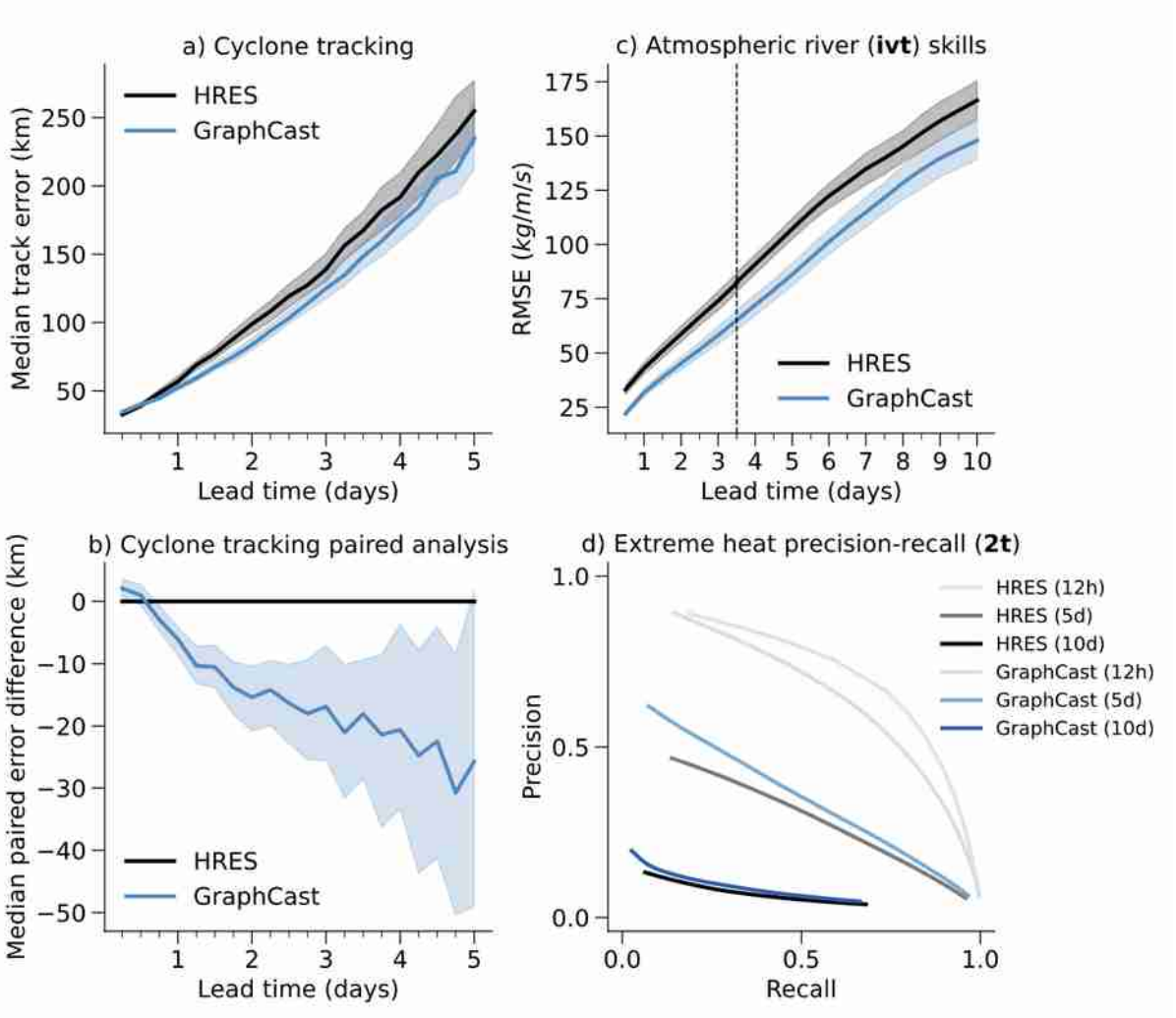
Fig. 2: An overview of the 3D Earth-specific transformer (3DEST). Based on the standard encode-decoder design, we (i) adjust the shifted-window mechanism and (ii) apply an Earth-specific positional bias – see the main texts for details.

- Vision transformer backbone.
- Input split into surface and upper-air variables.
- First operation is patch embedding: dimension reduction + introducing large number of latent channels.
- Then a standard encoder-decoder architecture is applied (8 layers) to each patch.

- Each encoder/decoder layer is a Earth-specific transformer block.
  - This is a type of vision transformer that is specifically designed to align with the Earth's geometry.
  - Window attention mechanism (to further reduce costs) with self-attention in each window.

