Data-driven weather forecasts

The future of weather forecasting?

Matthew Chantry

With thanks to Zied, Mariana, Simon, Mihai and many many more

ECMWF

© ECMWF March 21, 2024

What are we going to explore?

How are data-driven models trained?

What datasets are used?

What architectures are used, and why?

What are the current strengths/weaknesses?

Are data-driven models physical?

Other approaches for ML forecasting

2018
Exploring the concept

ECMWF staff
~500km ERA5 to predict future z500.
Similar work from Rasp and Weyn.

2020 WeatherBench

Defining the dataset, split, headline fields and metrics

Feb 2022
Full medium-range NWP

Keisler - GraphNN 1°, competitive with GFS

NVIDIA – FourCastNet
Fourier+ , 0.25°

O(10^4) faster & more energy efficient than IFS

Nov 2022
Tropical cyclones

Huawei – PanguWeather
0.25° hourly product

"More accurate tracks" than the IFS.

Dec 2022
Extensive predictions

Deepmind – GraphCast
0.25° 6-hour

Many variables and pressure levels with comparable skill to IFS.

Jan 2023
Global & Limited Area

Microsoft – ClimaX

Forecasting various lead-times at various resolutions, both globally and regionally

Apr 2023
7-day+ scores improve

FengWu – China academia + Shanghai Met Bureau
0.25° 6-hour product

Improves on GraphCast for longer leadtimes (still deterministic)

Jun 2023
Spherical harmonics

NVIDIA – SFNO
0.25° 6-hour product

Extension of FourCastNet to Spherical harmonics, improved stability

Diffusion modelling

Alibaba – SwinRDM
0.25° 6-hour product

Sharp spatial features

Last months
FuXi
AtmoRep
FuXi-extreme
NeuralGCM
GenCast
…

impossible to keep this figure up

What data do they train on?

- ERA5 dominates the landscape.
  - Long, global, self-consistent.
  - Easy and quick to access.

- Some work learning forecast trajectories, but this “limits” accuracy to that of the existing physical model.

- Now GraphCast & AIFS use a few years of IFS operational analysis. These improve the model, particularly for its use initialised from operational IC.

- Next steps, we can expect other centres to explore fine-tuning on their own IC.

- Beyond, combinations of observations and (re)analysis.

Design choices for learning a model from (re)analysis

What architecture? How should spatial relationships be encoded?

What to minimise?

How to handle time?

What variables do I use?

Architecture choices

Convolutions

Graphs

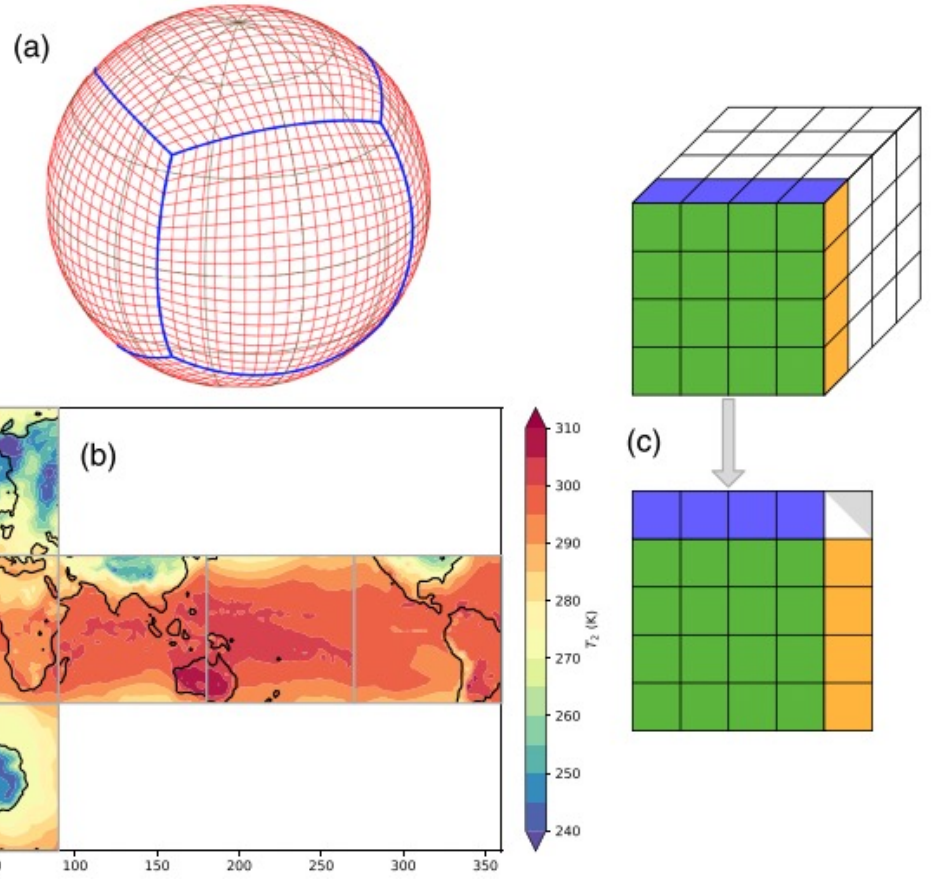
Transformers

Fourier Neural Operators

Convolutions

- Simplest design, treat the Earth as a cylinder, use convolutions with periodicity in longitude.
  - No treatment of the pole. How can flow easily pass over the artic?

- Weyn et al 2020 proposed a clever cubed-sphere approach.
  - One set of CNN for the side faces of the cube.
  - Another for the polar faces.



- Karlbauer et al (2023) do convolutions on the HealPix grid (see below)

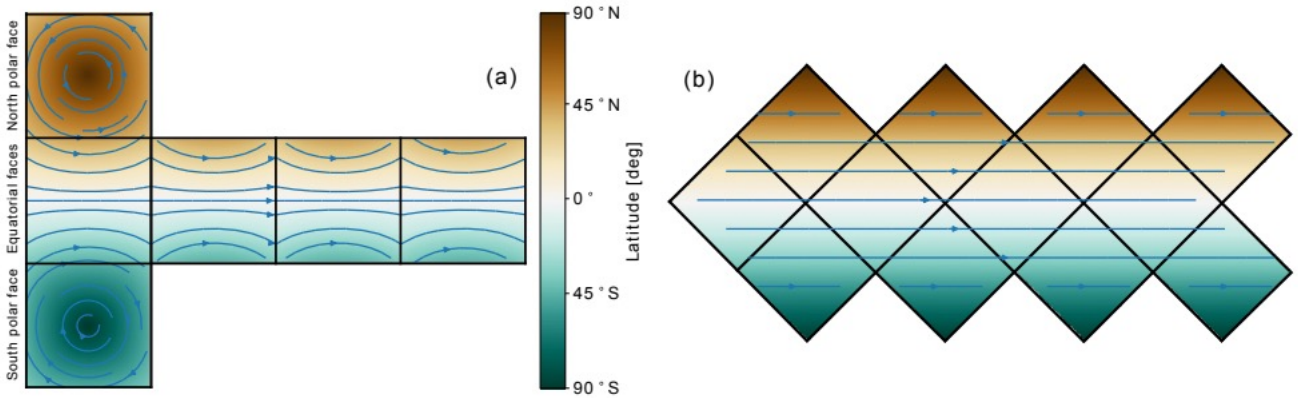
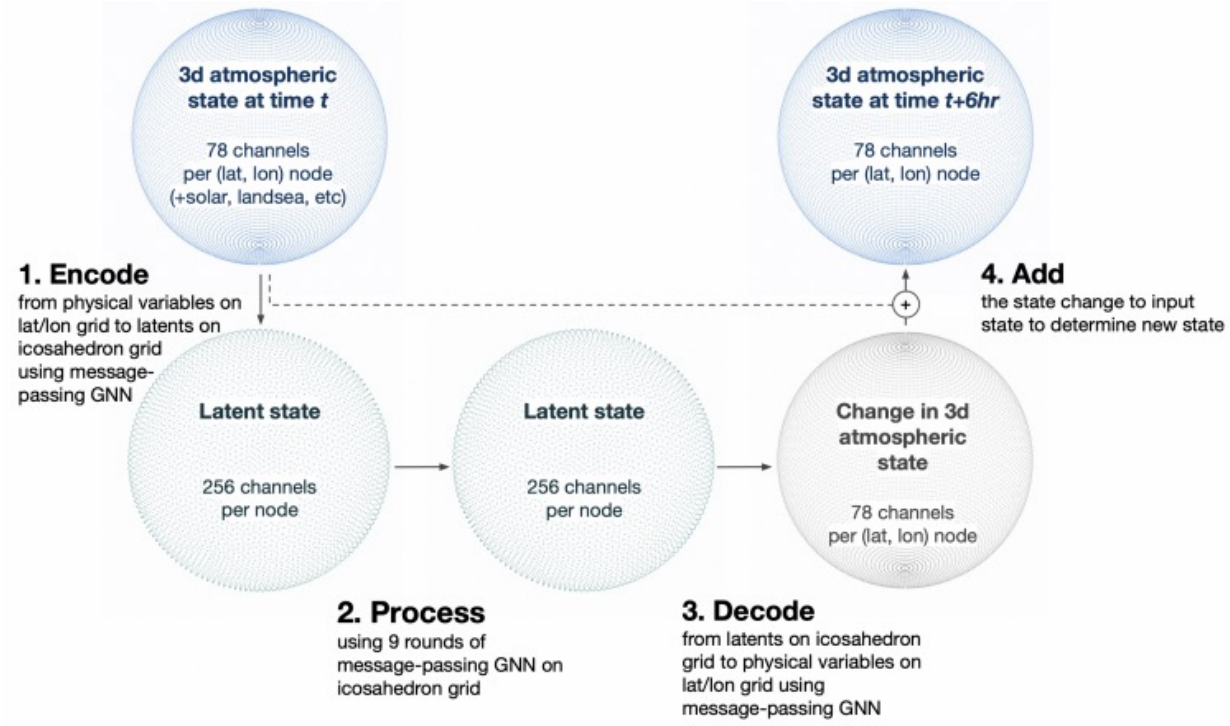
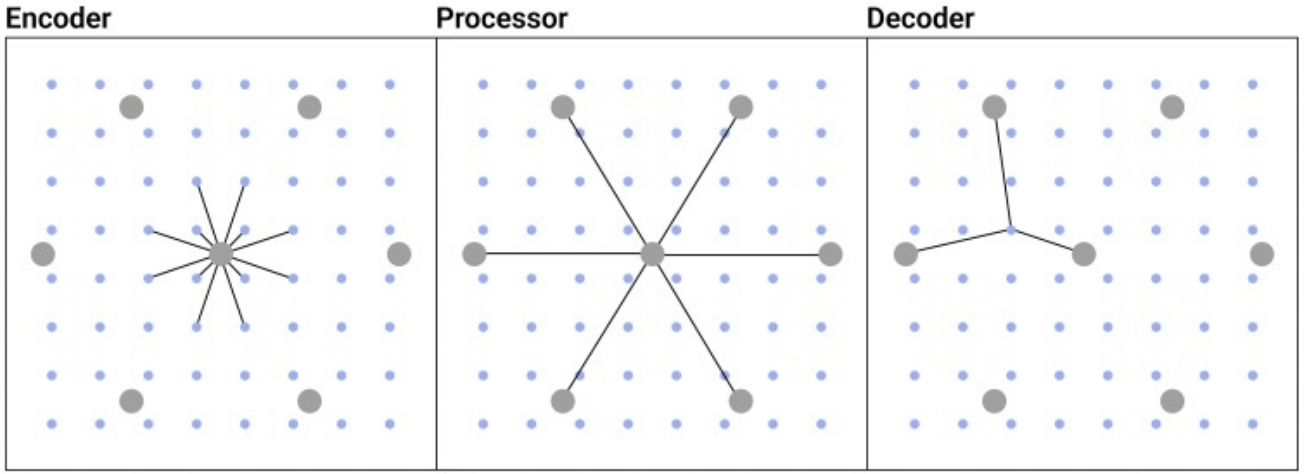


Figure 10: Lines of latitudes depicted as blue streamline arrows on the cubed sphere (a) and on the HEALPix (b). While the lines corresponding to constant eastward motion describe arcs of different radii on the cubed sphere mesh, the same motion translates to straight lines on the HEALPix mesh.

Graph Neural Networks

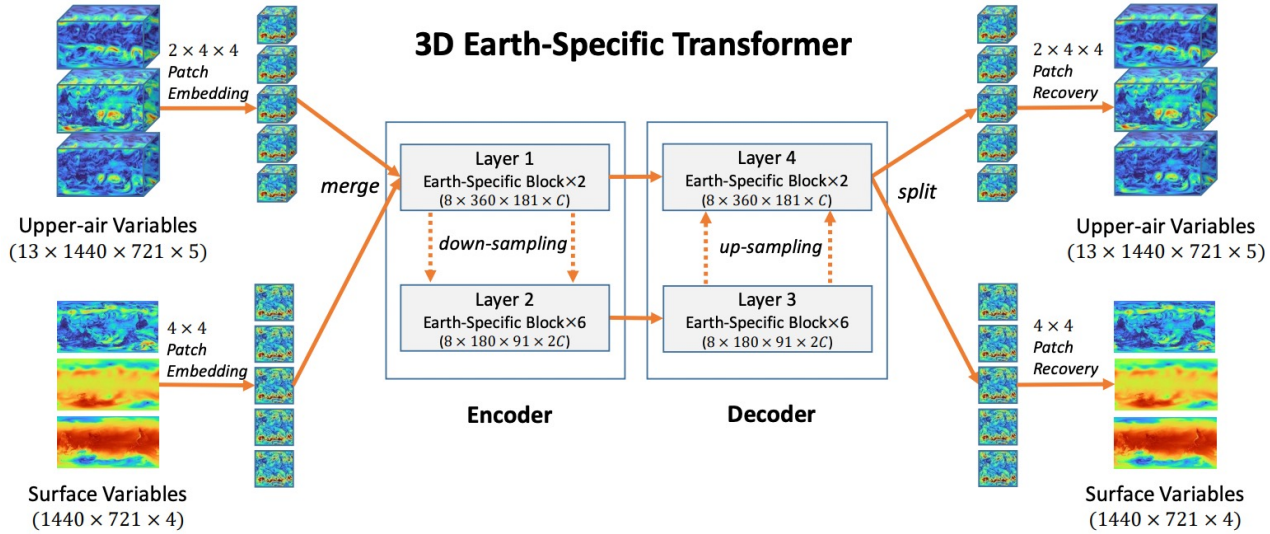
- First demonstrated by Keisler 2022.
- Most popular, the message passing GNN.
  - Involves MLPs on the edges, and nodes.
  - Alternates are GraphConvolutions & Graph Attention
- Further developed in GraphCast
  - And used in early versions of the AIFS.
- No issues at the poles.
- Can handle irregular data in space.



