Reanalysis: the bedrock of machine learning for the Earth system

Matthew Chantry

Zied Ben Bouallegue, Linus Magnusson, Simon Lang, Mark Rodwell, Mariana Clare, Mihai Alexe, Jesper Dramsch, Baudouin Raoult, Florian Pinault, Florian Pappenberger and many many more

Please write me an introduction for a presentation about machine learning using ERA5 for the european centre for medium-range weather forecasts annual seminar.

Ladies and gentlemen, esteemed colleagues, and distinguished guests,

I am honored to stand before you today at the European Centre for Medium-Range Weather Forecasts (ECMWF) Annual Seminar to discuss a topic of paramount significance in the field of meteorology and climate science—Machine Learning and its transformative role in harnessing the power of ERA5 data.

As we gather here to delve into the cutting-edge advancements in our discipline, it is evident that the intersection of meteorology, machine learning, and the invaluable ERA5 dataset has the potential to reshape the landscape of medium-range weather forecasting. ERA5, the latest and most comprehensive reanalysis dataset provided by the European Centre for Medium-Range Weather Forecasts, represents a treasure trove of high-quality atmospheric and surface information, spanning four decades. This vast repository of climate data offers an unparalleled opportunity to push the boundaries of meteorological science and improve the accuracy and precision of our weather predictions.

What’s behind the boom in machine learning?

- Big advances in machine learning architectures, training algorithms and frameworks.
  - Transformer architectures.
  - Self-supervised training, diffusion modelling.
  - PyTorch (and others) with huge community and big-tech support.
- Big advances in computational power.
  - GPUs and other accelerators.
  - NVIDIA (and others) investing in improving the efficiency of basic algorithmic components.
- Consuming huge amounts of data.
  - OpenAI (and others) have trawled the web for all text/images etc.
  - Bad data will lead to bad models.

Dall-e 2, prompt: European Centre for Medium-Range Weather Forecasts Annual Seminar

What kind of things are people using ML to do in the earth system?

Just about everything!

What’s behind the boom in machine learning?

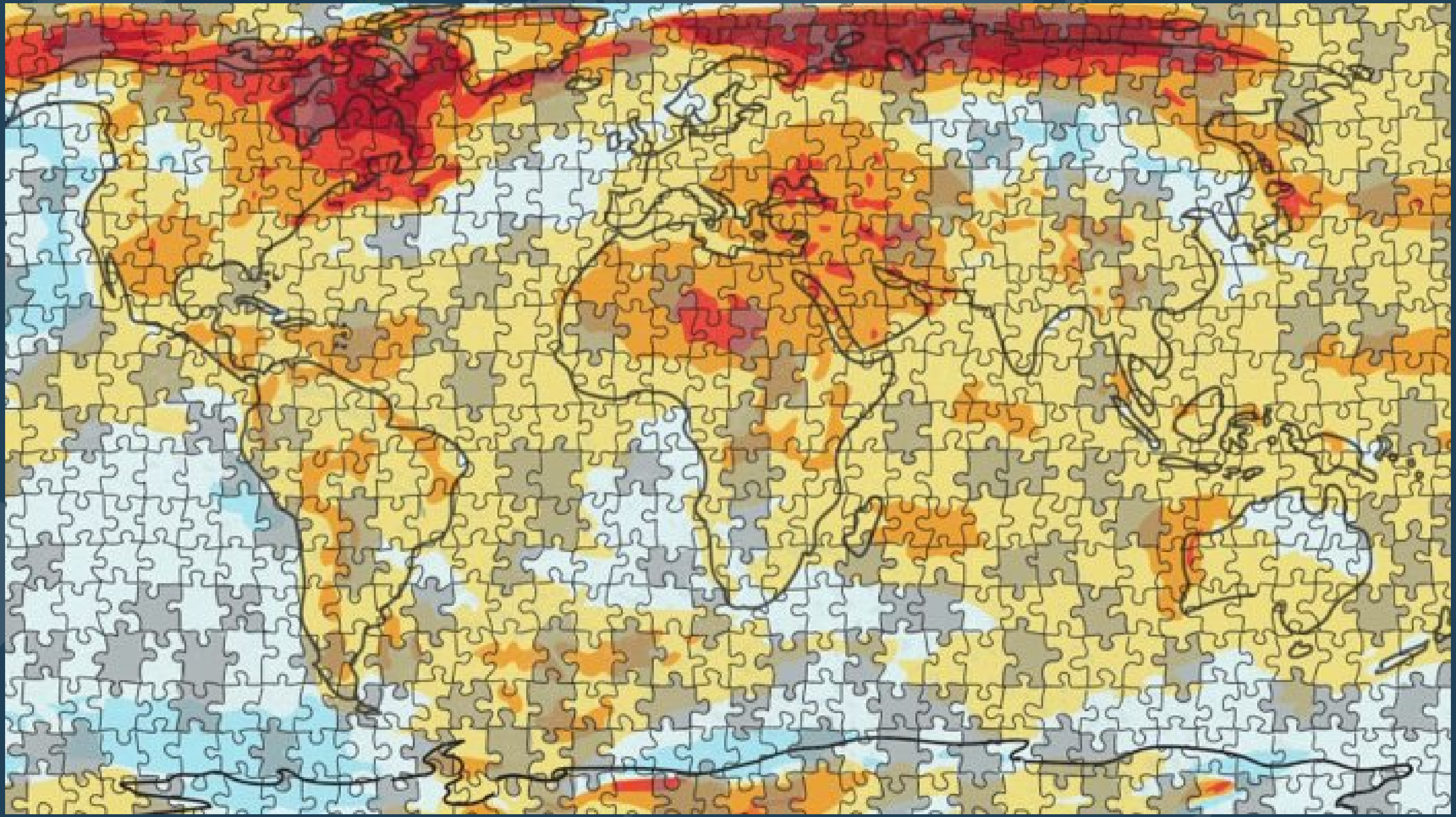
- Big advances in machine learning architectures and training algorithms.

- Big advances in computational power.

- Consuming huge amounts of data.
  - Learning from observations is hard.
    - Data is stored in many different places, and formats.
    - Observations change over time).
    - Data has biases and errors.
    - Requiring multiple variables can mean using multiple observation datasets.

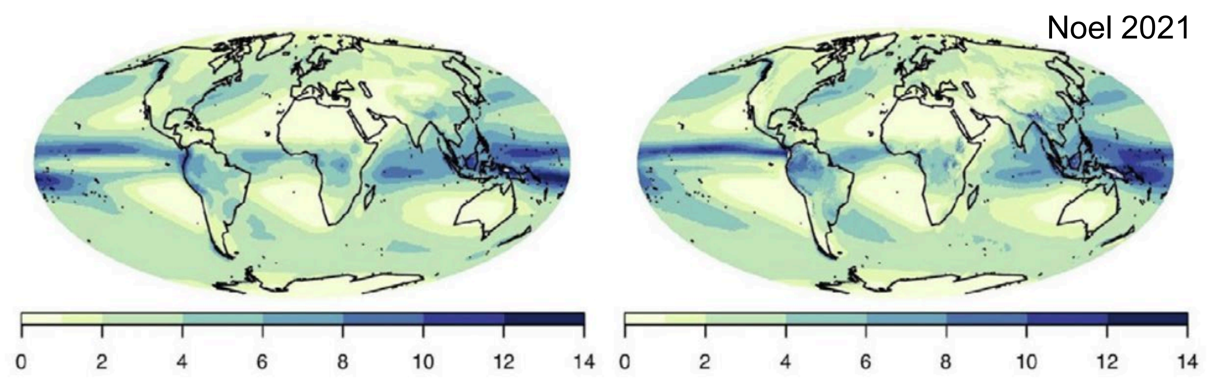
Reanalysis for machine learning

- Reanalysis provides singular point of “truth”.
- Many variables.
- All times.
- All points in space.
- All accessible from one access point.



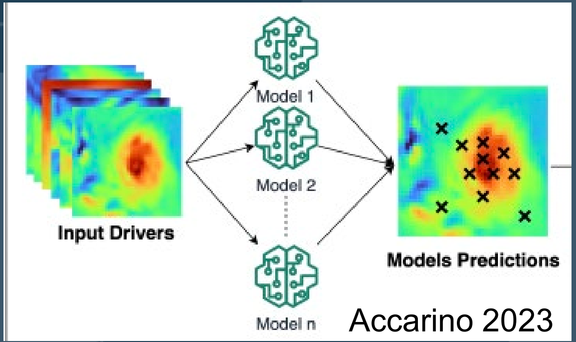
What are people doing with reanalysis and ML?

- Postprocessing to remove bias, calibrate forecasts…
- Downscaling to increased resolution of weather/climate models.



Noel 2021

- Learn observation operators.
- Hurricane forecasting.



Accarino 2023

- **Predict the weather…**

~10k references to ERA5, ~2k mention machine learning…

Is reanalysis sufficient to learn a global forecasting system?

Simple problem framing.

- Given state of ERA5 at a random point in time, $x(t)$.
- Construct a model F, a neural network parametrised by weights.
- Predict a future state of ERA5, $x(t+dt) \backsimeq F(x)$.
- Seek to minimise $[x(t+dt) - F(x(t))]^2$ using gradient descent.
  - i.e. change the weights in such a way to decrease the MSE.
- Randomly draw a new x and repeat.

Is reanalysis sufficient to learn a global forecasting system?

Simple problem framing.