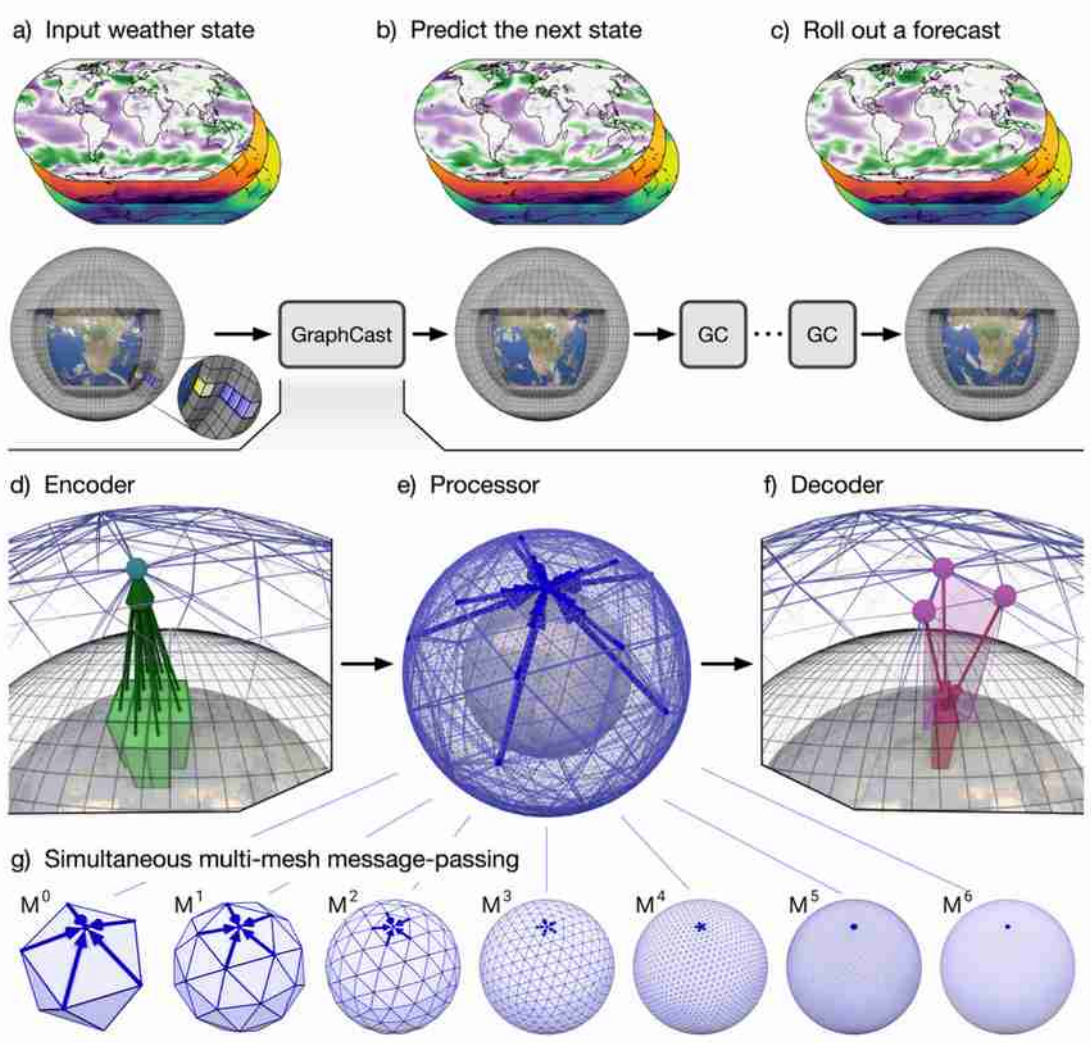
GraphCast (1)

- Released in Dec 2022 by Google.
- Predicts hundreds of weather variables, over 10 days at 0.25 deg resolution globally.
  - All this done in under 1 min on a single Google Cloud TPU v4 device.
  - As a reference, training took ~1 month on 32 Cloud TPU devices.
- Outperforms the most accurate deterministic systems in 90% of verification targets.
  - Also better extreme weather prediction (tropical cyclones, atmospheric rivers, extreme temperatures).



Lam et al. (2023)

GraphCast (2)