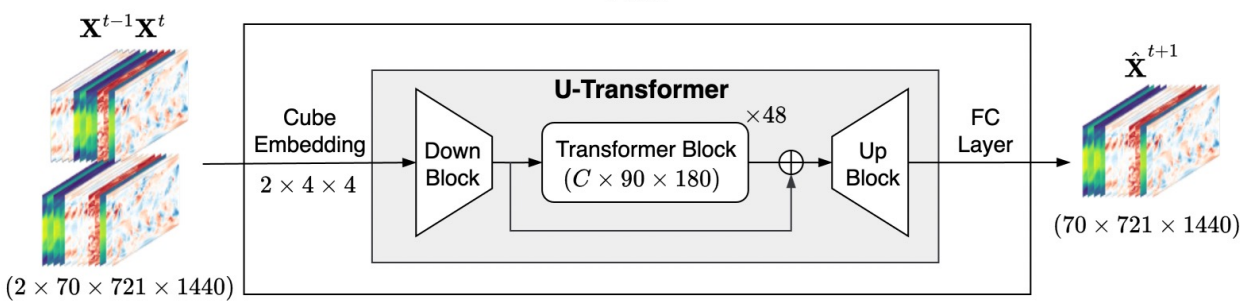


(Vision) Transformers

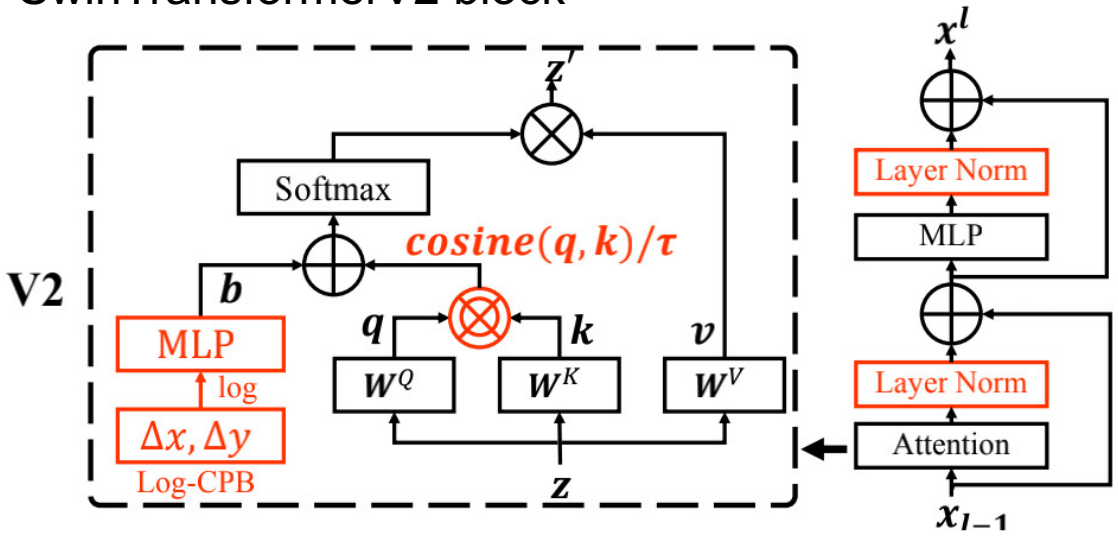
- Build heavily on advances in vision transformers
  - Specifically Shifted Window (SWIN) approaches.
- Pangu, FuXi, FengWu.
- Embed to a coarser resolution.
- Shifted-window approach adapted to include longitudinal periodicity.
  - But poles are not explicitly handled.



a) The overall architecture of FuXi model

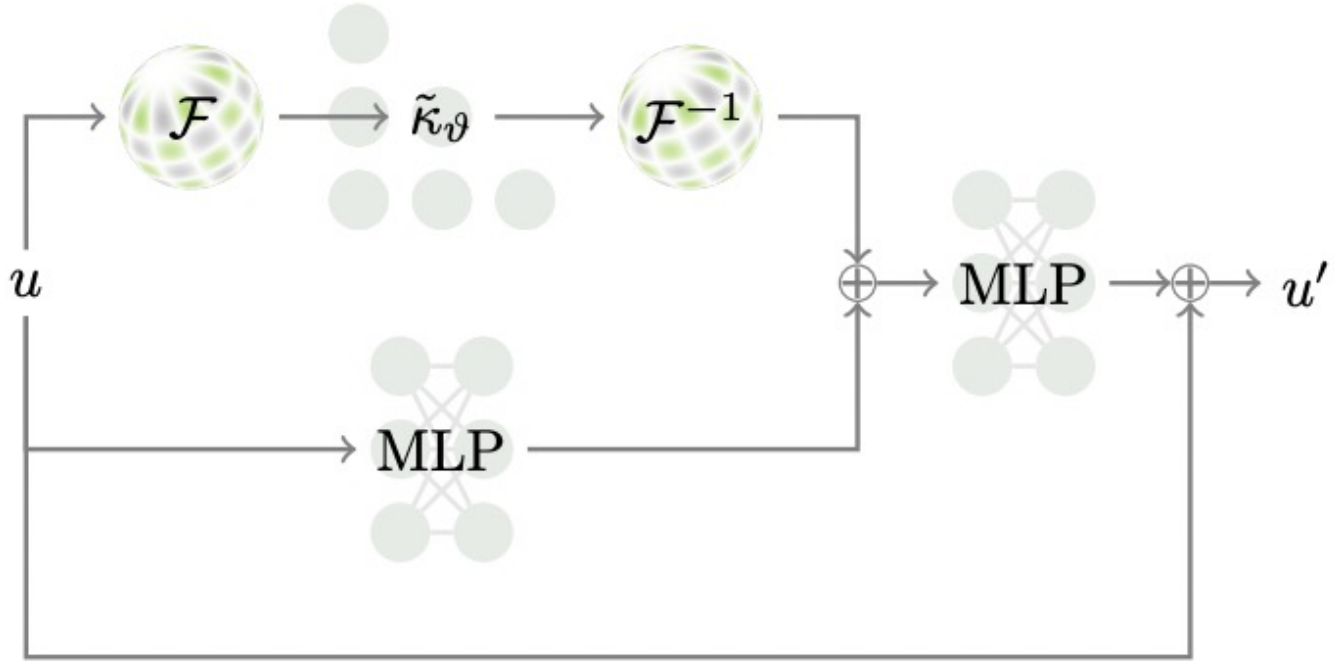


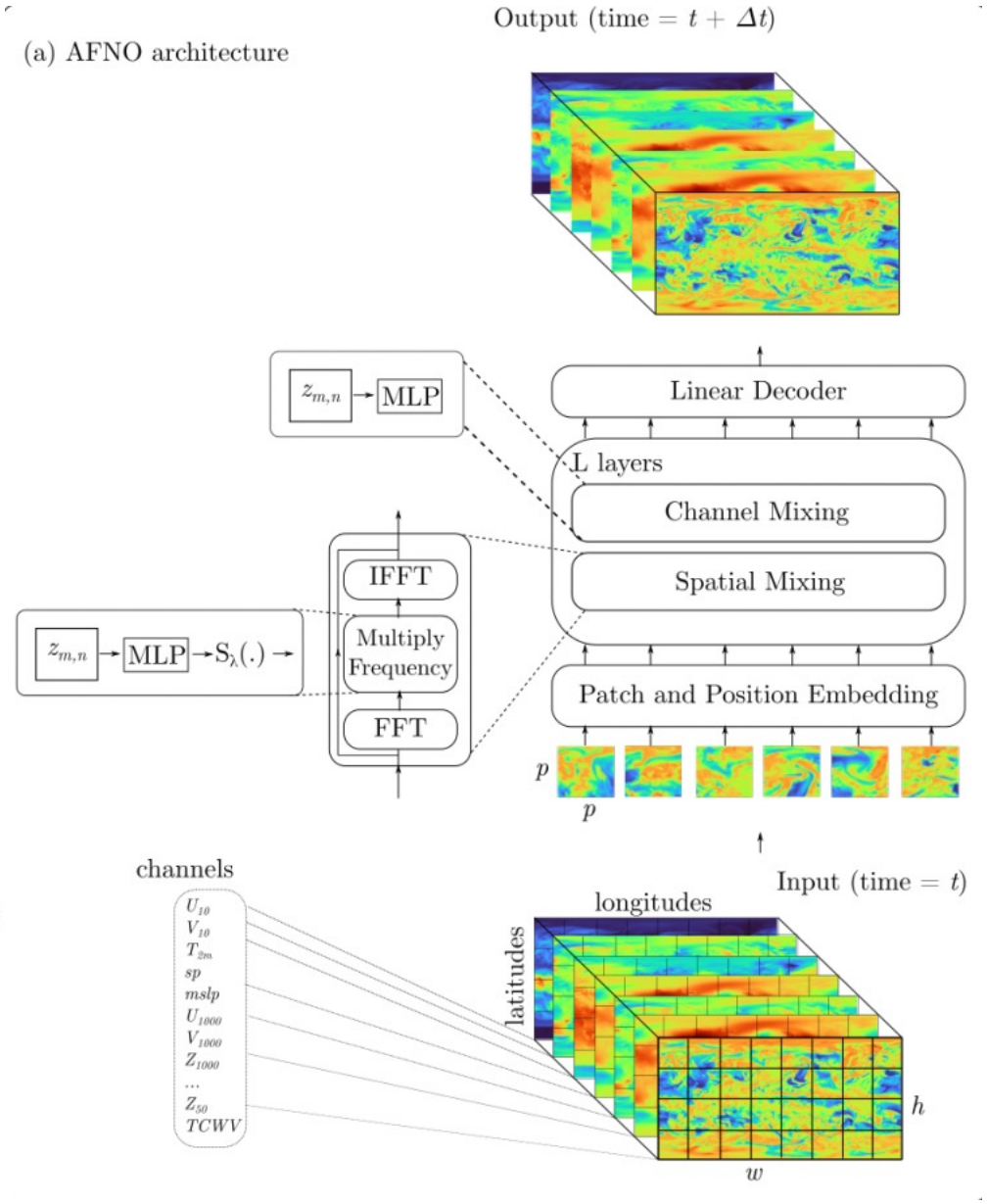
SwinTransformerV2 block



Fourier Neural operators

- e.g. FourCastNet, popularised by NVIDIA.
- Part of neural network carried out in frequency space.
  - Part in grid-point space.
- Grid-invariance built in.
- Spherical version encodes the symmetries of the sphere.
- Also used in “ACE”, the climate emulator.





So, which to choose?

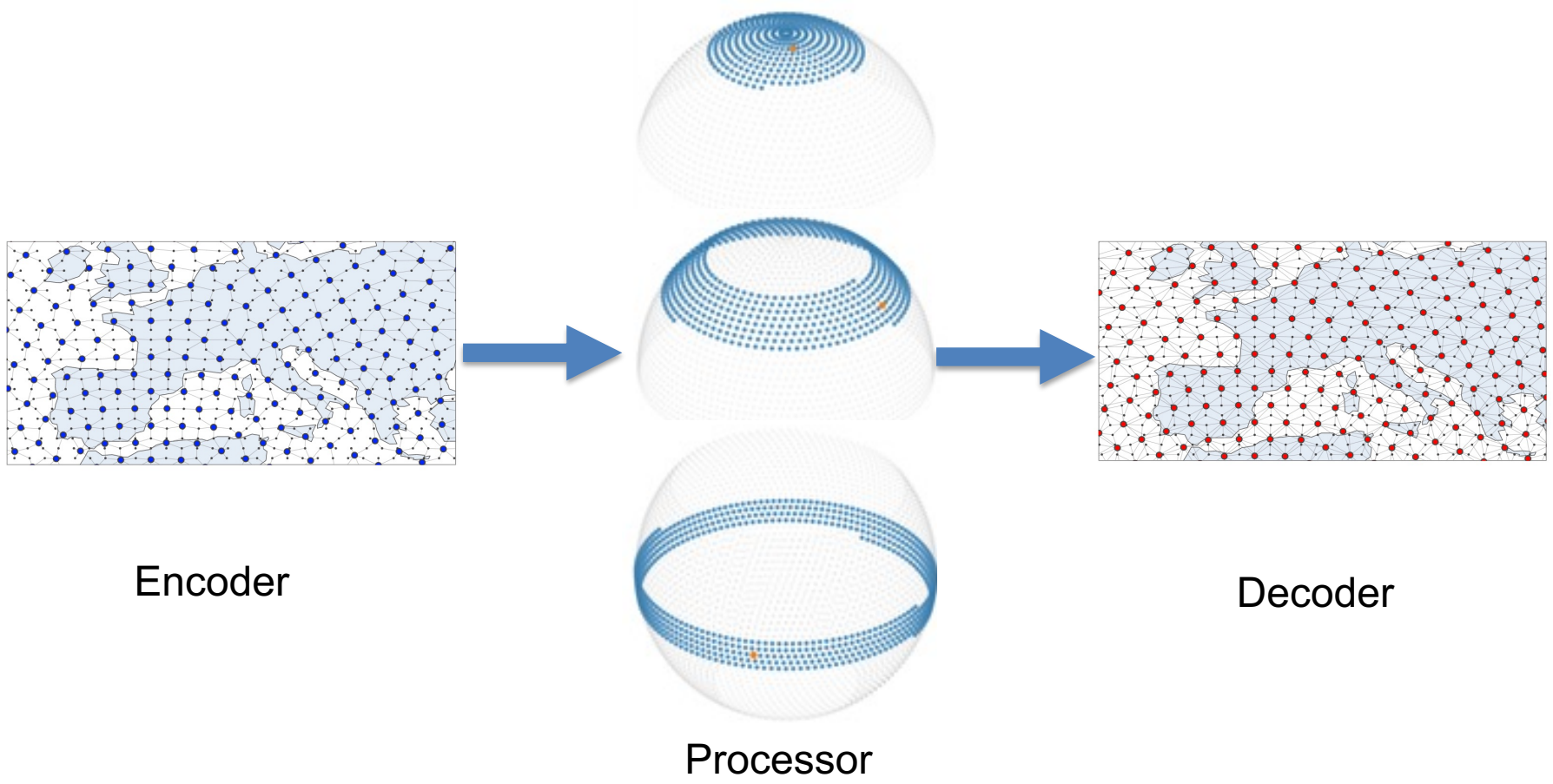
- Vision transformer and GNN solutions both hit comparable levels of skill.
  - SFNO a little further behind in skill, unclear why.
  - CNN not been implemented at the same scale.