- Given state of ERA5 at a random point in time, x(t). Typically u, v, t, z, q on ~10 pressure levels and 2t, 10u/v, sp.
- Construct a model F, a neural network parametrised by weights. Big models, $O(10^7)$ parameters.
- Predict a future state of ERA5, x(t+dt) $\simeq$ F(x). Typically 6-hour timestep!
- Seek to minimise $[\, x(t+dt) - F(\, x(t)\,)\,]^2$ using gradient descent.
  - i.e. change the weights in such a way to decrease the MSE.
- Randomly draw a new x and repeat. Many many times, passing through ERA5 O(100) times.

a very busy and FAST evolving landscape

2018 ECMWF's ML scientific publication

ECMWF's
Peter Dueben and Peter Bauer publish a paper on using ERA5 at ~500km resolution to predict future z500.

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

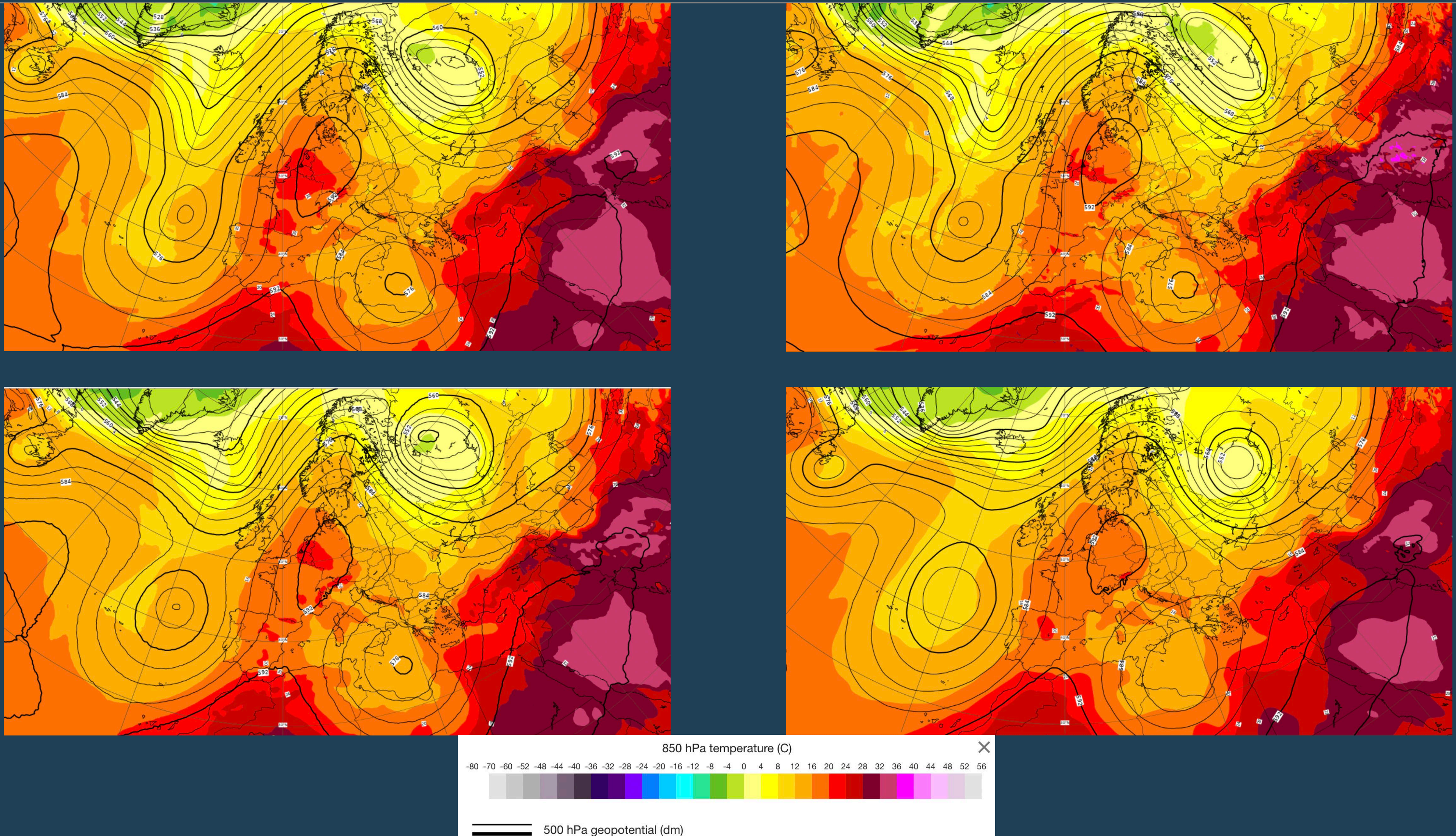
Diffusion modelling

Alibaba – SwinRDM
0.25° 6-hour product

Sharp spatial features

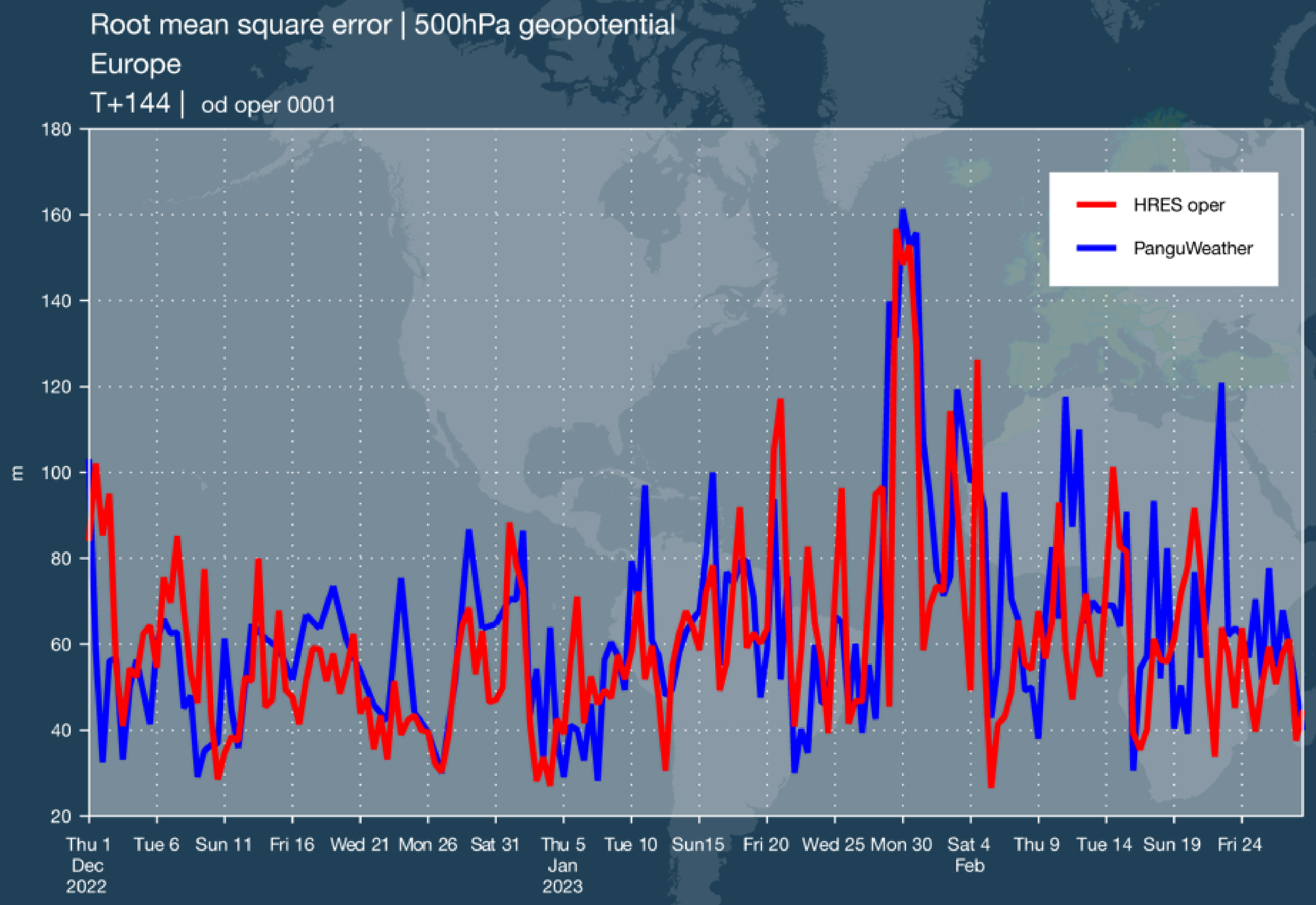
What ML models are showing...