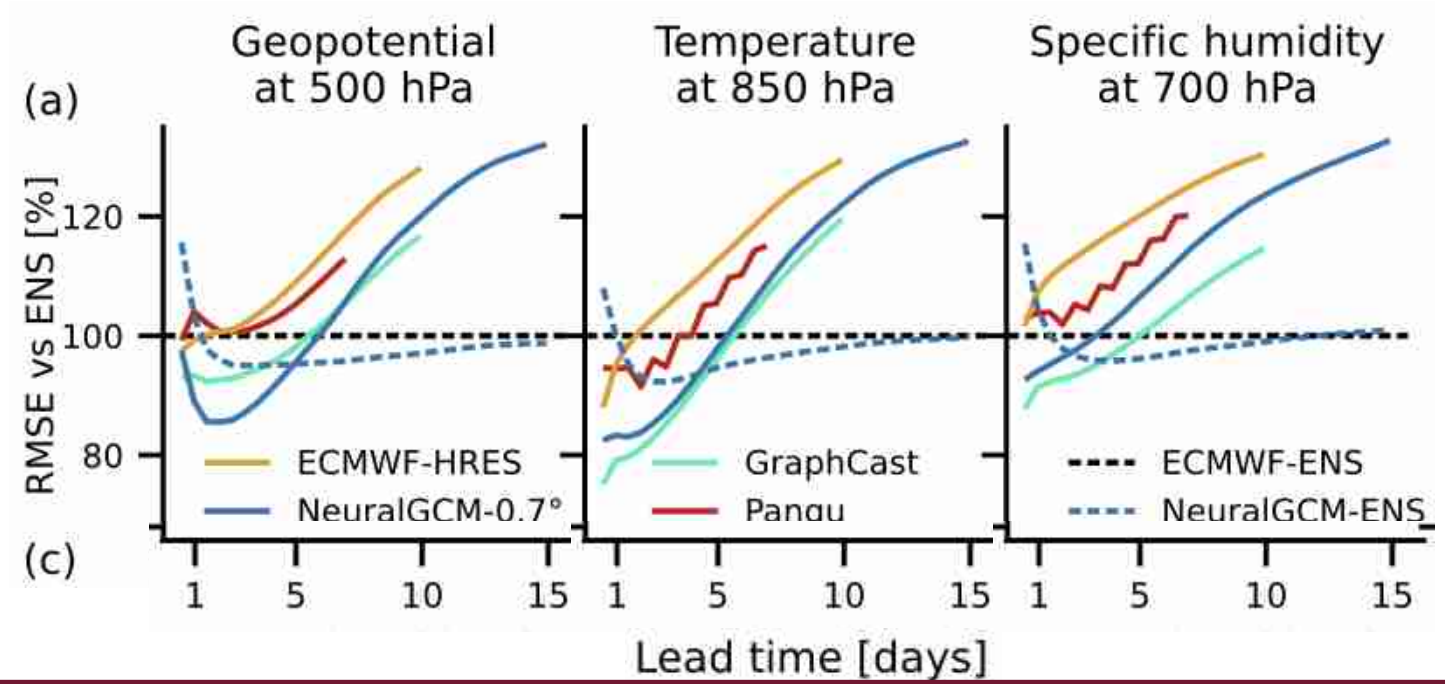


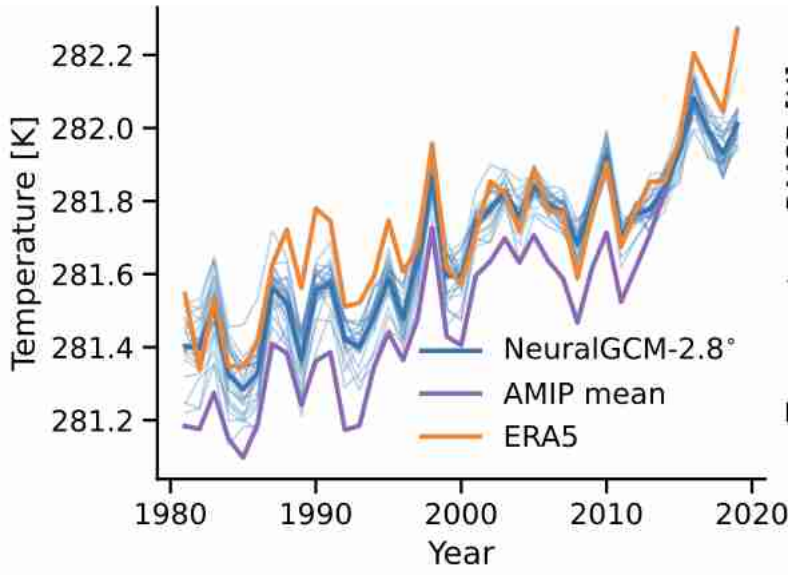
Lam et al. (2023)

- Based on Graph NNs in an encode-process-decode configuration (36.7 million parameters).
- **Encoder**: Single GNN layer to map variables (node attributes on input grid) to learned node attributes on multi-mesh representation.
  - Multi-mesh defined as the iterative refinement of a regular icosahedron 6 times.
- **Processor**: 16 unshared GNN layers perform learnt message passing (along graph edges) on the multi-mesh to propagate local and long-range information.
- **Decoder**: Single GNN layer to map output from final processor's layer back to the lat-lon grid.

Neural GCMs (1)

- First released in Nov 2023.
- General Circulation Models (GCMs): Combine a numerical solver for the large-scale dynamics and tuned representations of small-scale processes.
- Aside from competitive weather performance (left), a unique feature of NeuralGCM is its focus on climate simulations (right).





Kochkov et al. (2024)

Neural GCMs (2)

- This is a hybrid architecture.
  - *Differentiable* dynamical core solving the Navier-Stokes equations.
  - ML component emulating the physical parameterizations.
- Increased physical realism: NeuralGCMs do not suffer from the ``smearing'' at long forecast lead times.
  - In other words, increased stability for long-term weather and climate simulations.