- GNN naturally encodes the sphere and allows use of equi-spaced grids.

- Vision transformers (and SFNO) appear to converge faster than GNN.

Latest AIFS – hybrid of graphs & transformers

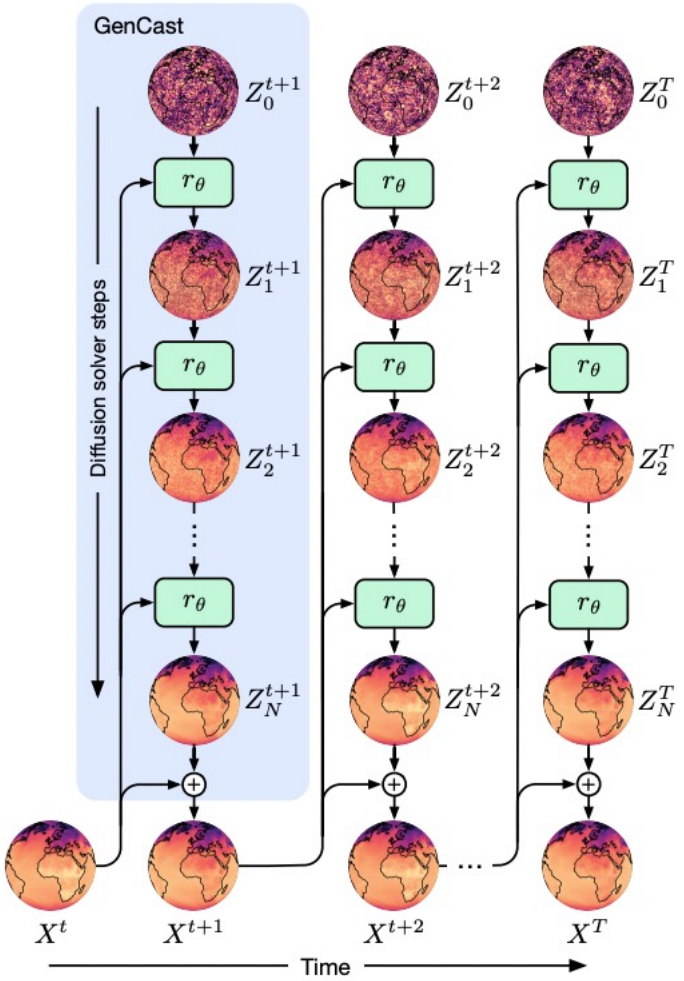
- Encoder/decoder: graph attention.
- Processor: Transformer blocks, attention across regional bands.



What loss to optimise?

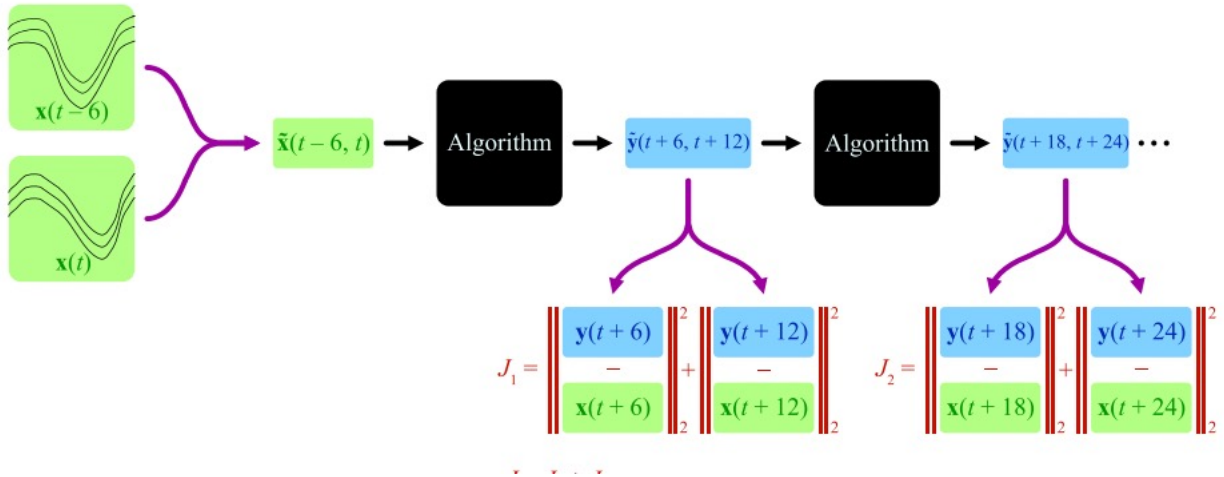
- MSE and MAE make for very popular choices.
- But the problem is inherently probabilistic.
  - Generative techniques can help here!
  - GenCast (right) builds a denoising network that is inherently uncertain.
  - NeuralGCM, directly aims to minimise the CRPS (a probabilistic skill score).

- For all losses, you need to decide how to aggregate over variables & heights.
  - This introduces many parameters…
  - FengWu predict mean & standard-deviation, **optimise log-likelihood** argue that homoscedastic uncertainty balances the loss.

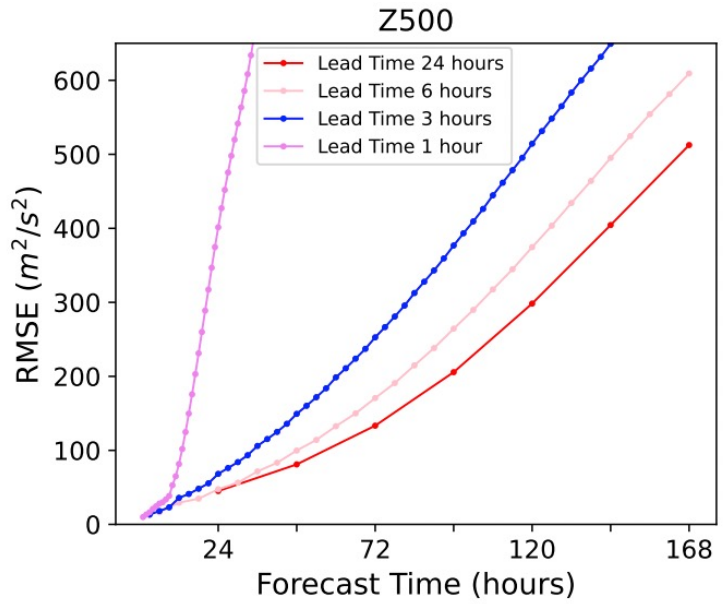


How to handle time?

- How many time slices to provide as input?
  - Weyn et al (2020) provide 2 time slices and get out the next two.
  - This is now used by many others.

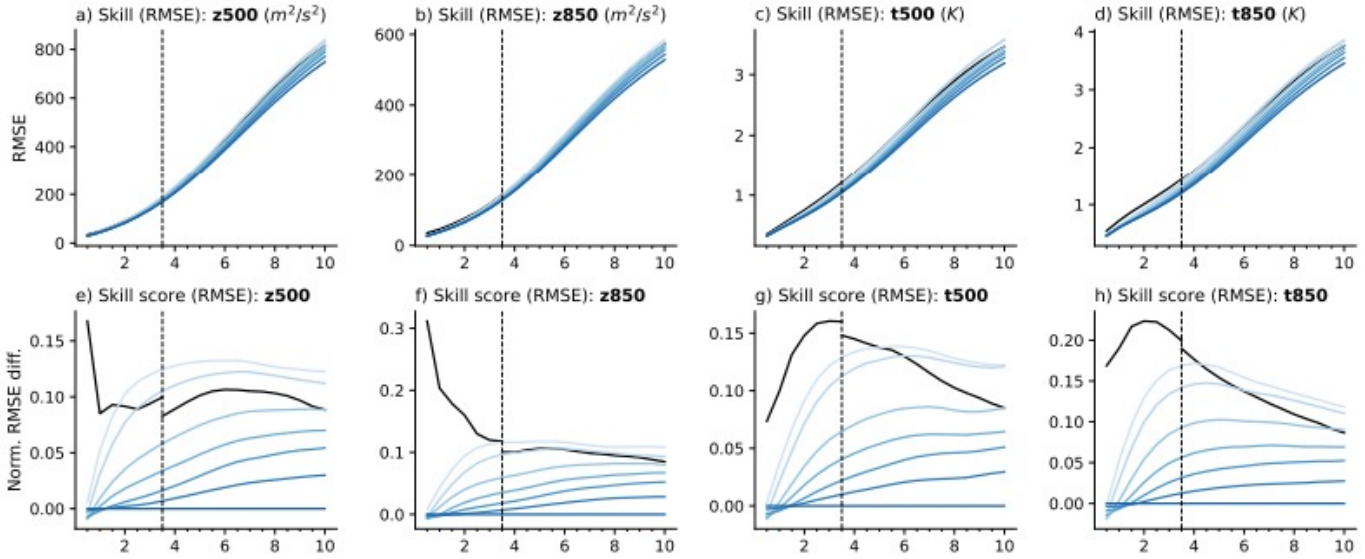


- How big of a timestep to make?
  - Early work tried 3-day steps but failed to compete with physics models.
  - Many choose 6-hours, but this limits the granularity of the output.
  - Pangu Weather created multiple models for different timesteps…
  - but this leads to inconsistencies in time.

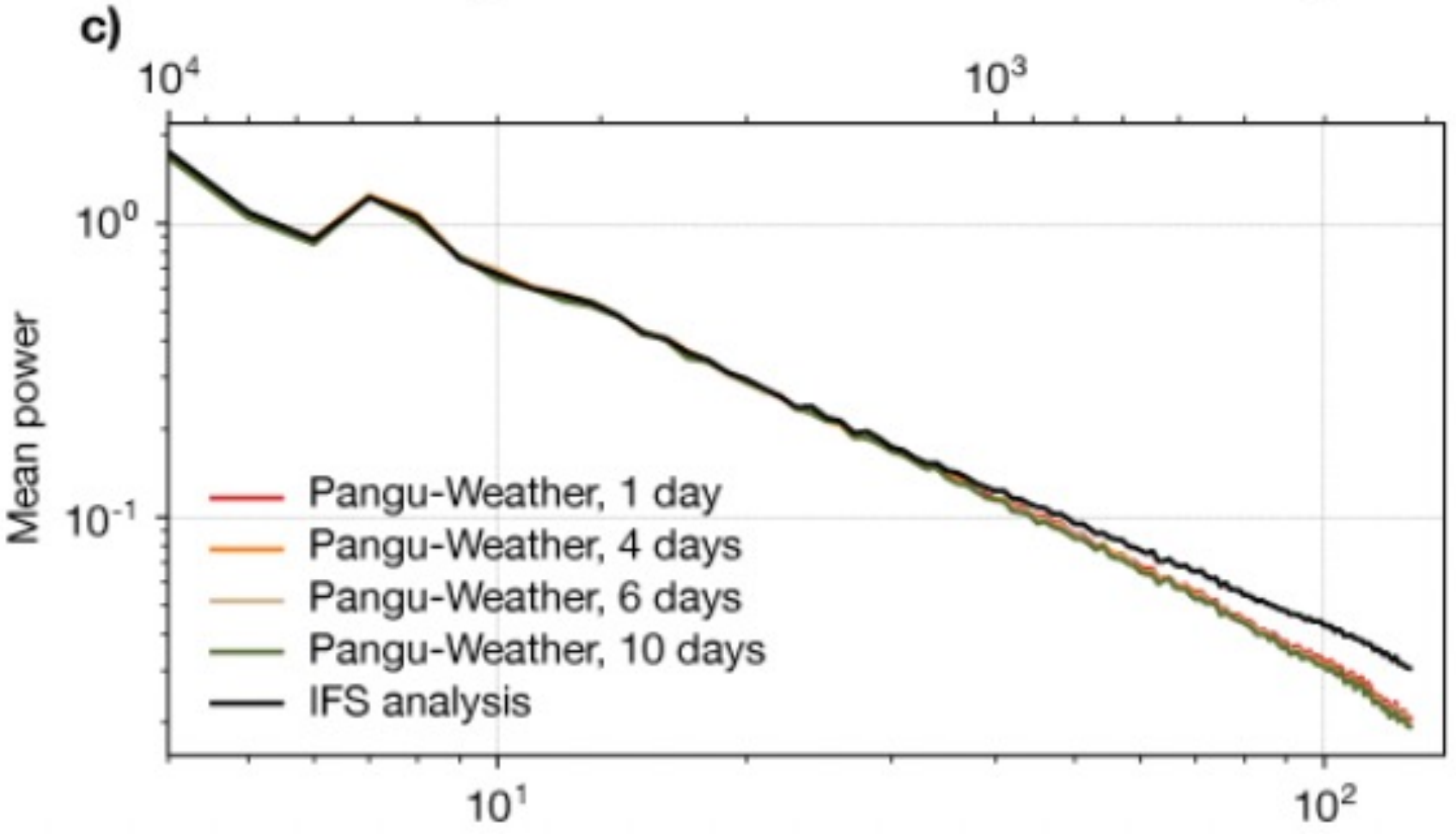


Minimise over long time windows?