Day 4 forecasts over Europe (valid today, 7 Sept 2023 12UTC)

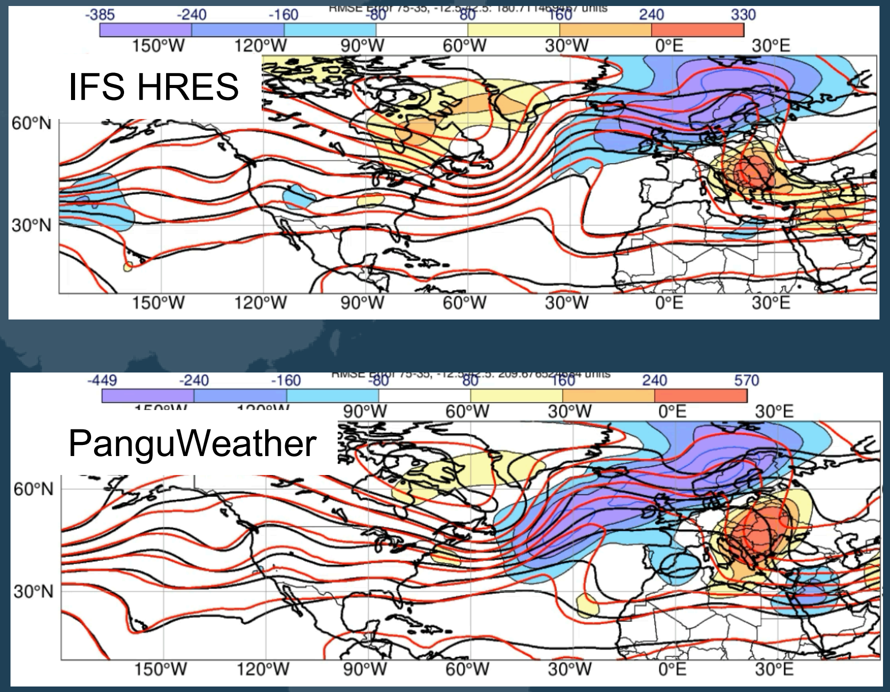


Time-series of day 6, RMSE over Europe

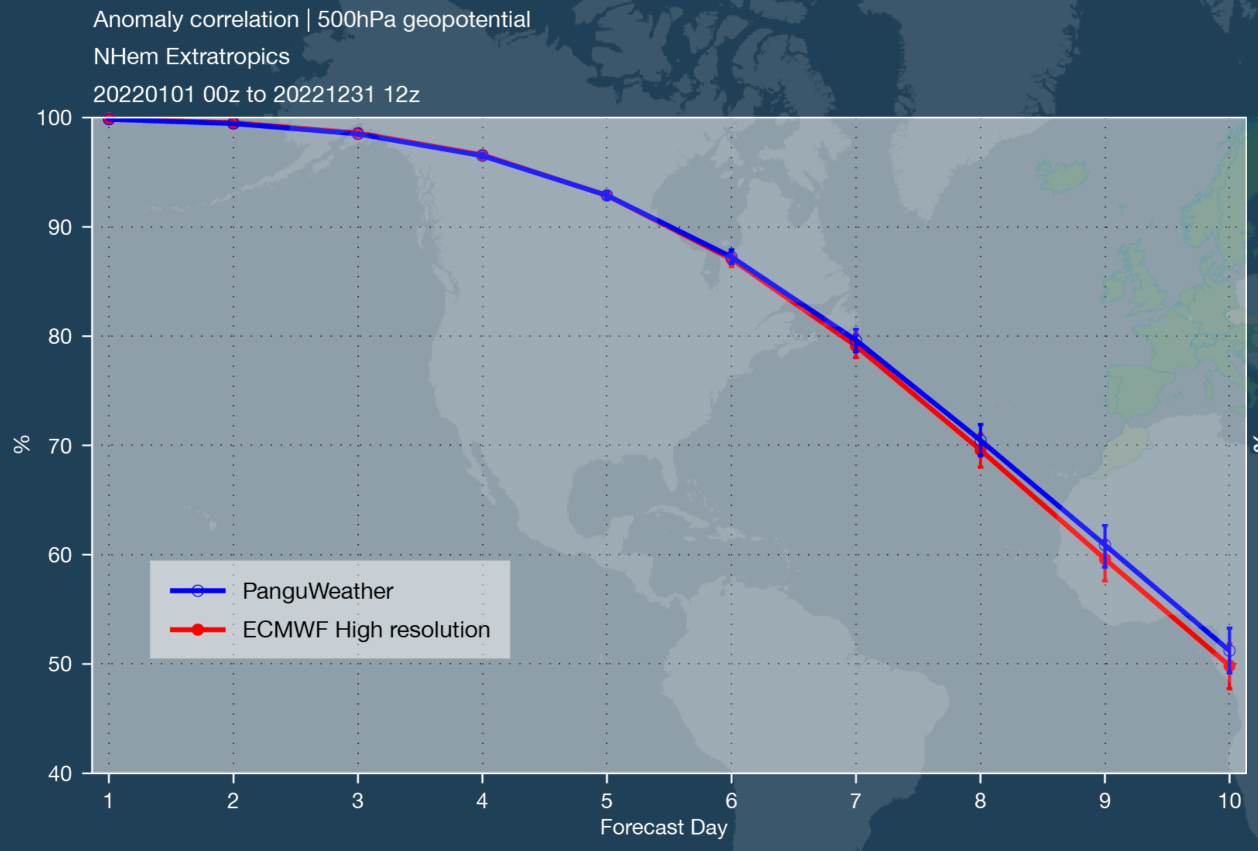
Same starting point....similar results

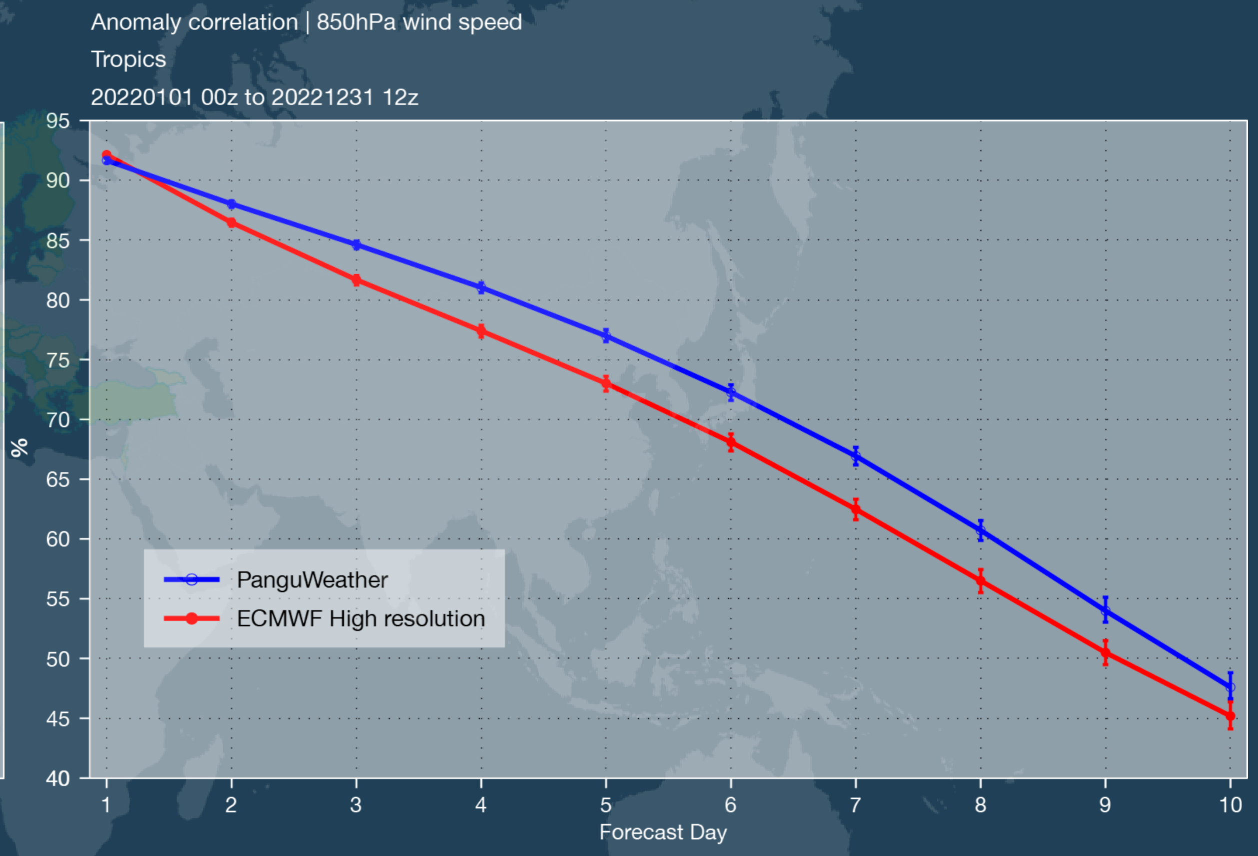


+144h forecast errors 30 January 00UTC

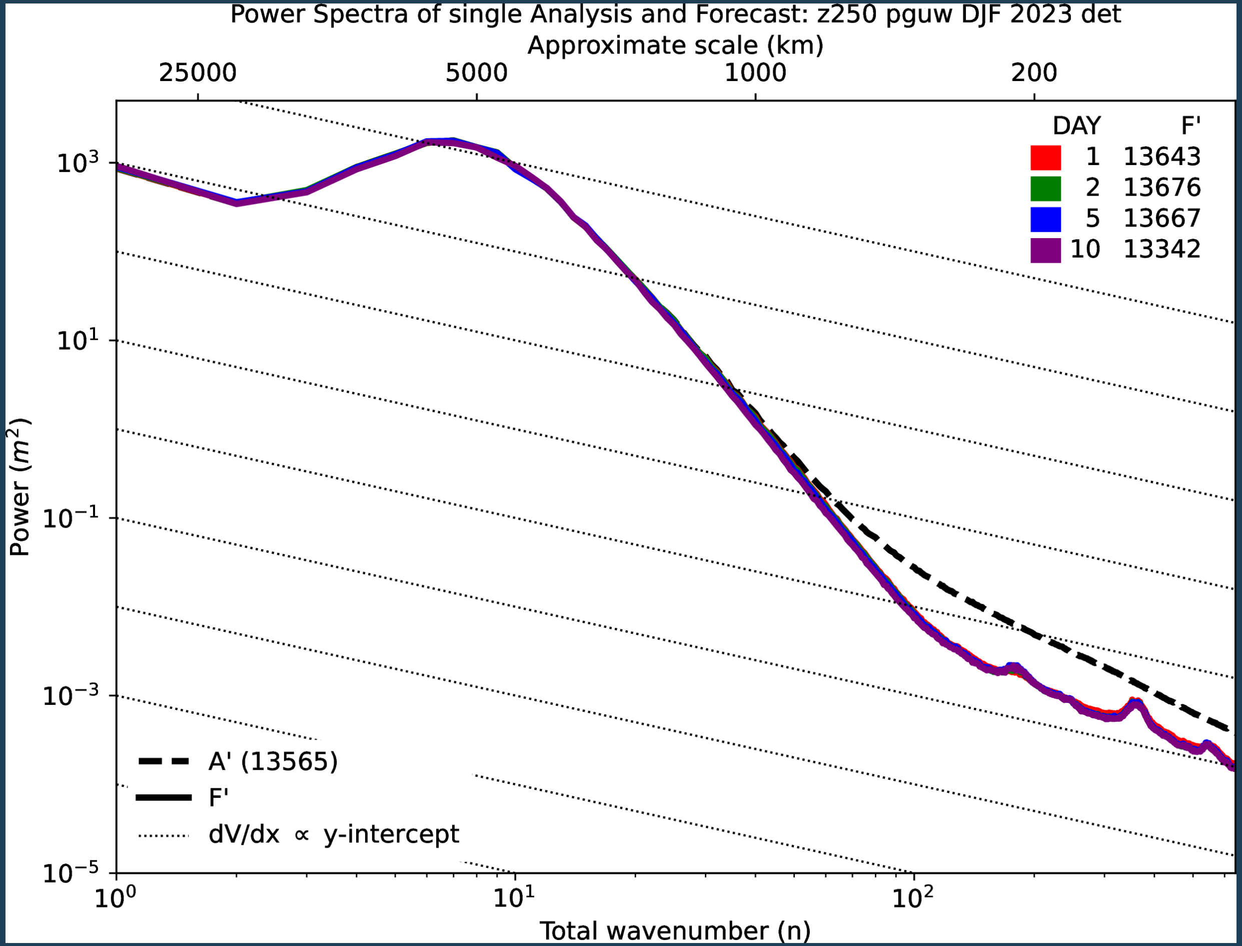


What the analysis is showing: **an undeniable skill**