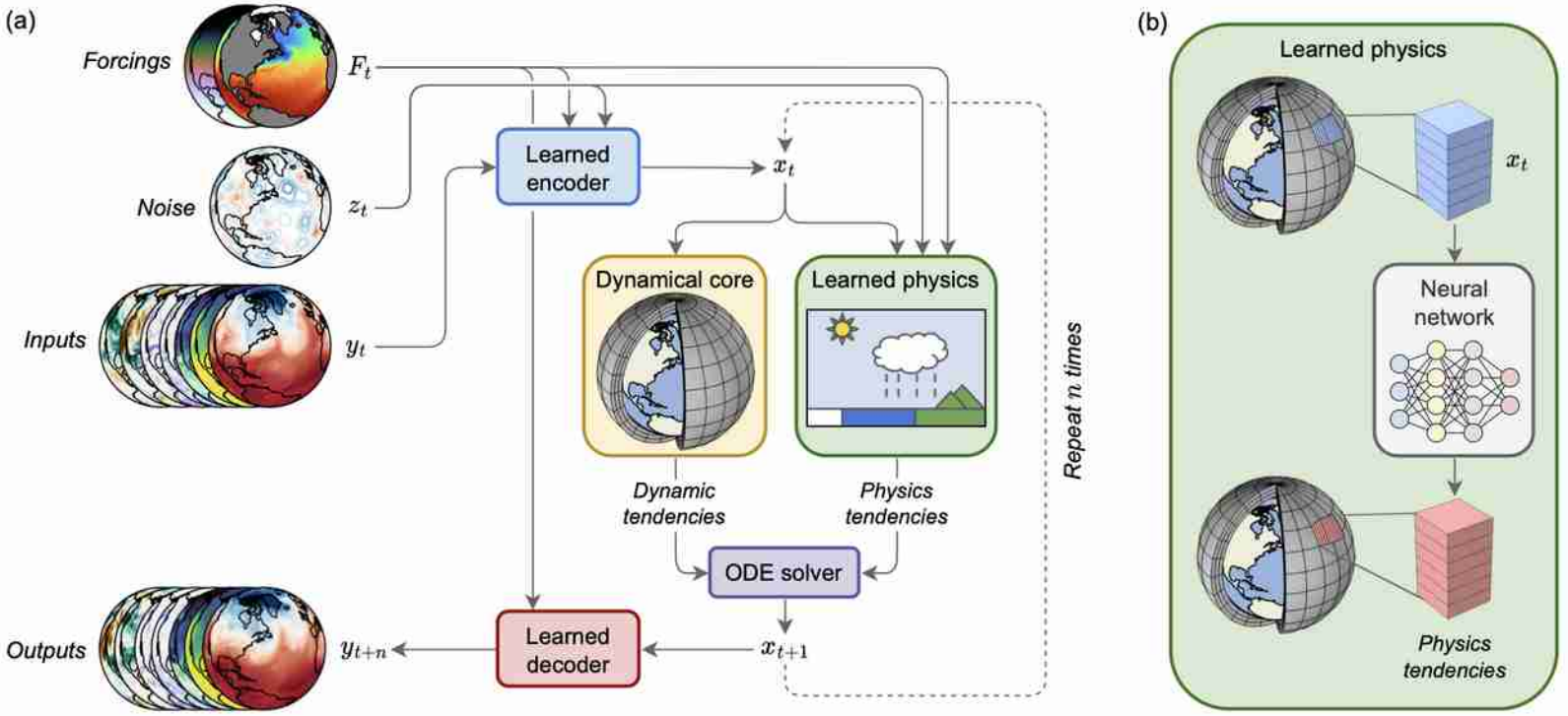
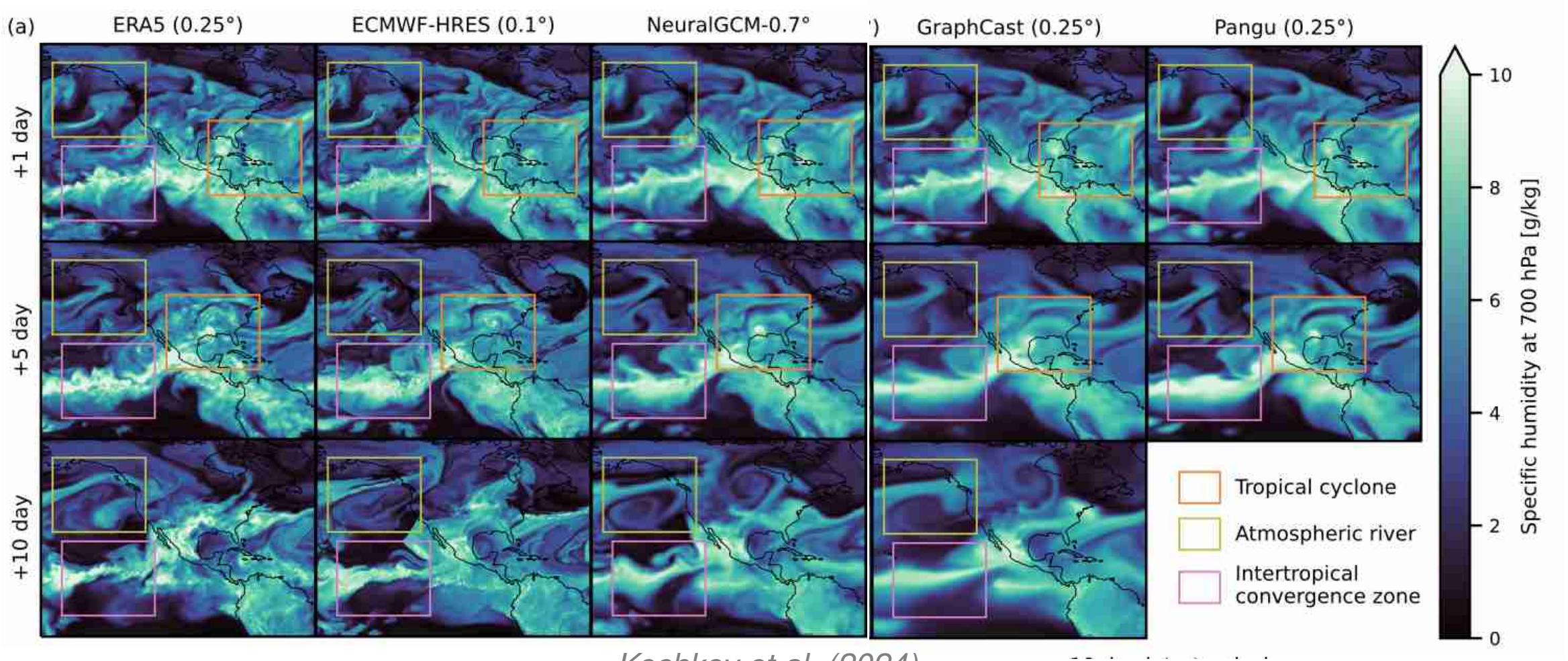


Fig. 1 Structure of the NeuralGCM model. (a) Overall model structure, showing how forcings F_t, noise z_t (for stochastic models), and inputs y_t are encoded into the model state x_t. Model state is fed into the dynamical core, and alongside forcings and noise into the learned physics module. This produces tendencies (rates of change) used by an implicit-explicit ODE solver to advance the state in time. The new model state x_{t+1} can then be fed back into another time step, or decoded into model predictions. (b) Inset of the learned physics module, which feeds data for individual columns of the atmosphere into a neural network used to produce physics tendencies in that vertical column.

Kochkov et al. (2024)