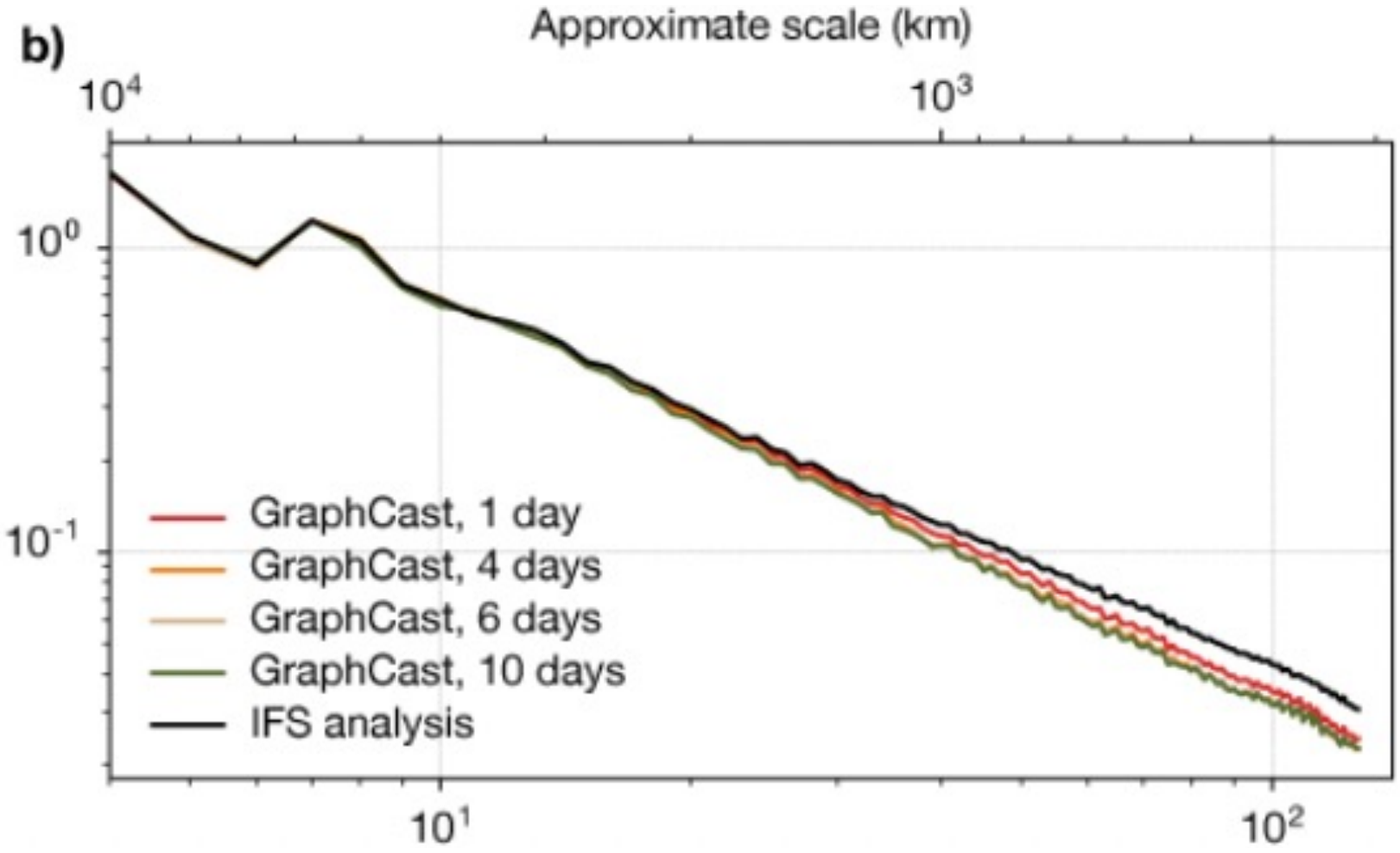
- FourCastNet & Keisler proposed to minimise the loss over multiple applications of the model.
  - i.e. not just minimise f(x(t)) against x(t+6), but also f(f(x(t)) against x(t+12).
  - This leads to stable and accurate results.
  - GraphCast came up with an efficient algorithm for minimising out to 72h.
  - FuXi took it to the extreme, minimising out to 15 days.
- FengWu made a “buffer” of predictions from their model and used this as training data.
  - Thereby training the model to deal with its own output.



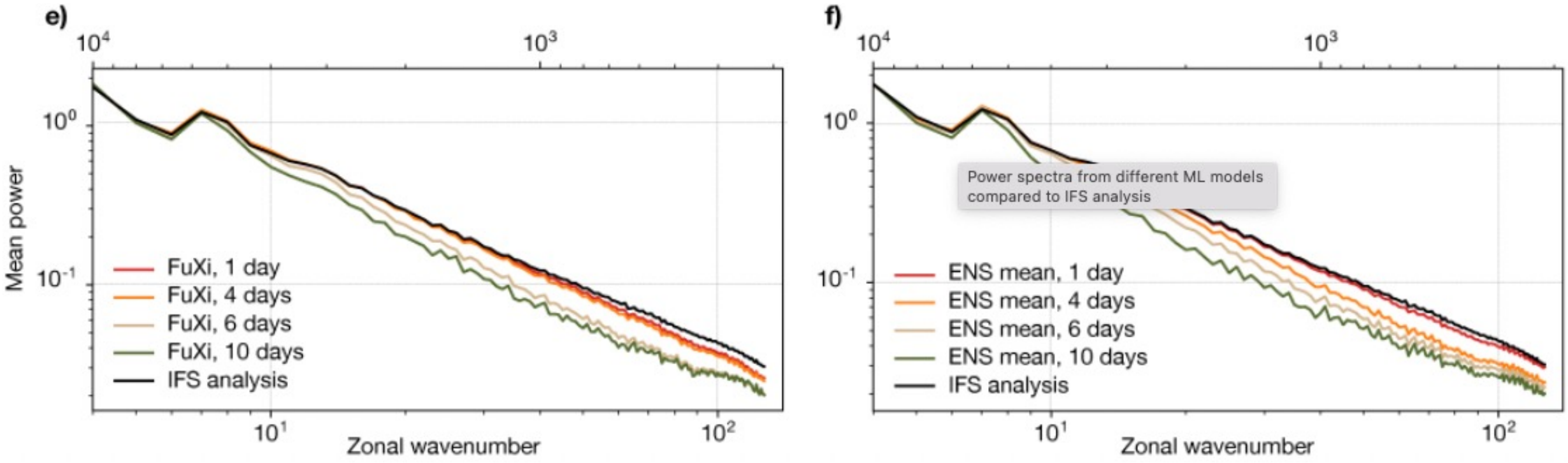
“Cost” of minimising MSE/MAE



“Cost” of minimising MSE/MAE – over a few days

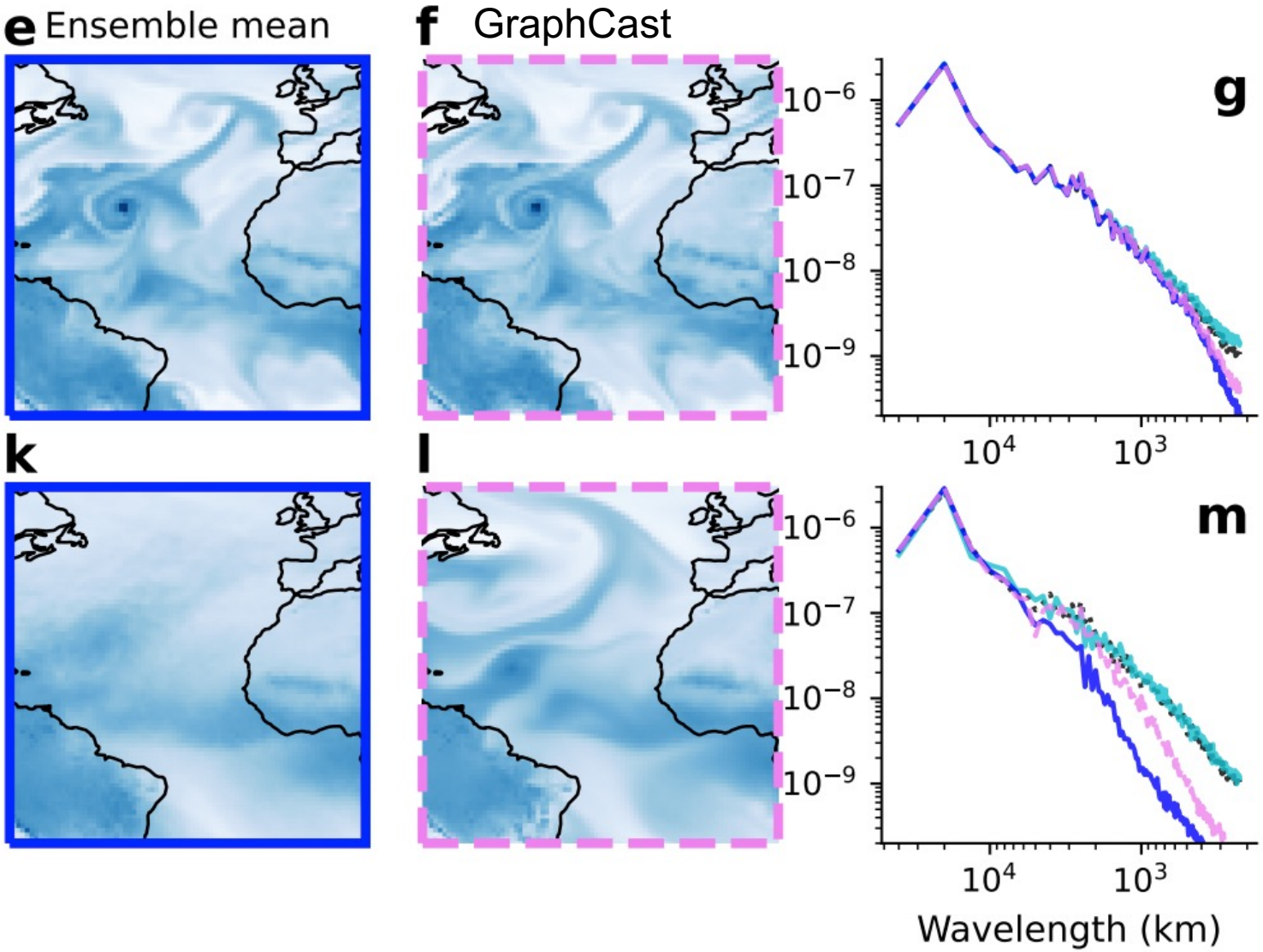


“Cost” of minimising MSE/MAE – over week(s)



Not behaving like a forecast member, but like an ensemble mean. Useful to fewer users?

Value of a probabilistic loss



From GenCast paper
Graphcast in green.

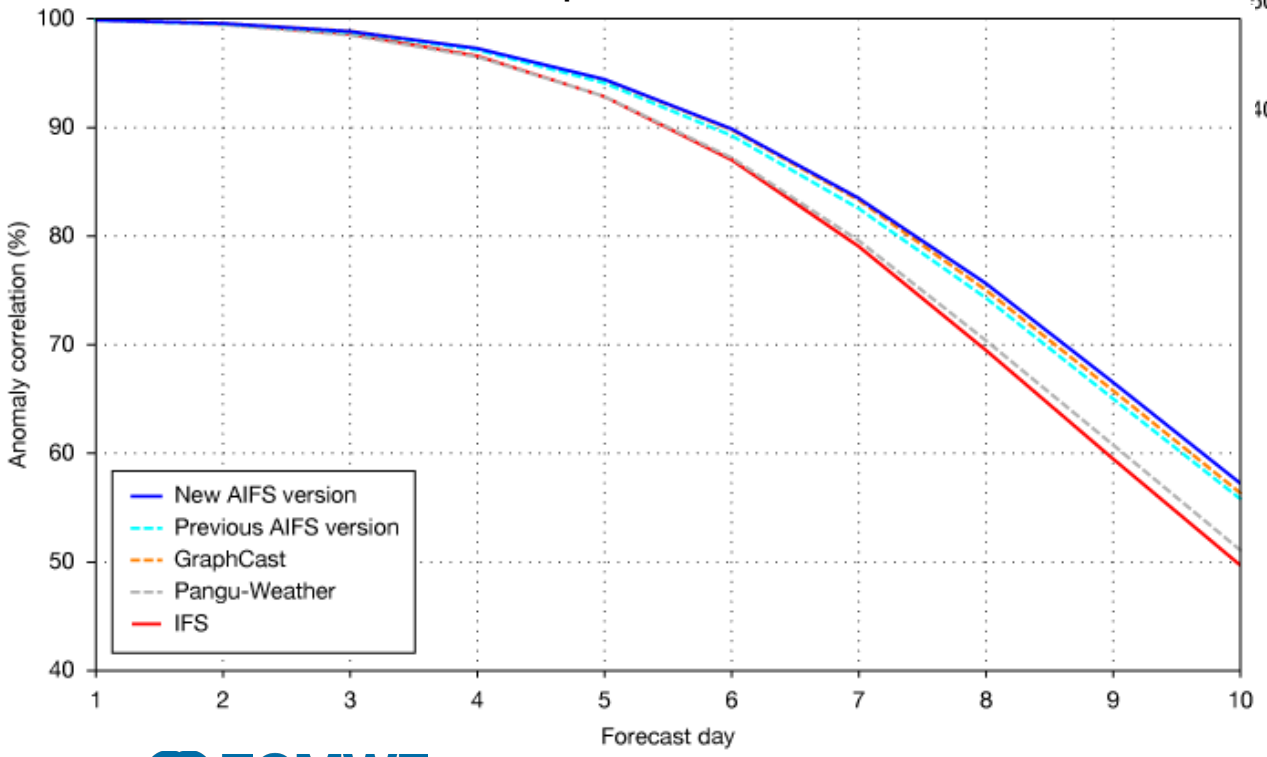
What variables to use?