Digging into the scores: RMSE, bias and forecast variability DJF 2022/2023



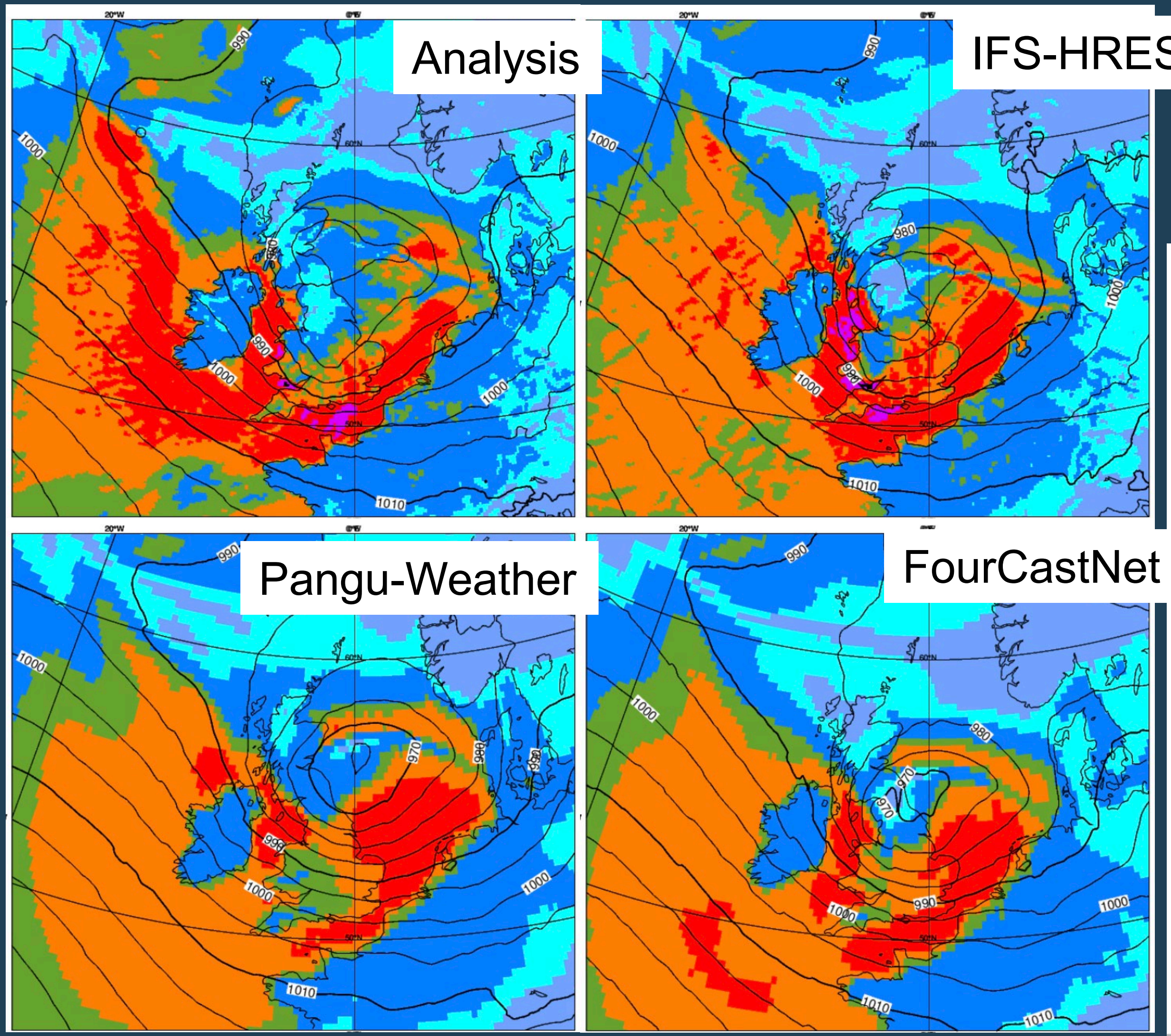
Why is the RMSE lower in PanguWeather?

- Not a clear reduction in forecast activity in PanguWeather
  - …but smoothing of small scales
- Strong model drift in Pangu
  - …but regional biases are improved

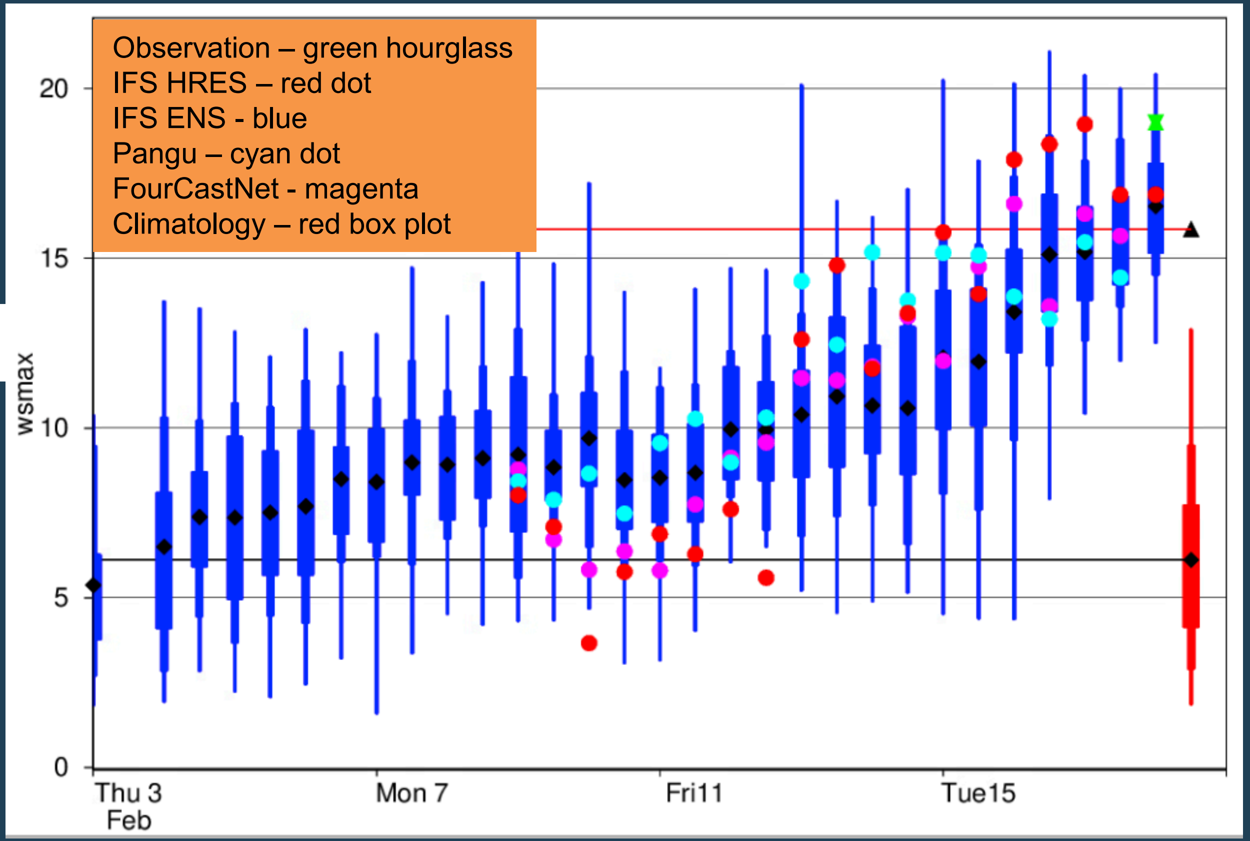
Scope for further investigations to understand the differences

What about high-impact events?