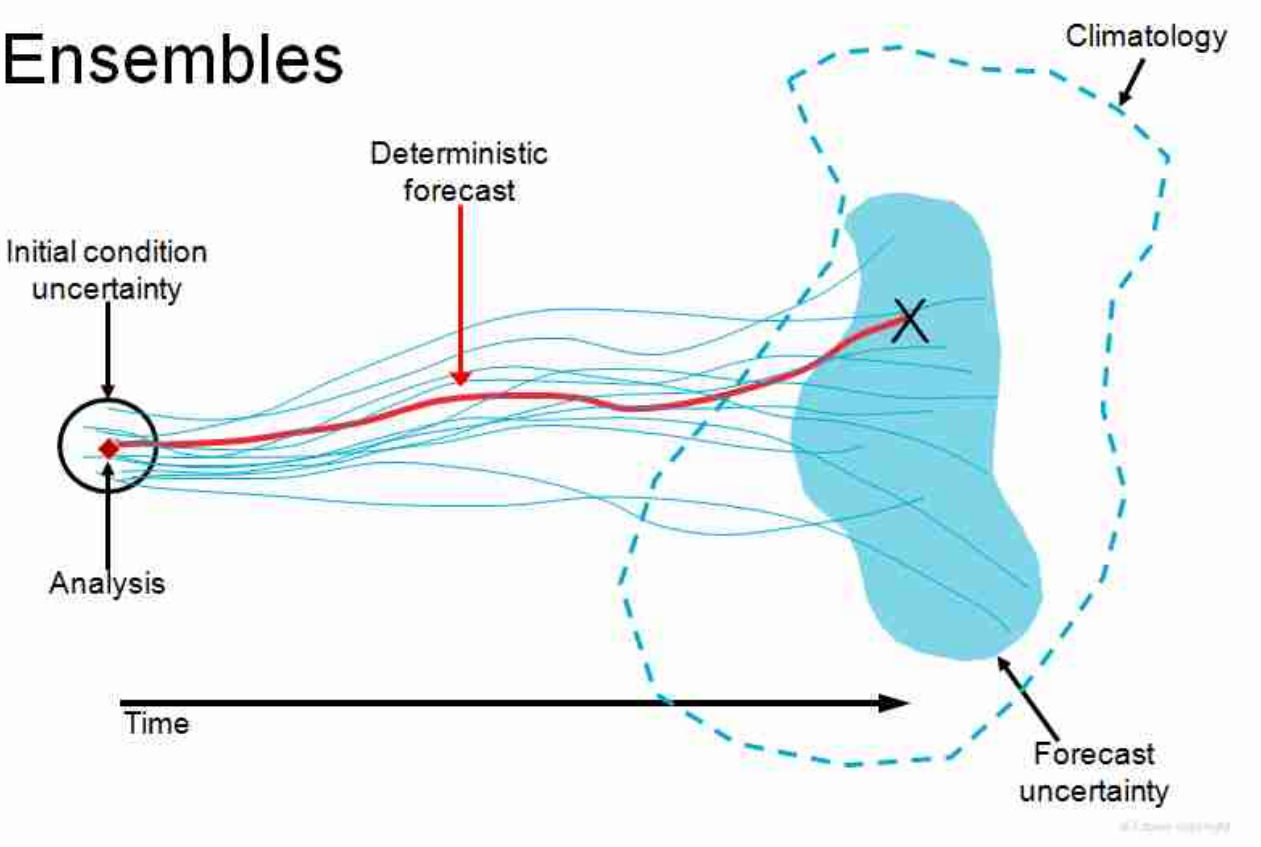
Neural GCMs (3)

- Example showing how the long-range NeuralGCM forecasts preserve their spatial granularity.



Kochkov et al. (2024)

Ensemble forecasting

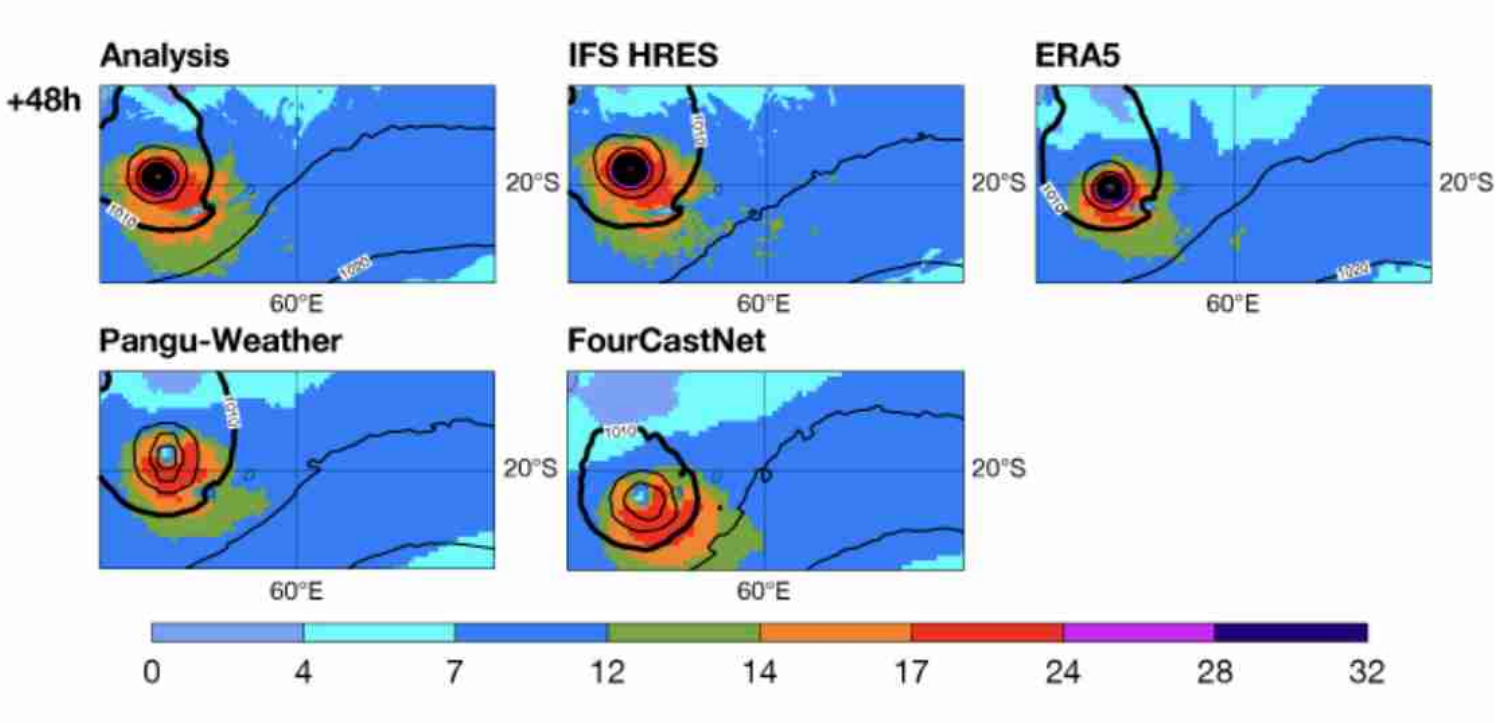


Courtesy of Met Office (UK).