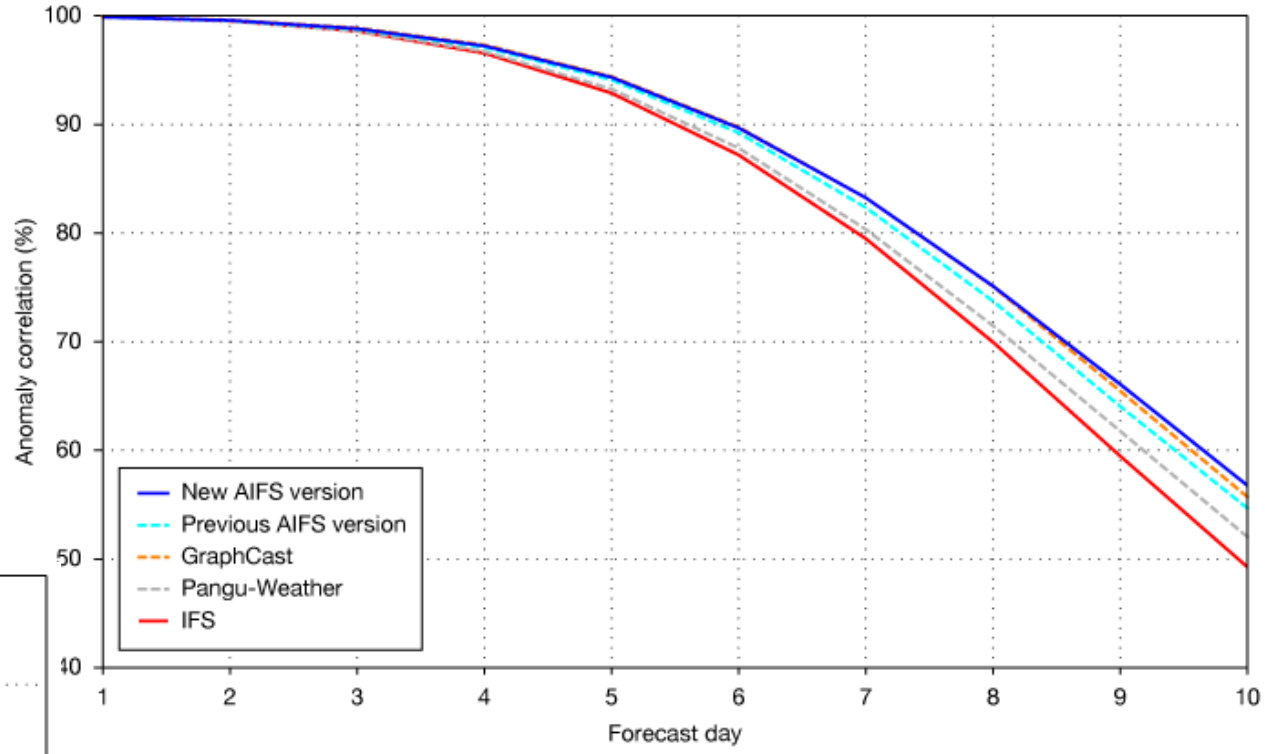
- Two driving motivations:
  - What helps me predict better.
  - What do users want from the system.
    - If high quality data exists, then it can be added directly to the training…

- Typical set used by many models:
  - ~13 pressure level (with model top at 50hPa)
    - Contrast with 137 model levels in the IFS.
    - GraphCast version with 37 levels isn't more skilful than 13 level version.
  - q, t, u, v, z
    - But no direct cloud information.
  - At the surface: 2t, msl/sp, 10u/v and precipitation for some.

How good are these models?