Storm Eunice (2.5-day forecasts valid18th Feb 2022 12UTC)

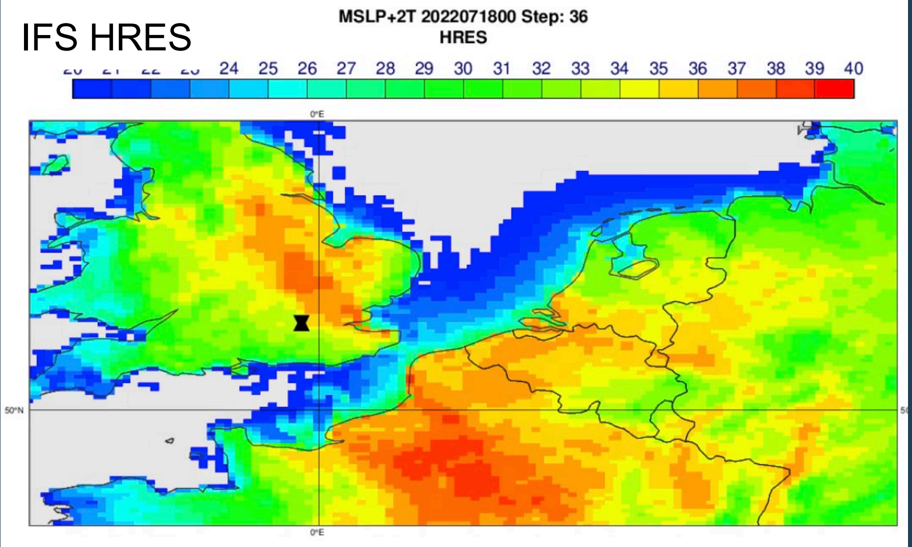


Maximum mean wind Heathrow 18 Feb (00, 06, 12 and 18

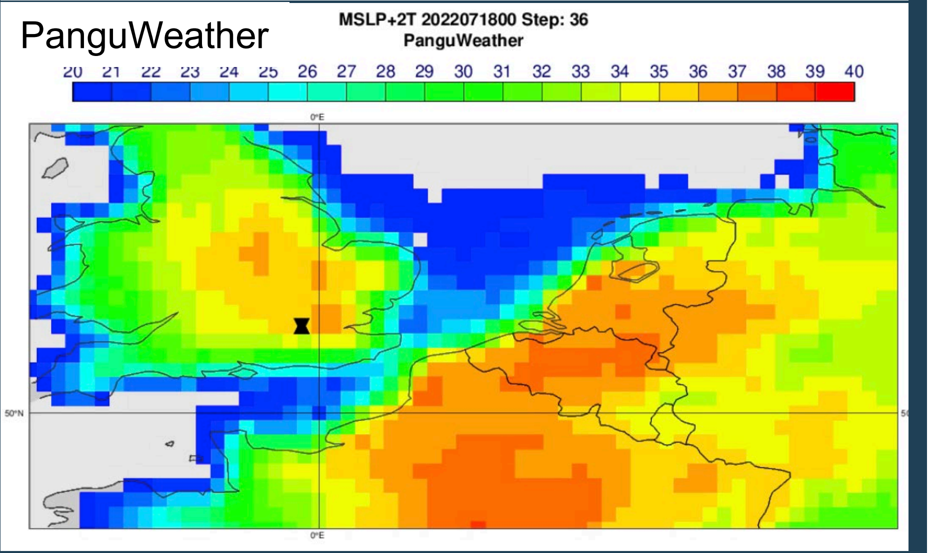


UK heatwave 2022

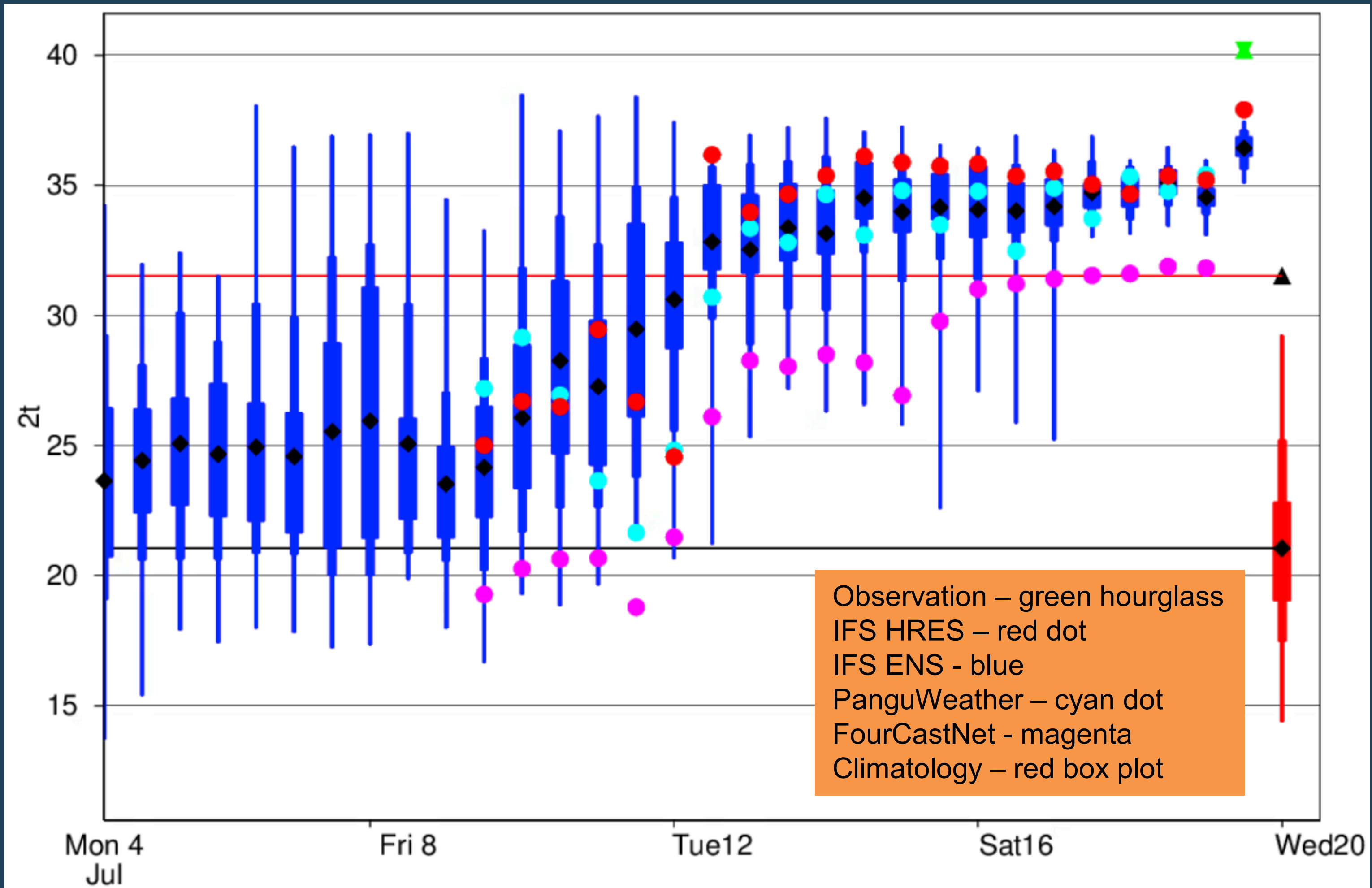
IFS HRES



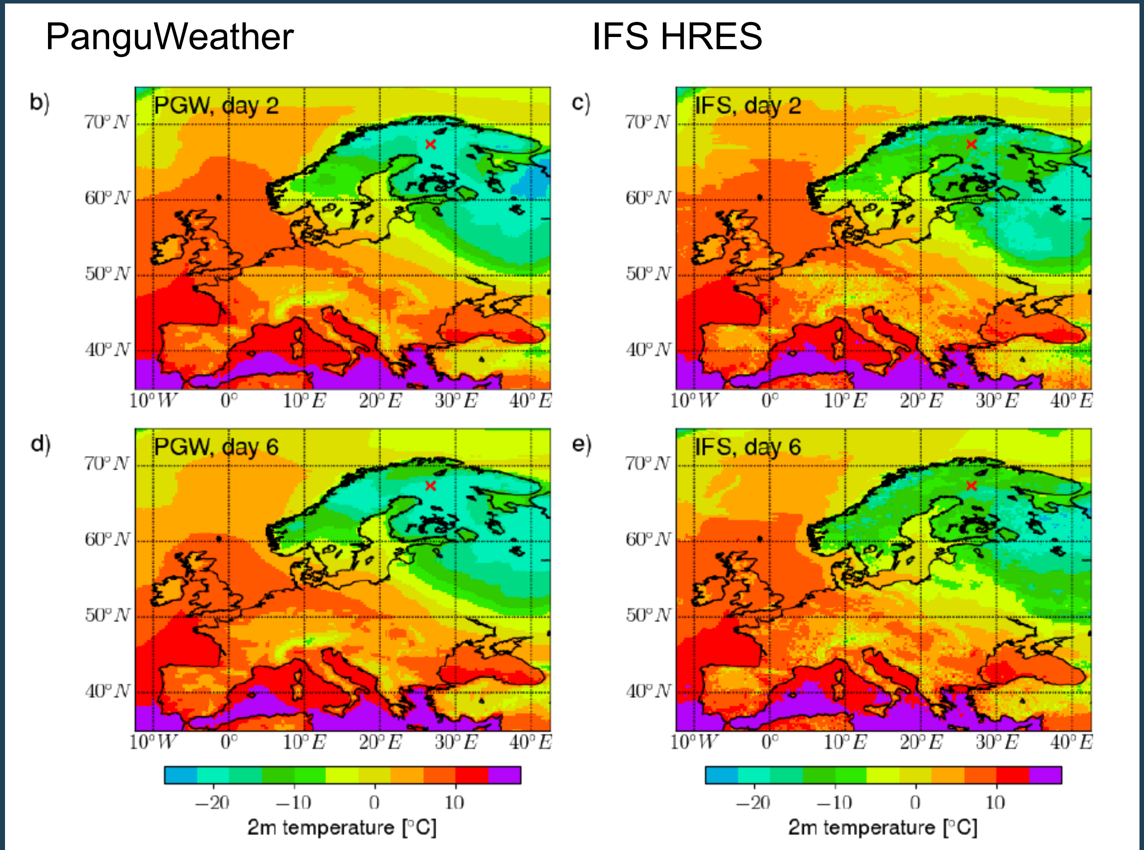
PanguWeather