- Early 1990s were marked by a paradigm shift in NWP due to the advent of ensemble forecasting.
- Idea is simple: run multiple forecasts starting from perturbed initial conditions.
- Due to the chaotic nature of atmospheric dynamics, ensemble members will begin to diverge over time => ensemble spread allows us to describe forecast uncertainty.

GenCast (1)

- Released in December 2023 by Google.
- The first DWP model specifically designed for probabilistic forecasting.
  - GenCast more skilful than ECMWF's ensemble system 96% of the time.



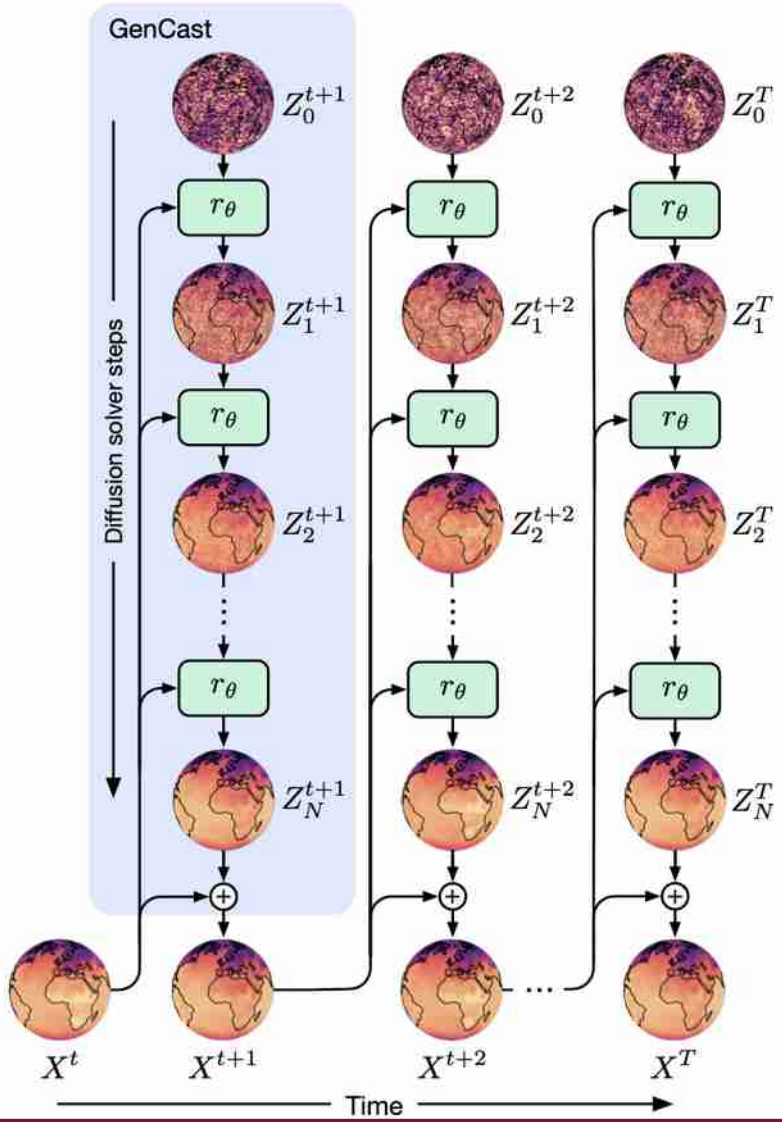
Chantry et al. (2023) [link]

- Prior to GenCast, ML models largely trained to produce deterministic forecasts which minimize MSEs.
  - The problem of such training is the lack of physical consistency at longer forecast lead times: predictions are smoothed out and forecasts of extremes are penalized.

GenCast (2)

- GenCast uses a radically different training approach based on diffusion models.
- Just like regular diffusion models, a realistic image of the Earth's atmospheric state is produced by learning to denoise a noise-corrupted image.
- The main difference is that the learnt denoiser is conditioned on the previous 2 atmospheric states (graphic is simplified as it only shows conditioning on the previous state):

$$Z_{i+1}^{t} = r_\theta\left(Z_i^t; X^{t-2:t-1}, \sigma_{i+1}, \sigma_i\right)$$



Price et al. (2023)

Lessons learnt and next steps

- First DWP models met with a lot of scepticism, but recent progress has been really encouraging.
  - For most purposes AI models perform equally well or better than traditional NWP.
  - The purely data-driven nature of most DWP models is still amusing: ML just works with a few minor tweaks and few decades of reanalysis data seems to be sufficient.

- The second NWP revolution is still ongoing: the discussed AI models (mostly based on GNN and ViT architectures) are just a first generation. More work needed in terms of the:
  - Need for higher resolution models comparable to operational models.
  - Inclusion of societally relevant variables (e.g., precipitation).

- One area of particular importance is research at the intersection of DWP and data assimilation: very much a low-hanging fruit still.
  - DWP demonstrate impressive performance but still rely on initial conditions generated from the optimal blending of physical models and observations.

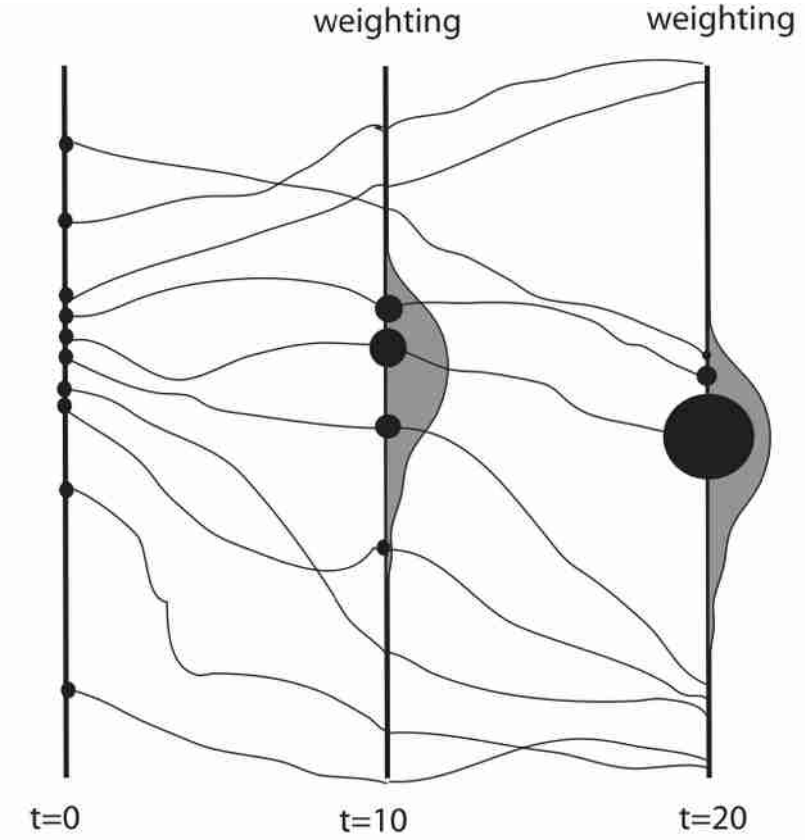
Ongoing work: Particle filtering with ML ensembles