Northern hemisphere z500 ACC



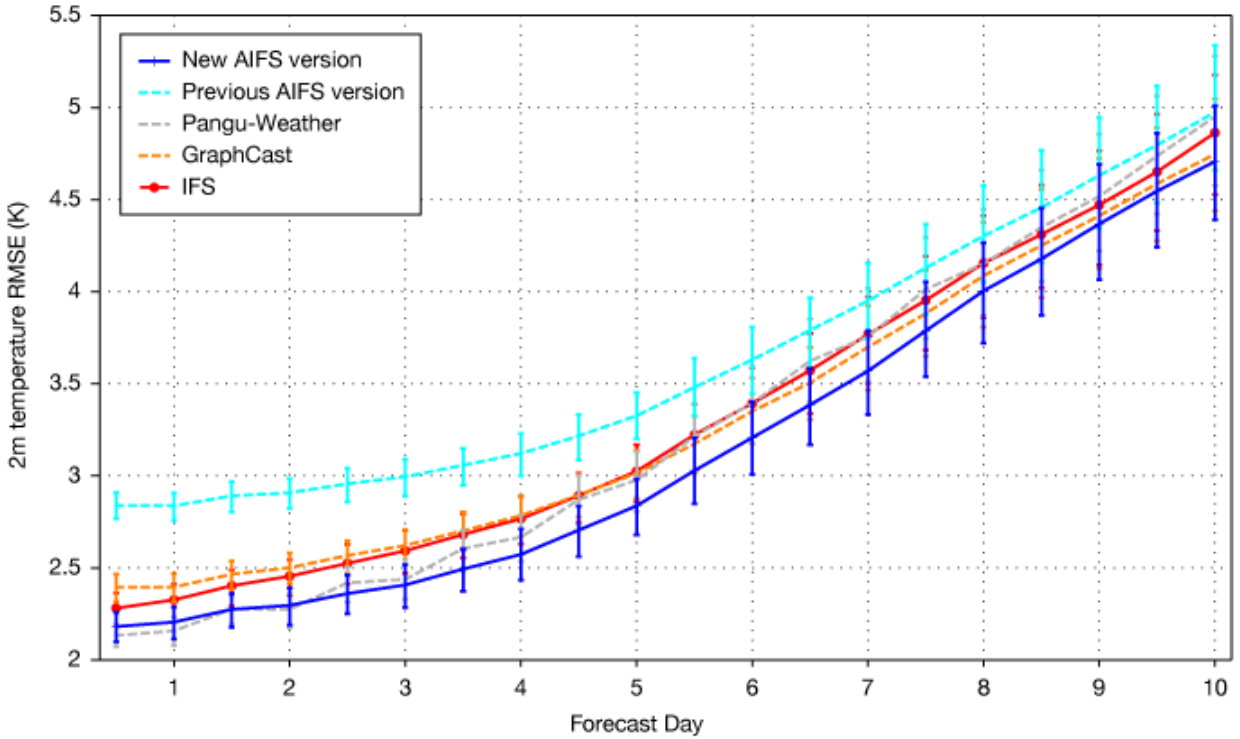


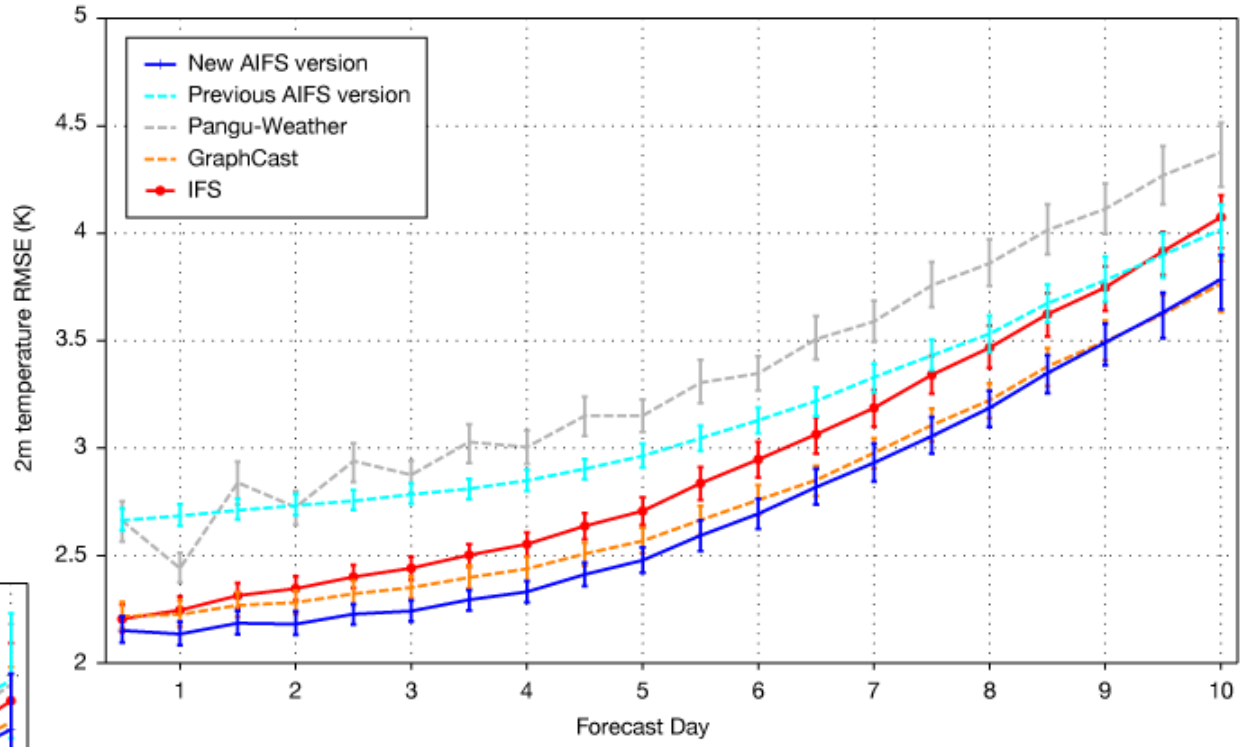
Southern hemisphere z500 ACC

Higher = better

AIFS v0.2 – surface against observations

Northern hemisphere 2m-temperature



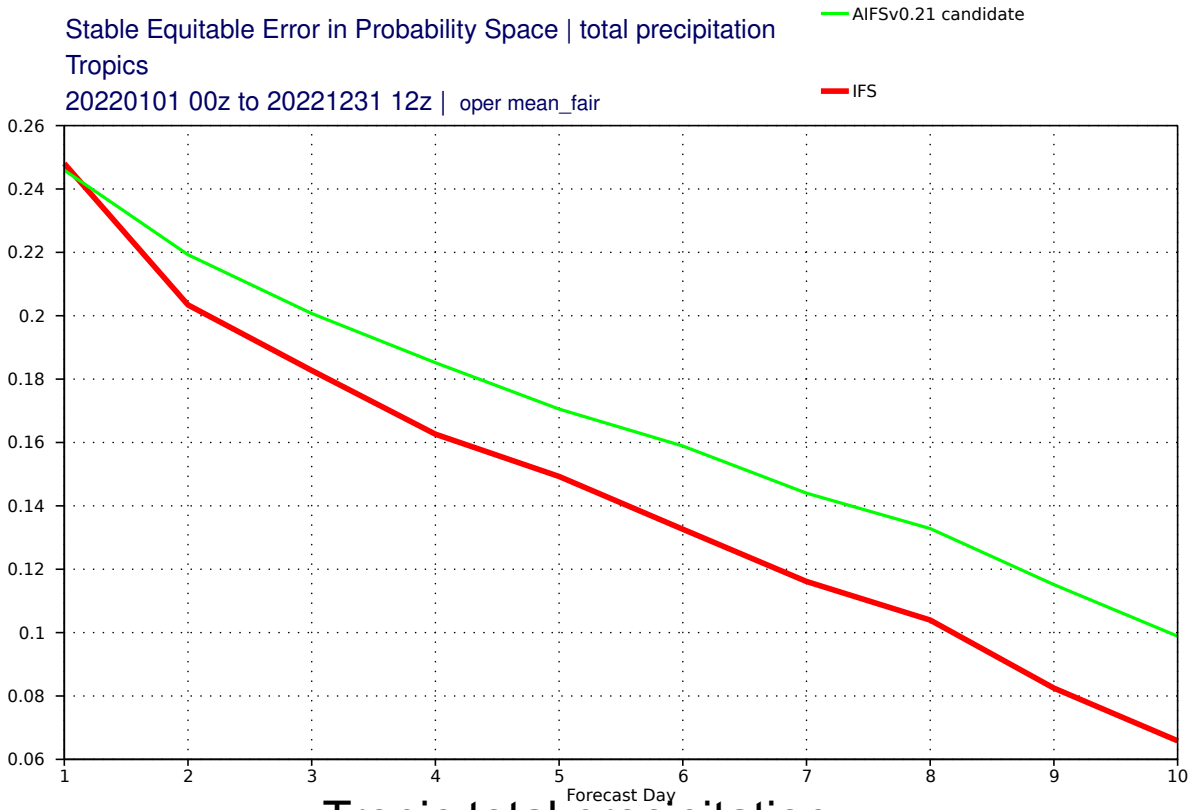


Southern hemisphere 2m-temperature

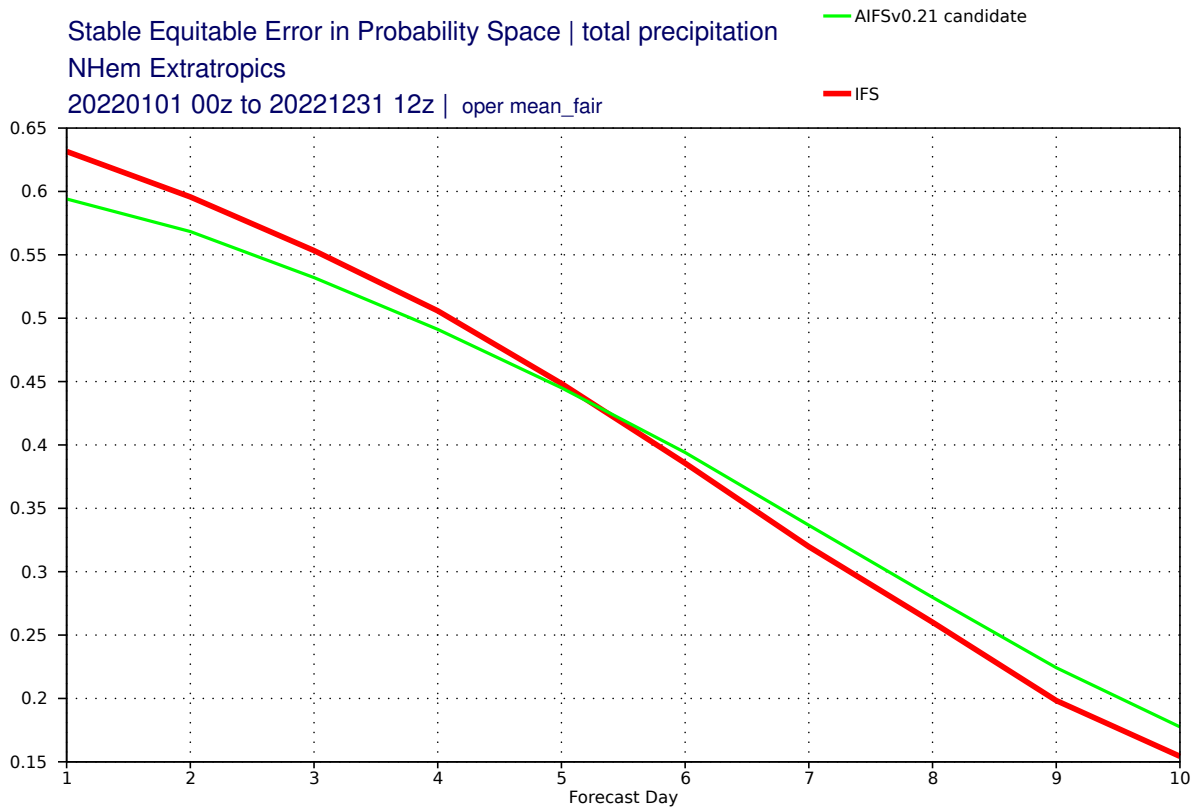
Lower = better

AIFS v0.2.1 – Adding precipitation

- Live from 28th Feb.
- SEEPS for 24h accumulated precipitation against observations.



Tropic total precipitation

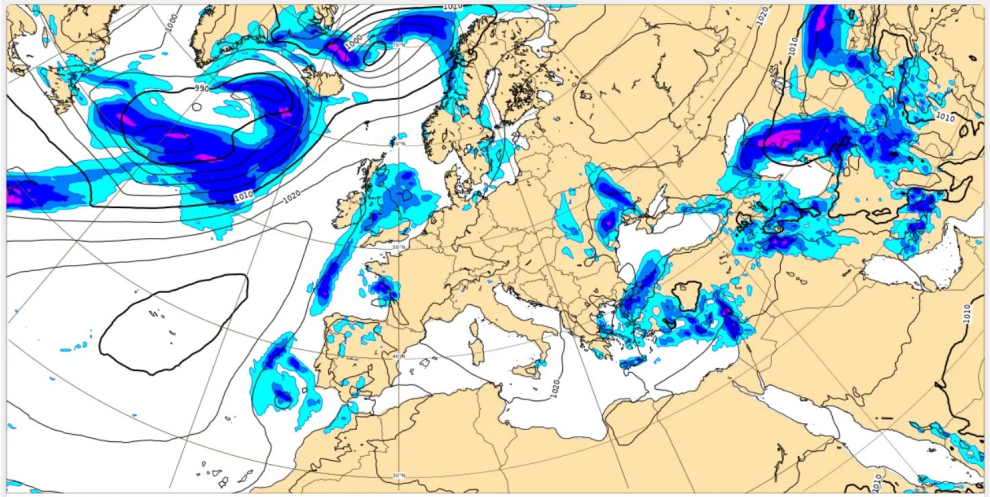


Northern hemisphere total precipitation

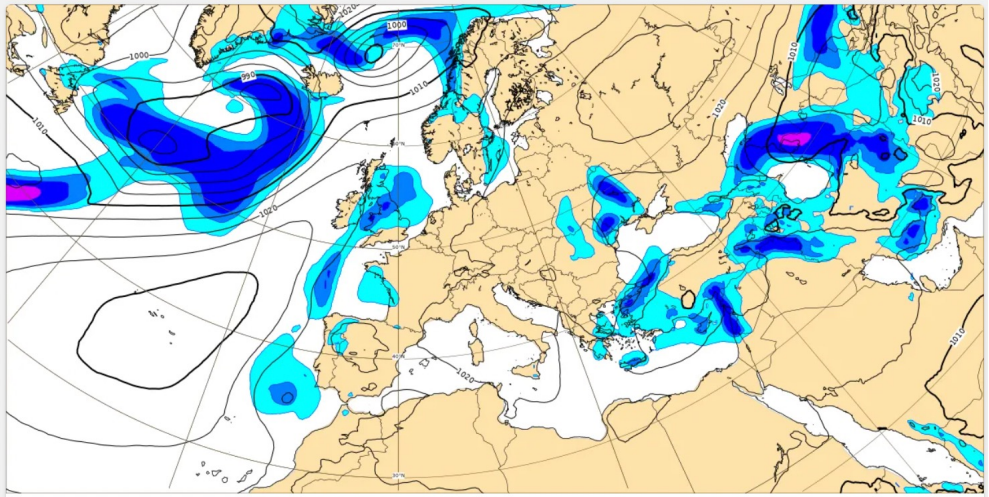
Higher = better

Precipitation currently lacking intensity and small scale structure

+6h

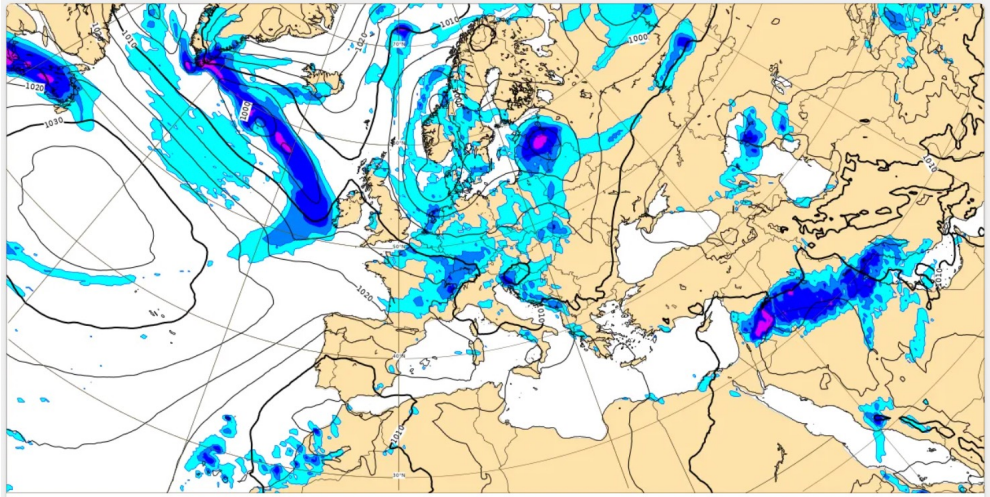


Rain and mean sea level pressure

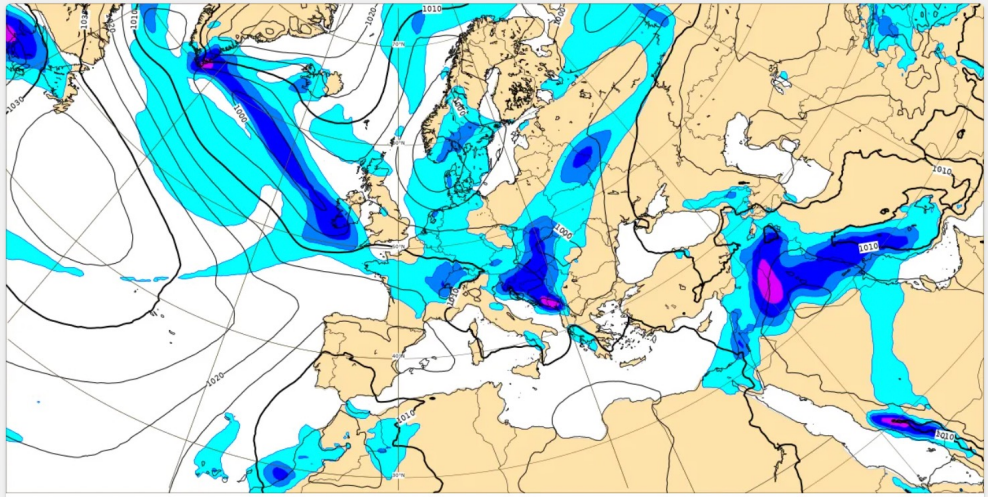


Experimental: AIFS (ECMWF) ML model: Rain and mean sea level pressure

+120h



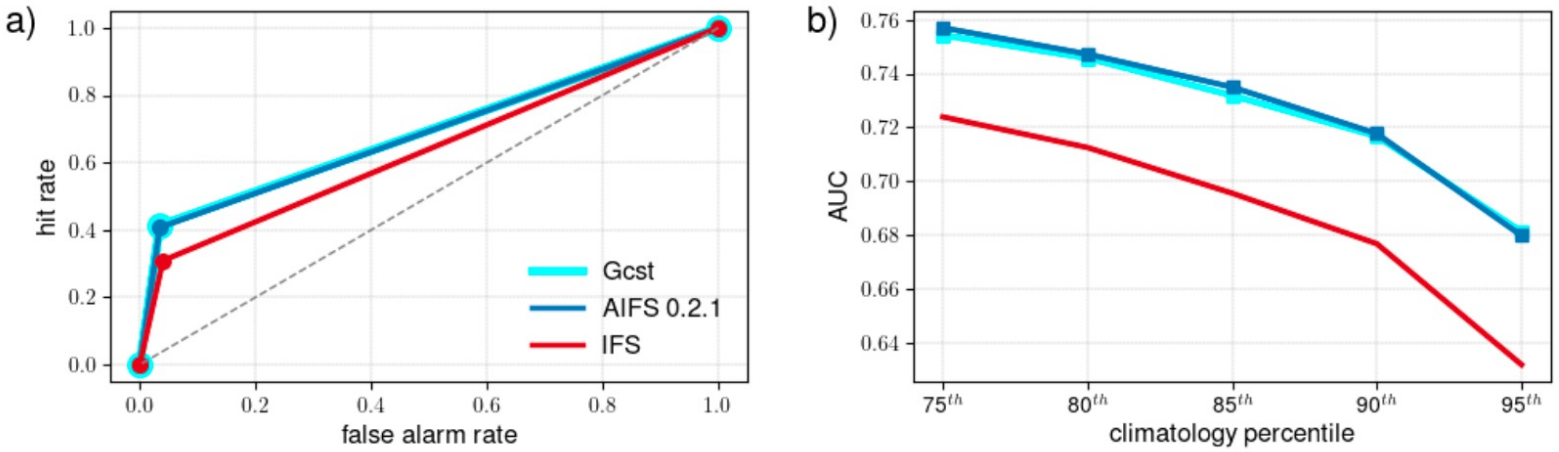
Rain and mean sea level pressure



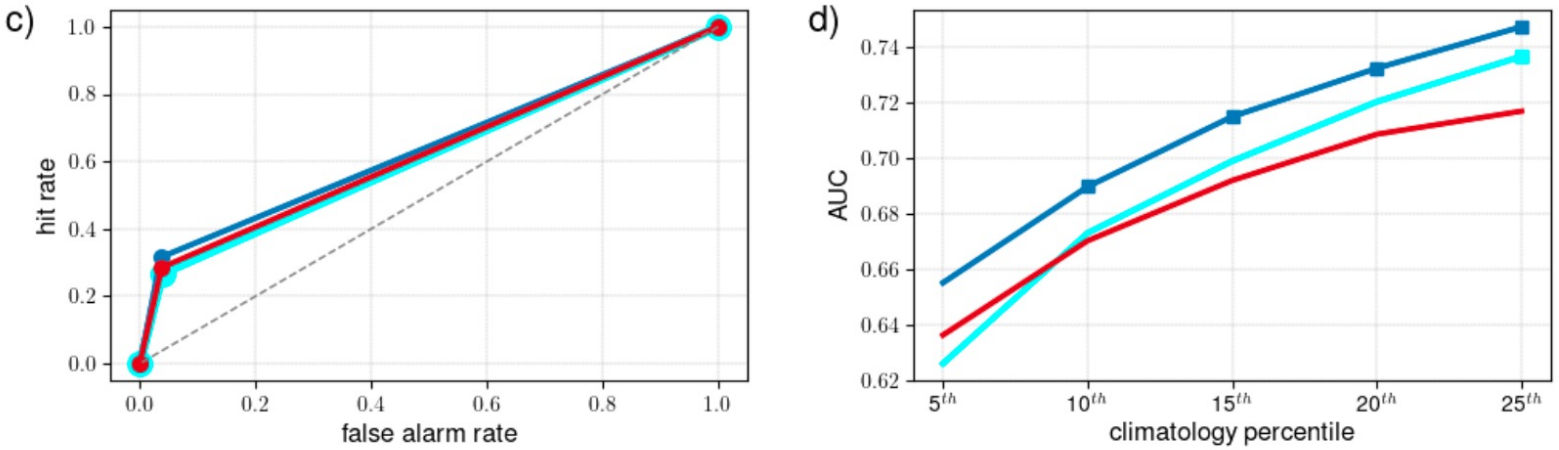
Experimental: AIFS (ECMWF) ML model: Rain and mean sea level pressure