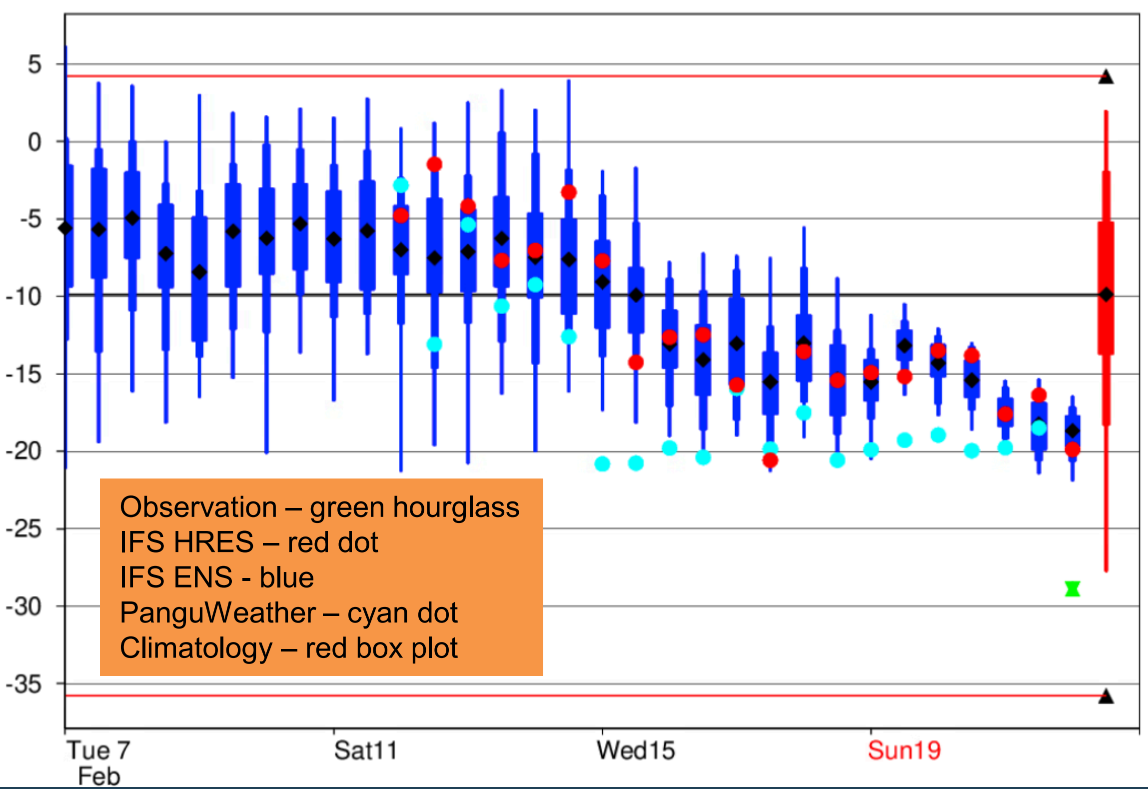
2m temperature Heathrow 19 July 12UTC



Cold snap over northern Europe Feb 2023



2m temperature Sodankyla 22 February 00UTC

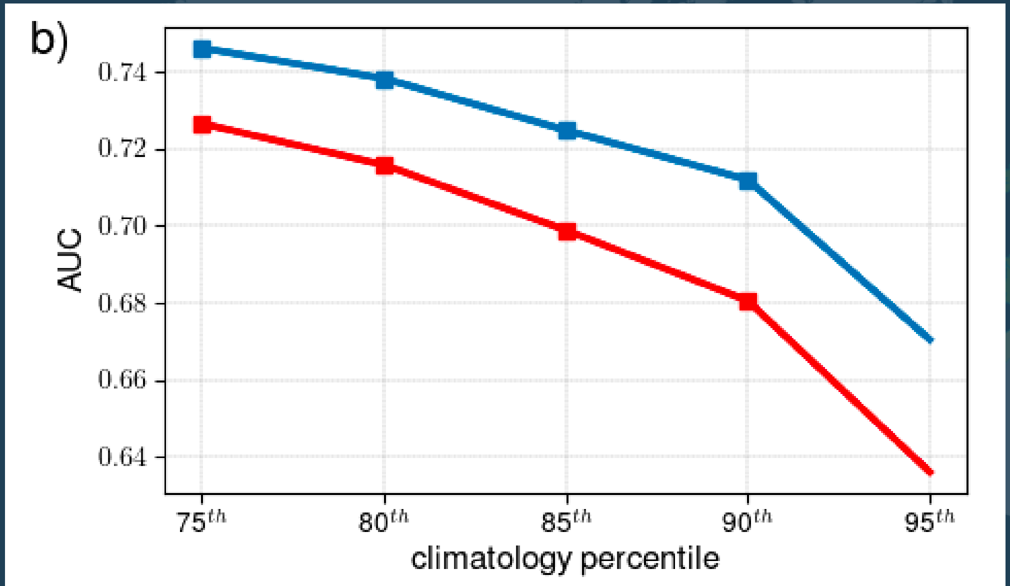


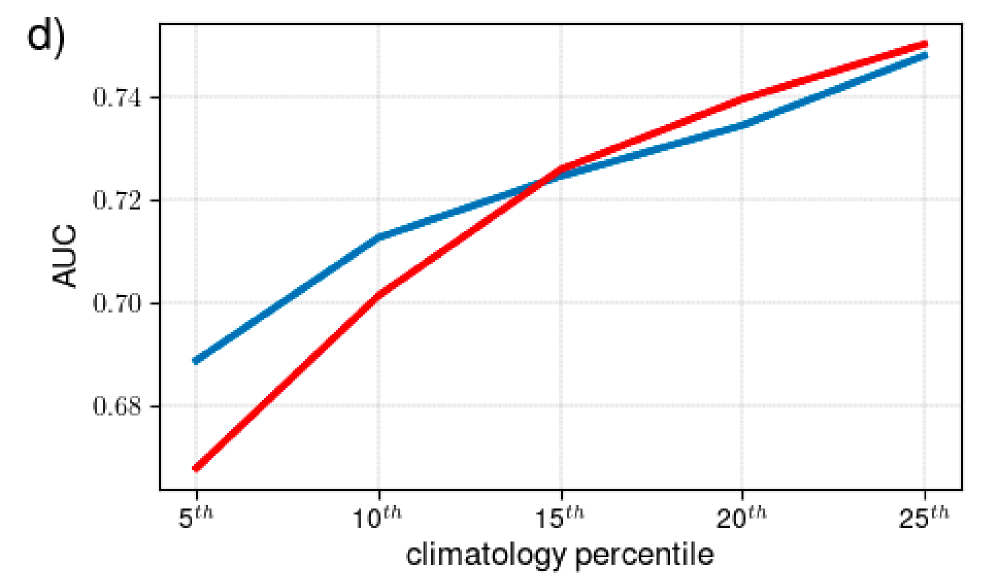
Potential discrimination ability (ROC area) for day 6 forecasts