- My student Katy Merritt is currently working on addressing this gap in the context of particle filters, which are advanced (nonlinear/non-Gaussian) DA methods.

- One of the main problems with particle filters is that they are susceptible to the curse of dimensionality in high dimensions.
  - When the number of particles is too low, weight collapses to one particle.

- The problem can be alleviated with larger ensembles, but operational centers can only afford O(100) with physics-based NWP.

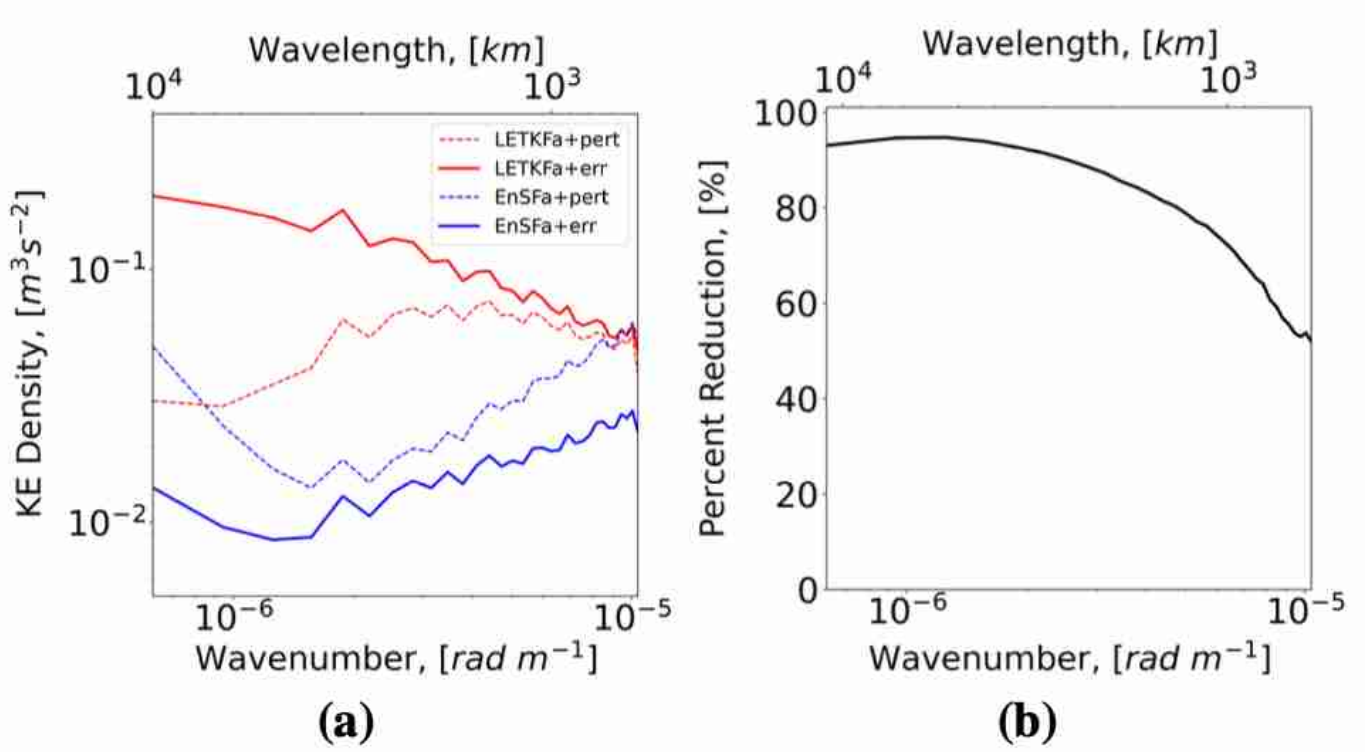
- **Project**: AI models can generate much larger ensembles much faster => can we use them for particle filtering?



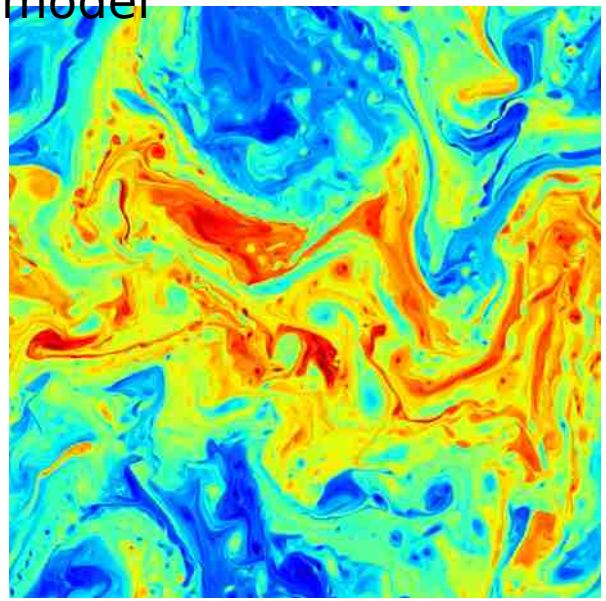
van Leeuwen (2009)

Ongoing work: Diffusion models for data assimilation

- Training-free (Monte Carlo) implementation of diffusion models for non-Gaussian data assimilation.
- Significant improvements over reference LETKF method in SQG model with arctan observations.