What about “extreme” events? – 2t Summer/Winter extremes

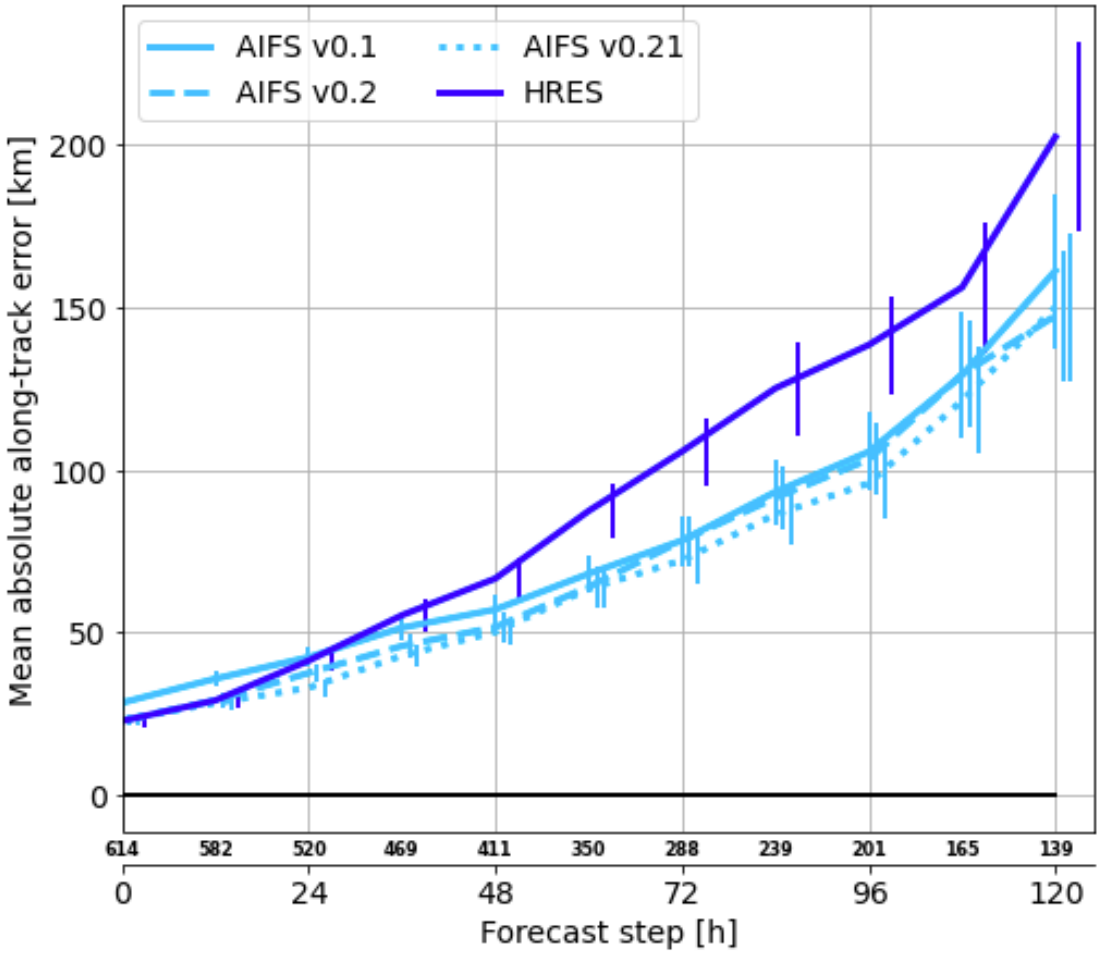
Summer 2022

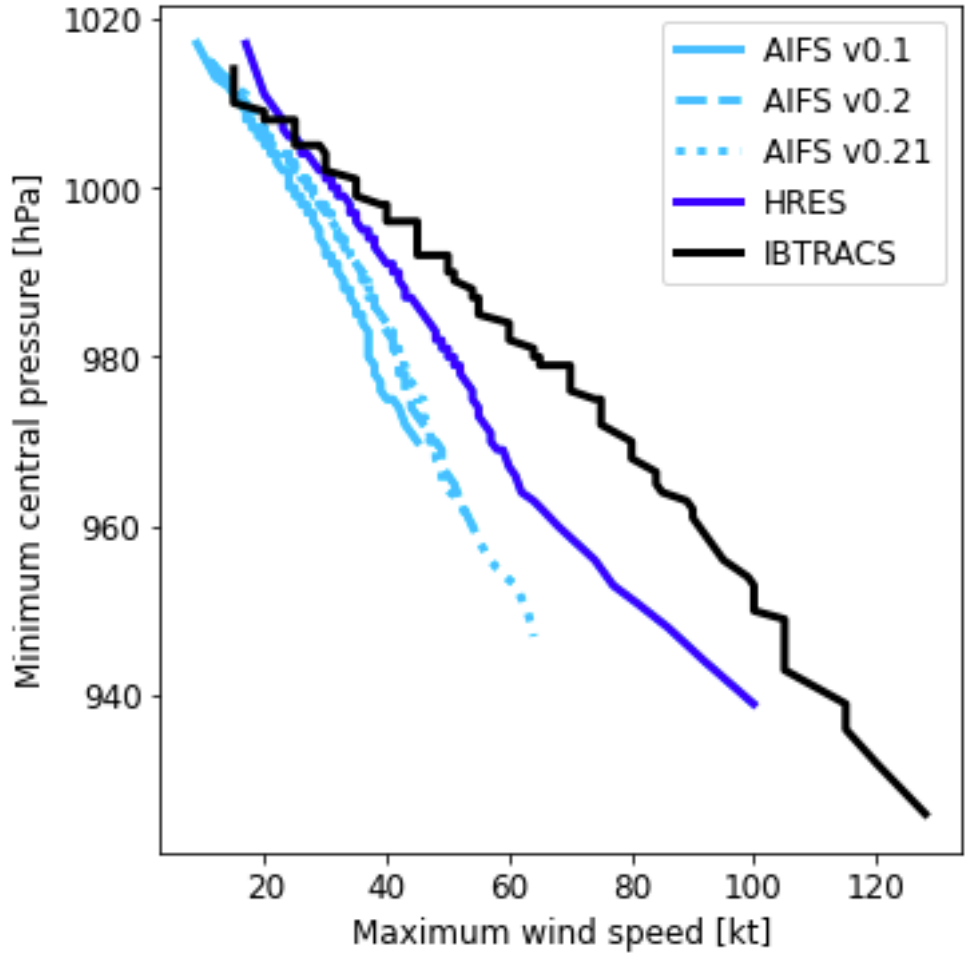


January/February 2022



Tropical cyclones: a tale of two plots





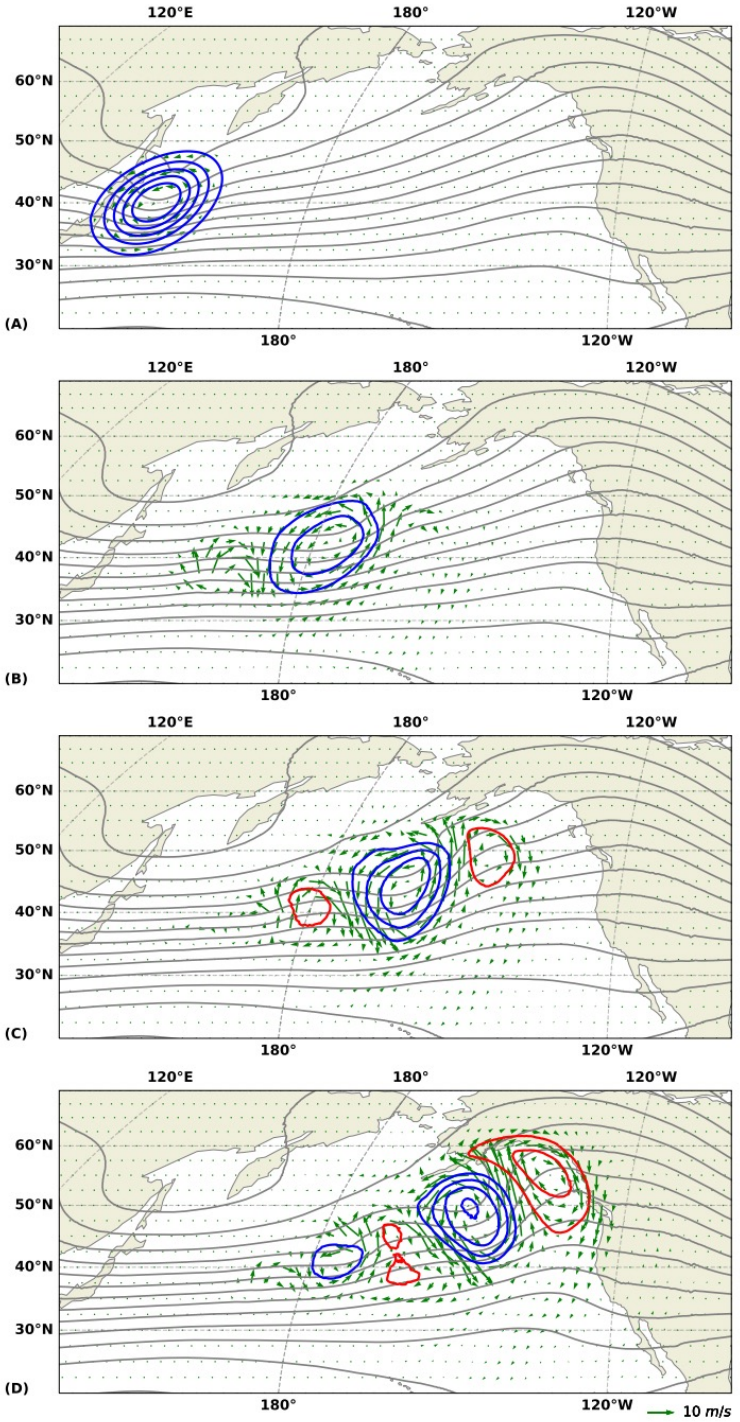
Are data-driven weather forecasts physical?

- Highly recommend reading Hakim & Masanam 2023:
  Dynamical Tests of a Deep-Learning Weather Prediction Model

- Take Pangu weather, and test it on a series of classical dynamical core test cases.
  - Cases need to be applied as deviations from climatology.
  - Apply localised disturbances and study the reaction of the system.

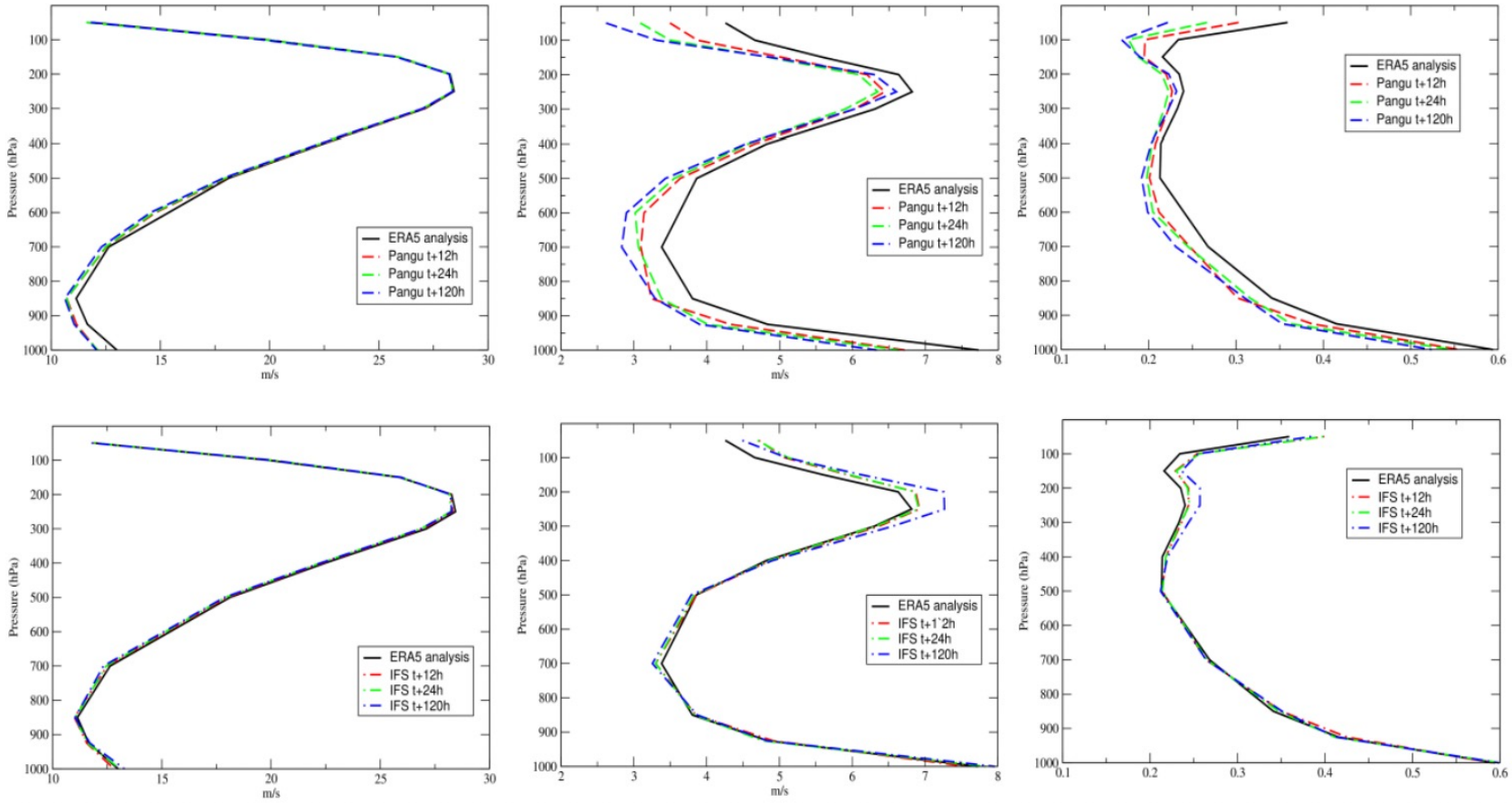
- Overall, Pangu behaves as expected, compared with theory.
  - The 1h model is best for the faster evolving processes, which aren't well captured by the longer timestep model.

- Hopefully we see lots more studies of this type.



Are data-driven weather forecasts physical?

- Bonavita 2023 explore other tests with Pangu Weather.
- Geostrophic balance fairly well represented, but not as well as the IFS.