warm **summer** day

IFS –red
PanguWeather - blue

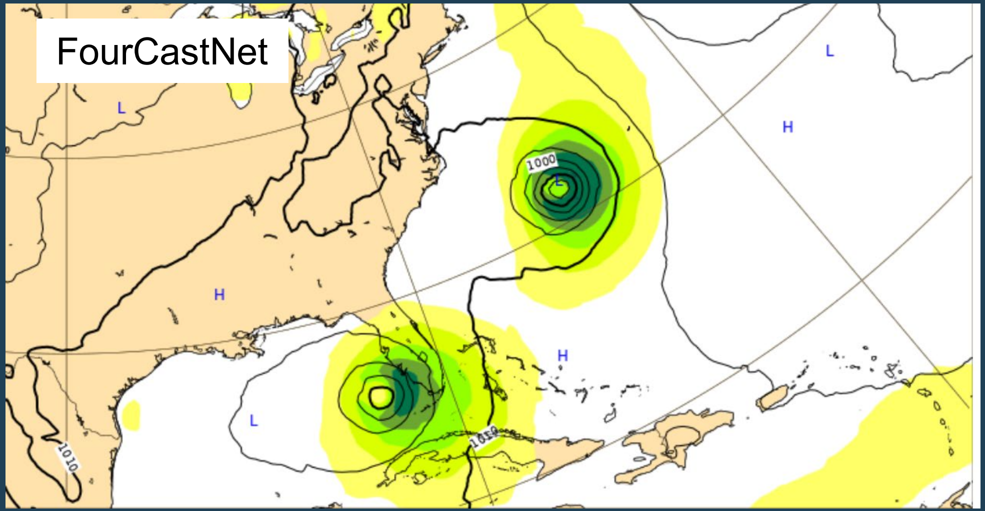
cold **winter** days

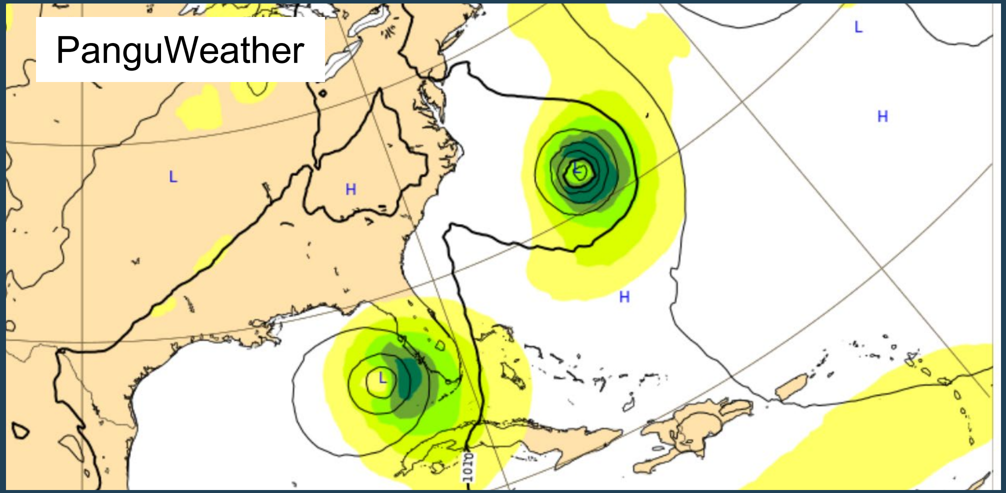


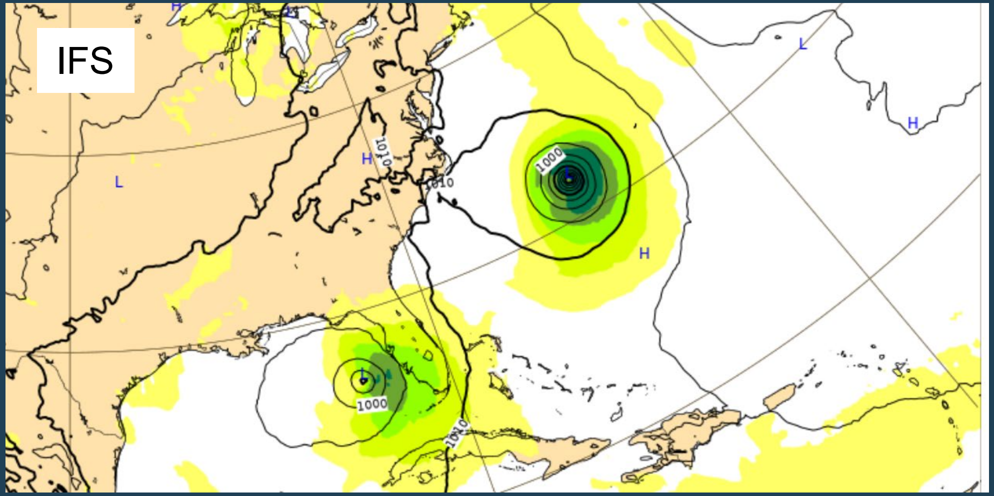


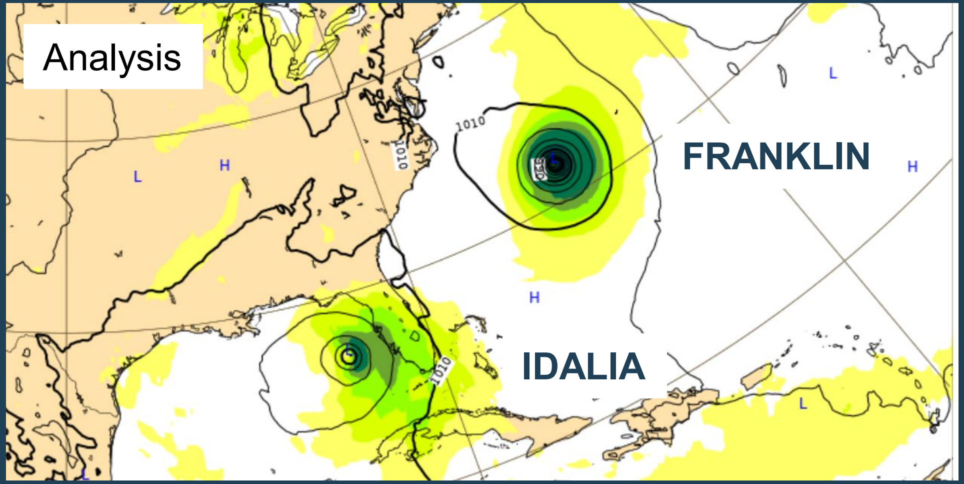
PanguWeather better for both warm and cold extremes, based on climatological threshold from own climate

Tropical cyclones Idalia and Franklin (day 2 forecasts, valid on 30 Aug 2023 00UTC)