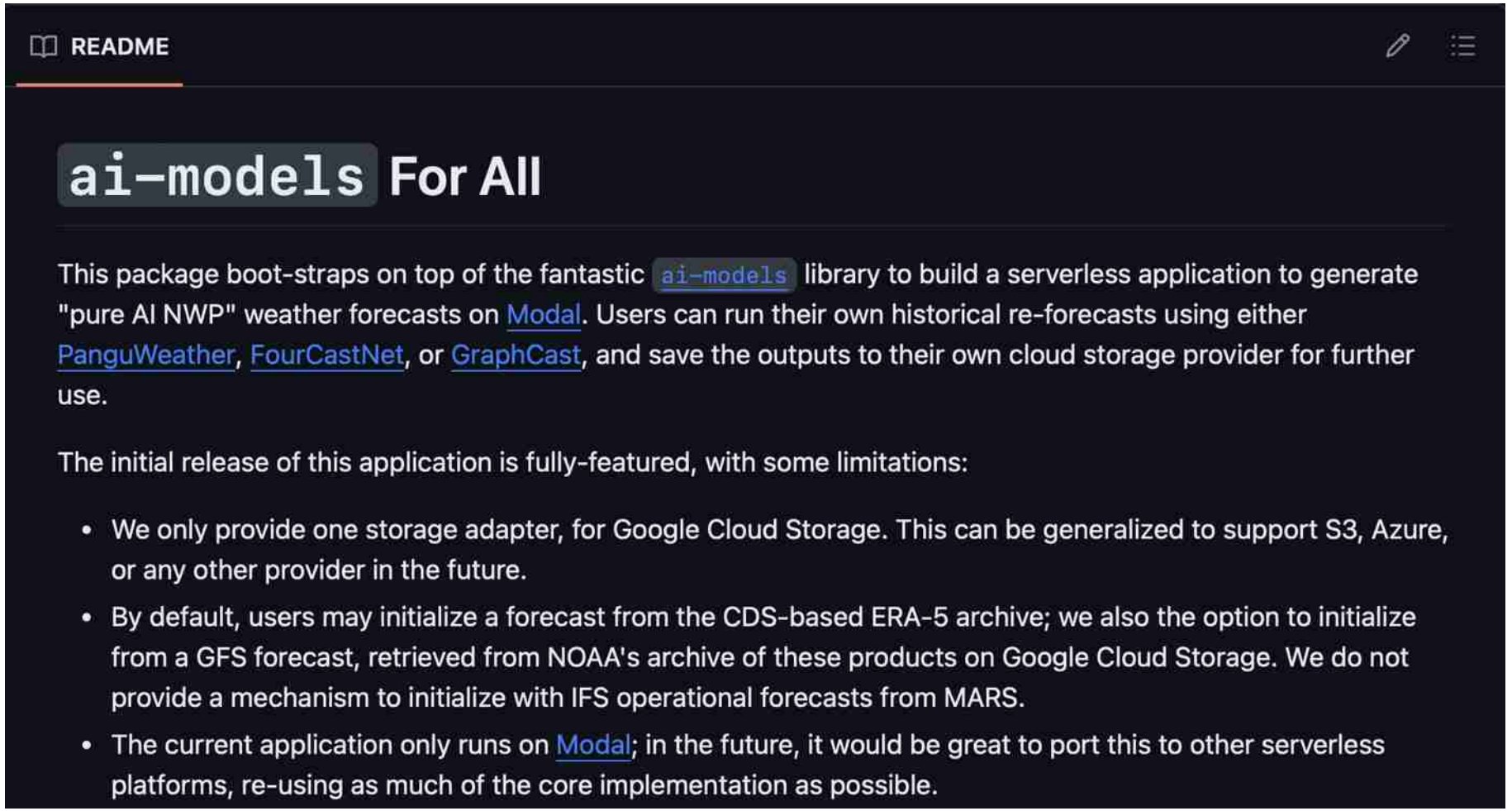


Snapshot of the SQG model



Run your own AI model

- Link: https://github.com/darothen/ai-models-for-all.



README

`ai-models` For All

This package boot-straps on top of the fantastic `ai-models` library to build a serverless application to generate "pure AI NWP" weather forecasts on Modal. Users can run their own historical re-forecasts using either PanguWeather, FourCastNet, or GraphCast, and save the outputs to their own cloud storage provider for further use.

The initial release of this application is fully-featured, with some limitations:

- We only provide one storage adapter, for Google Cloud Storage. This can be generalized to support S3, Azure, or any other provider in the future.
- By default, users may initialize a forecast from the CDS-based ERA-5 archive; we also the option to initialize from a GFS forecast, retrieved from NOAA's archive of these products on Google Cloud Storage. We do not provide a mechanism to initialize with IFS operational forecasts from MARS.
- The current application only runs on Modal; in the future, it would be great to port this to other serverless platforms, re-using as much of the core implementation as possible.

FourCastNet notebook

- Link: https://colab.research.google.com/drive/1HoP1Jn55rm4YjzhDve_X0PMYQwrcBxNW?usp=sharing#scrollTo=3zclV4lUiAo3

```
# visualize spatiotemporal predictions
fig, ax = plt.subplots(nrows=1, ncols=2, figsize=(15, 5))
t = 2 # at 2x6 hours lead time
ax[0].imshow(predictions_cpu[t,0], cmap="bwr")
ax[1].imshow(targets_cpu[t,0], cmap="bwr")
ax[0].set_title("FourCastNet prediction")
ax[1].set_title("ERA5 ground truth")
fig.tight_layout()
```