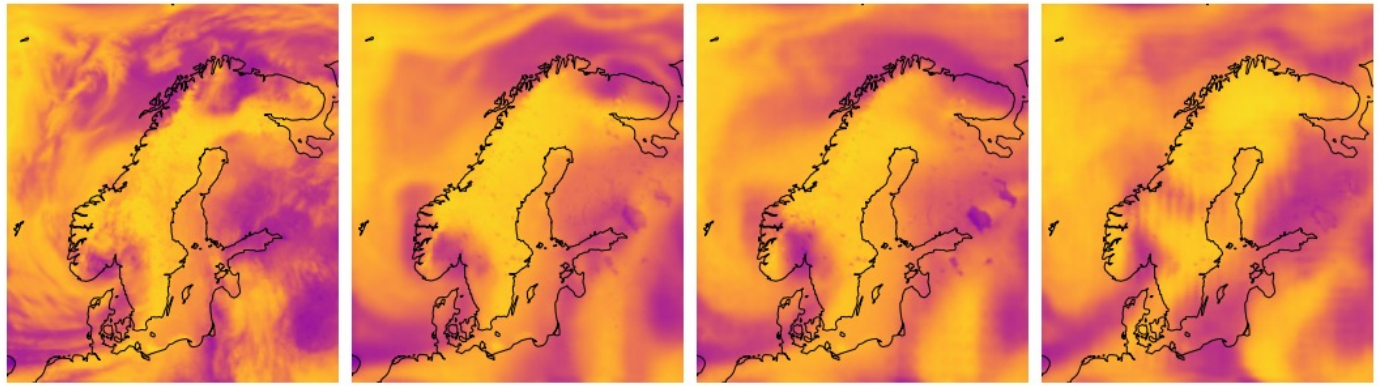
Geostrophic winds | Ageostrophic winds | Ratio

Limited area data-driven models

- NeuralLAM, the first of likely many regional data-driven models.
  - An emulator of the MEPS model, learning to mimic trajectories
  - Uses Graph NN.

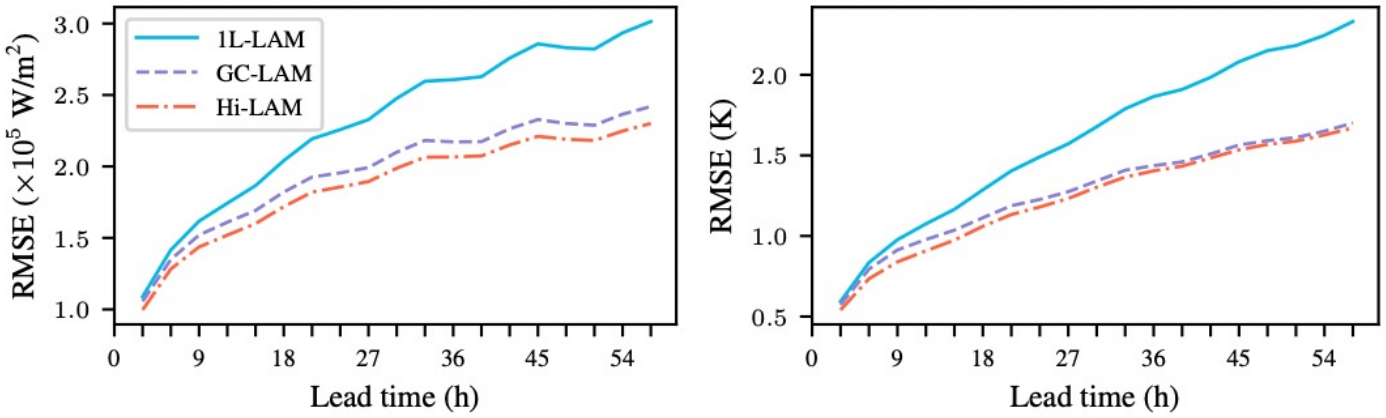
- Several alternate approaches exist for LAM with ML.
  - Stretching a global grid for high resolution over a region.
  - ML downscaling to add resolution.

- Anemoi, the framework used to create the AIFS is being expanded to enable these different methods side-by-side.



(a) Ground Truth (b) Hi-LAM (c) GC-LAM (d) 1L-LAM

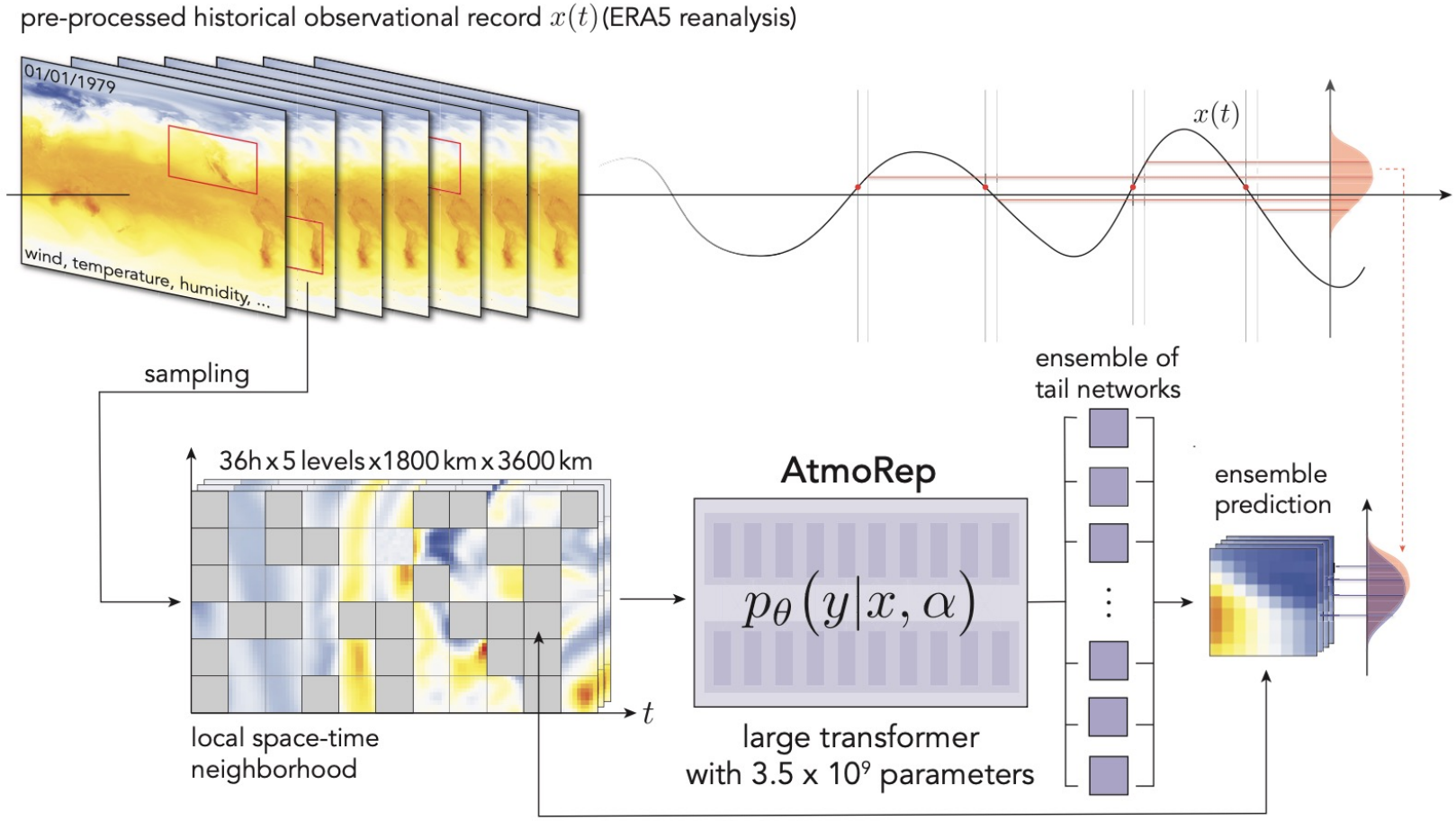
Figure 3: Ground truth and example forecasts of `nlwrs` at lead time 57 h.



(a) `nlwrs` (b) `t_2`

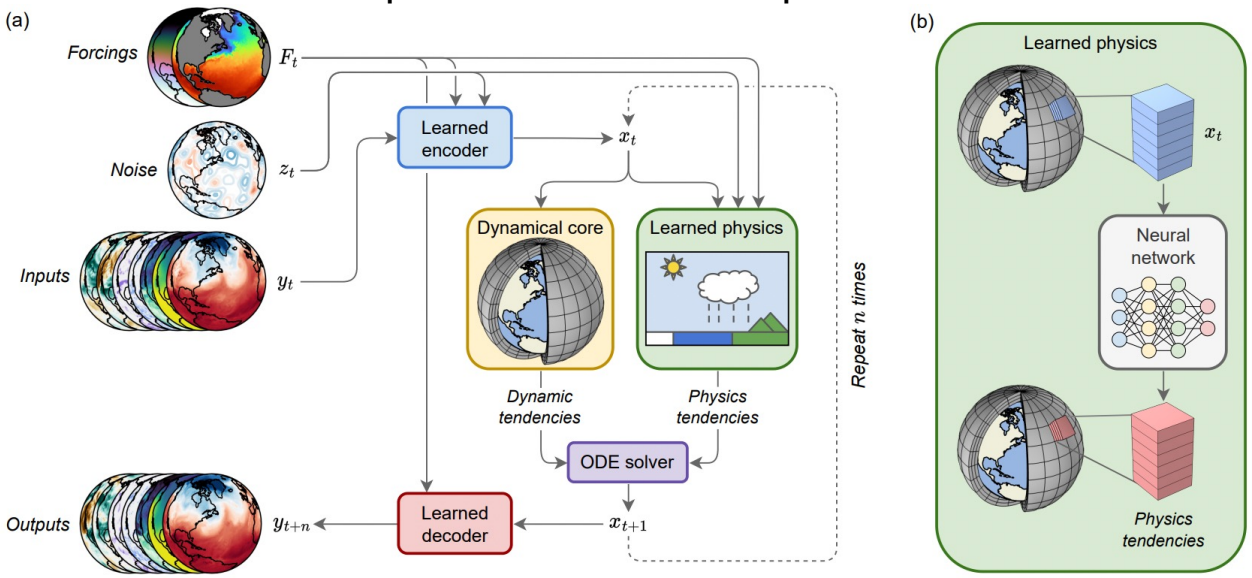
Learning more abstractly – representation learning

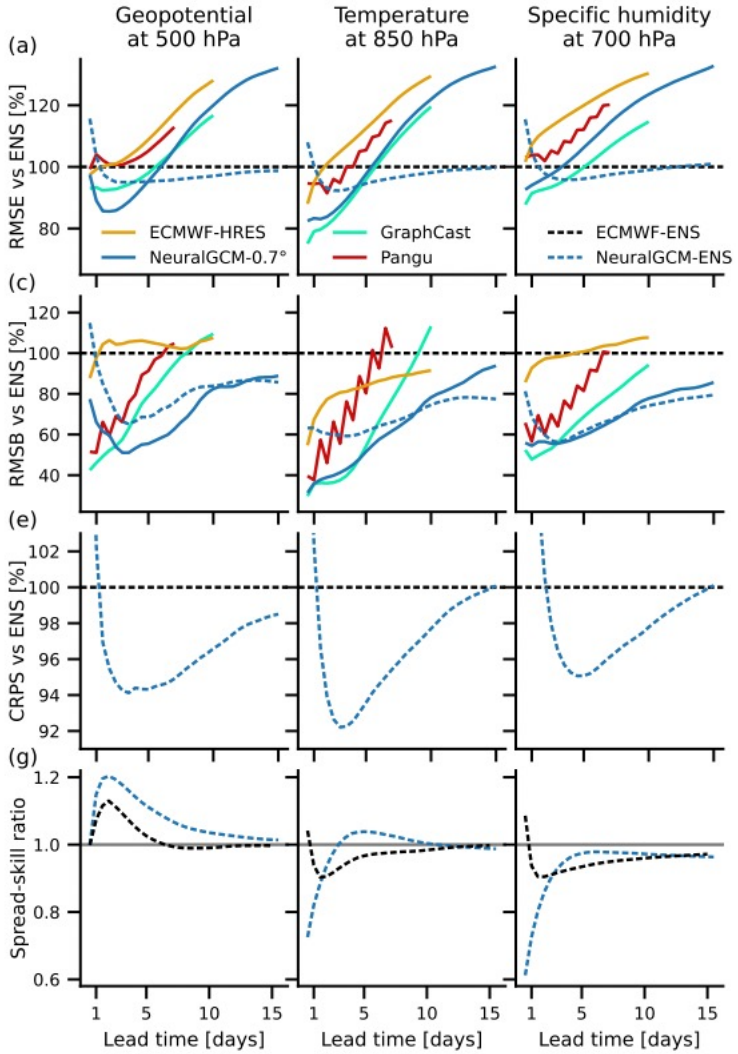
- Learn to fill in space-time gaps of ERA5.
- Forecasting becomes a subtask.



Do you need to learn everything?

- NeuralGCM (Google).
- Write a dynamical core in JAX (differentiable ML framework).
  - Add neural networks with the connectivity to learn local parametrisations.
  - Train the whole thing over time windows.
- Learn very skilful model (including an ensemble).
  - But requires a lot more compute than a data-driven weather model.





Where are we?

- For headline scores, data-driven models are best.
- Don’t represent all the spatial scales correctly when trained deterministically.
  - Still useful, despite this, and probabilistic framing appears to solve this.
- Extreme events
  - a mixed bag, with much more evaluation needed.
- Ensembles
  - first models show a lot of promise
  - but again, more evaluation needed

Where is the field going?

- Earth system data-driven models
  - Capture land, ocean and more processes.

- Extended range predictions, pushing beyond 2 weeks.

- Use of observations to predict the future state.
  - Incorporate data-assimilation into the training.

- Collaboration between ECMWF and MS on data-driven models.
  - Opportunities to share code/infrastructure whilst still having bespoke models.

What do you think the future will hold?