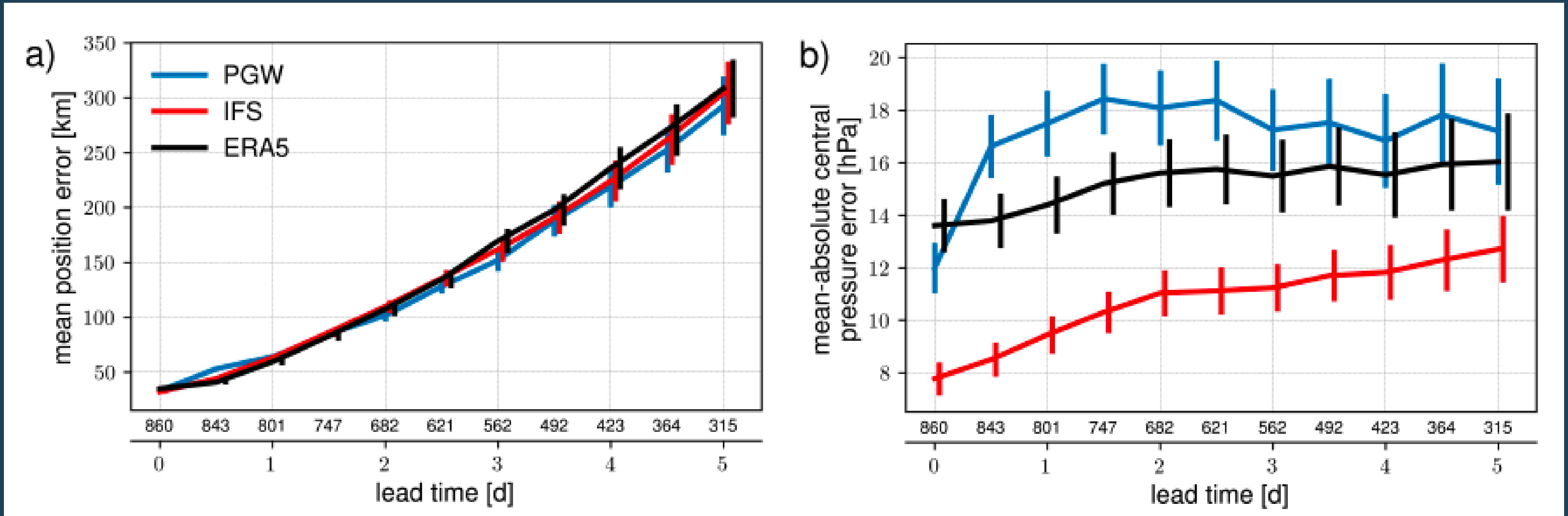




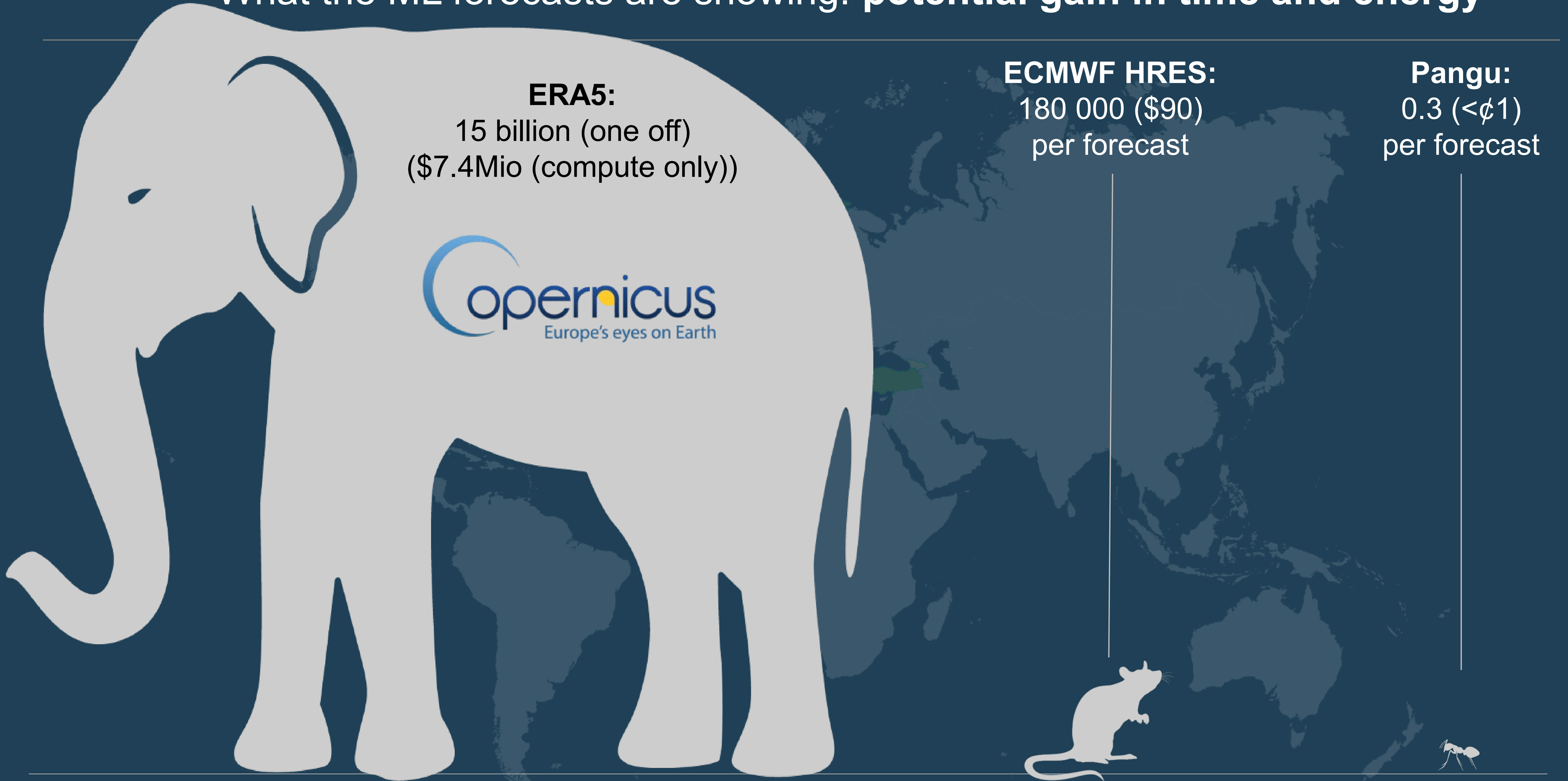
Tropical cyclone verification

Position error

Intensity bias



What the ML forecasts are showing: **potential gain in time and energy**



Summary/Outlook

- Very good scores for PanguWeather initialised from ECMWF analysis
- Temperature extremes, cyclogenesis of both extra-tropical and tropical cyclones can be captured
- Similar perturbation growth rate from initial perturbations on synoptic scales

- Problem with structure of very intense cyclones
- Smooth small scales

- Currently not directly predicting precipitation
- Missing model uncertainty in ensemble mode

What’s next?

Project overview: different paths **towards a ML ensemble prediction at ECMWF**

The hybrid model

Enhanced and accelerated implementation of ECMWF ML Roadmap

Delivering results

Development of a ML ensemble forecast

Data-driven model initialised with NWP analysis hence requiring conventional data assimilation.

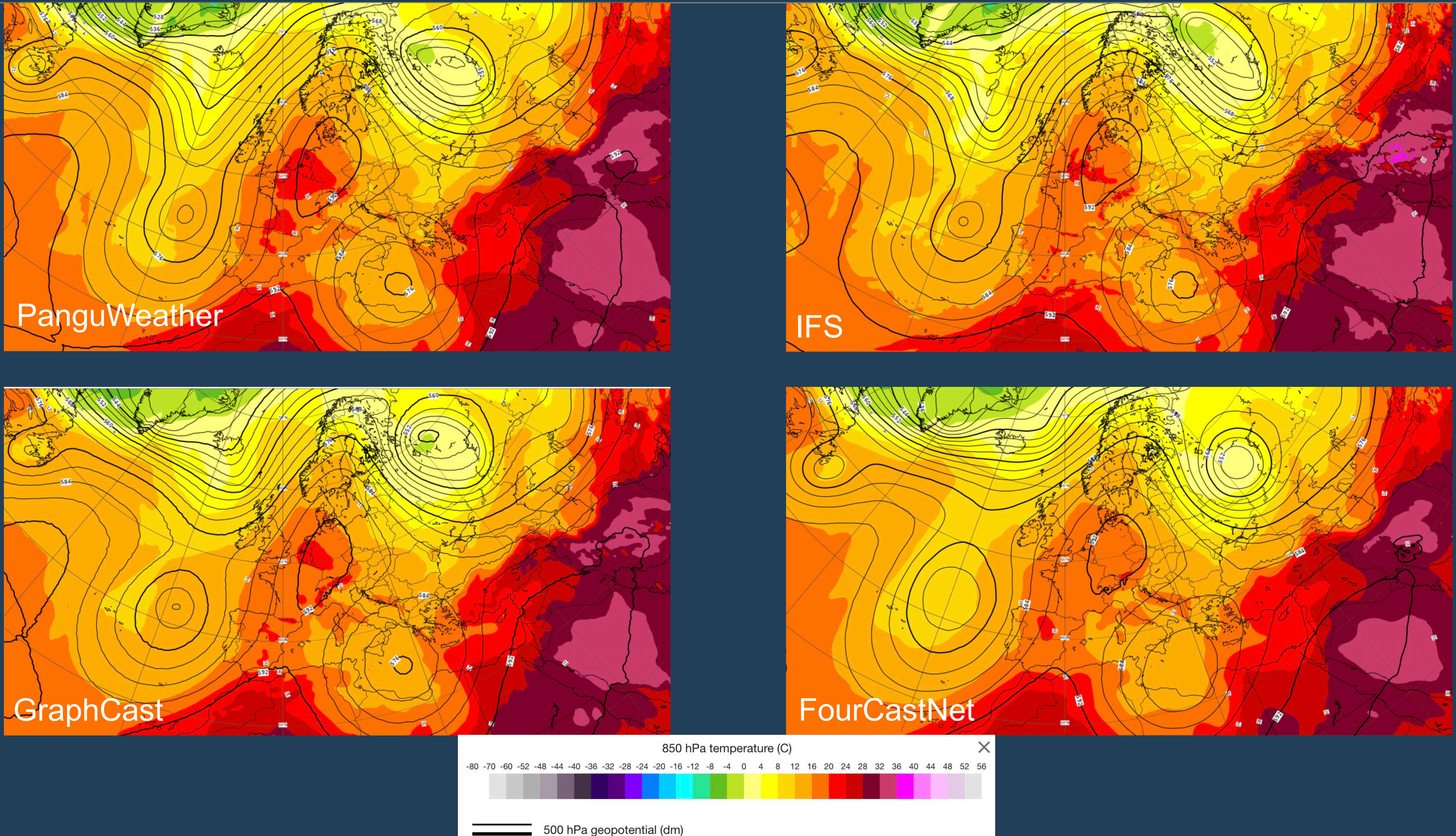
Embracing novelty

Observations-driven ML system

A whole system reinventing the path from observations to predictions.

A scientific challenge

Day 4 forecasts over Europe (valid today, 7 Sept 2023 12UTC)



Results in this talk are mainly from:

arXiv > physics > arXiv:2307.10128

Search...
Help | Advanced S

Physics > Atmospheric and Oceanic Physics

[Submitted on 19 Jul 2023]

The rise of data-driven weather forecasting

Zied Ben-Bouallegue, Mariana C A Clare, Linus Magnusson, Estibaliz Gascon, Michael Maier-Gerber, Martin Janousek, Mark Rodwell, Florian Pinault, Jesper S Dramsch, Simon T K Lang, Baudouin Raoult, Florence Rabier, Matthieu Chevallier, Irina Sandu, Peter Dueben, Matthew Chantry, Florian Pappenberger

ECMWF Newsletter 176 • Summer 2023

news

Exploring machine-learning forecasts of extreme weather

Linus Magnusson

Acknowledgements

We appreciate the researchers at ECMWF for their efforts in collecting, archiving, disseminating, and maintaining the ERA5 reanalysis dataset, HRES, and ensemble, without which this study would not have been feasible.

ACKNOWLEDGMENTS

We would like to thank ECMWF for offering the ERA5 dataset and the TIGGE archive. Without such selfless dedication, this research would never become possible. We

Acknowledgements

We acknowledge the use of the ERA5 dataset on both pressure levels and single level provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). Without their great efforts in collecting, archiving, and disseminating the data, this study would not be possible.

We thank the researchers at ECMWF for their open data sharing and maintaining the ERA5 dataset without which this work would not have been possible.

We also thank ECMWF for providing invaluable datasets to the research community.