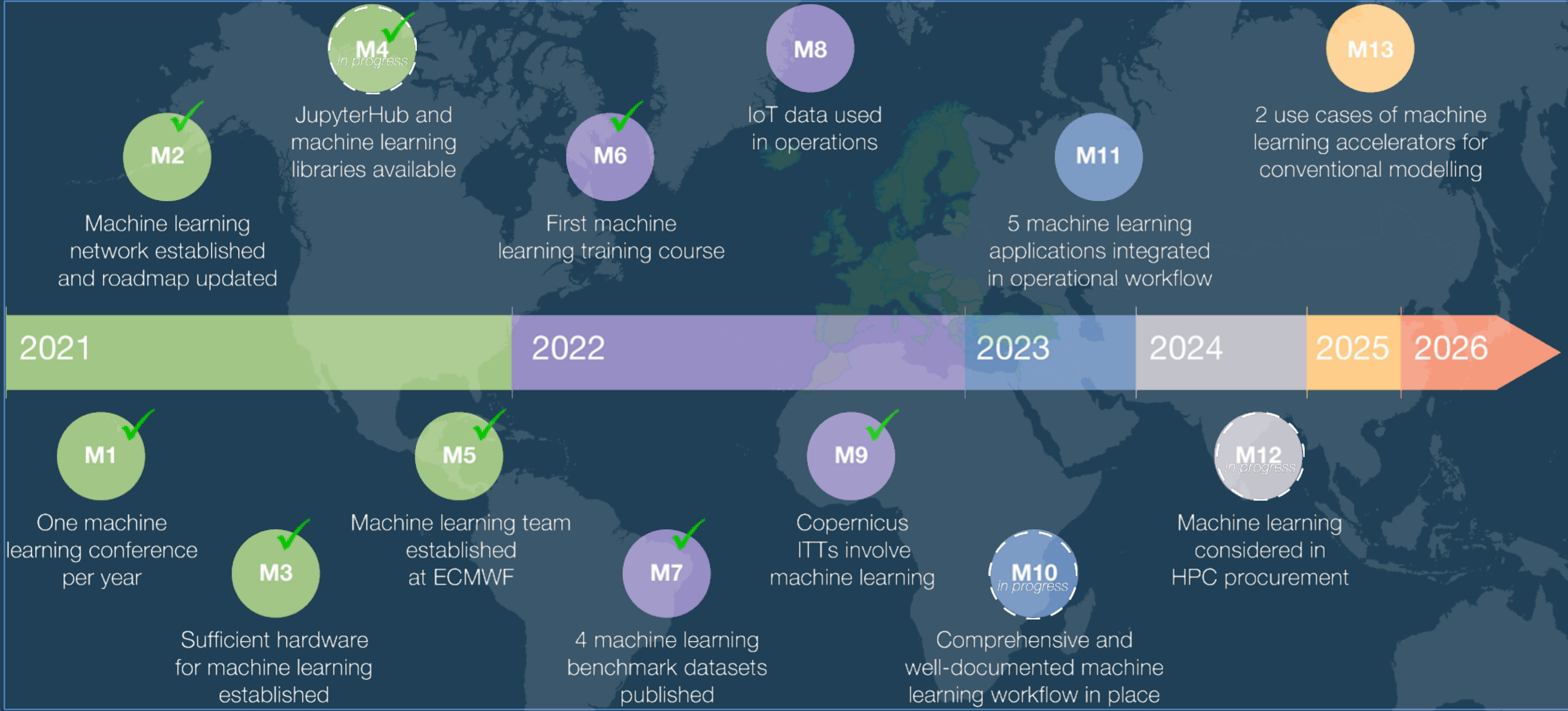
Machine Learning at ECMWF

ECMWF HPC Workshop 2023

Matthew Chantry
Machine Learning Coordinator
matthew.chantry@ecmwf.int

2021 ECMWF ML Roadmap

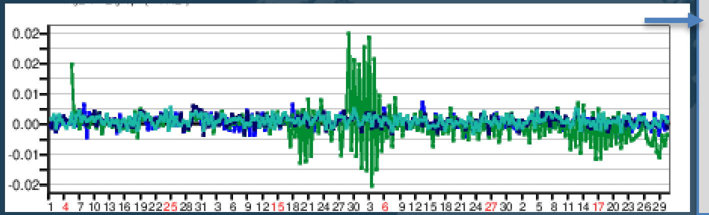


Vision 2031

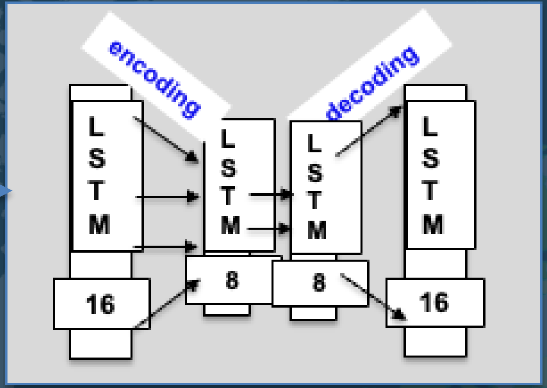
- It is difficult to distinguish between machine learning and domain sciences
- Data handling fully capable to serve machine learning needs
- Fully supported diagnostic tools via trustworthy AI
- Physical constraints can be represented in deep learning
- Use of machine learning as easy and normal as data re-gridding
- Unsupervised learning and causal discovery used on a regular basis
- Machine learning solutions from end-users integrated in workflow

Machine learning **automated anomaly attribution**

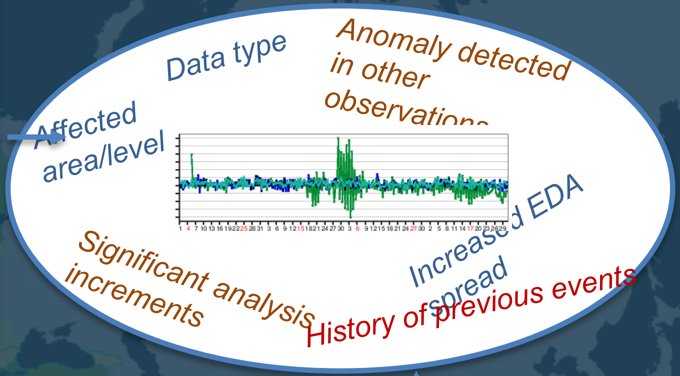
Observation anomaly



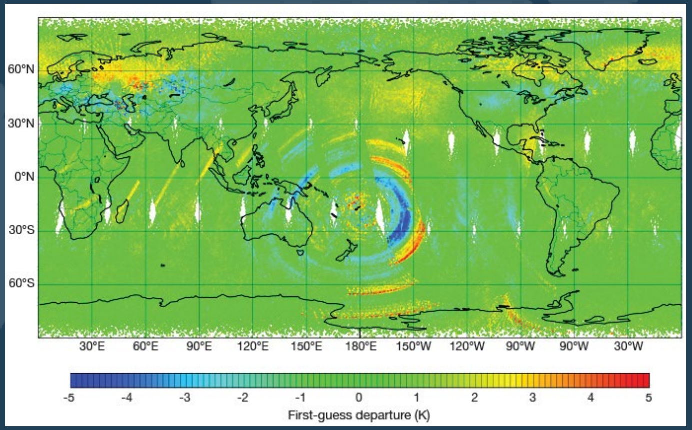
LSTM autoencoder
Detect Anomalies



Random forest classifier
Improves classification



Hunga-Tonga eruption



M. Dahoui

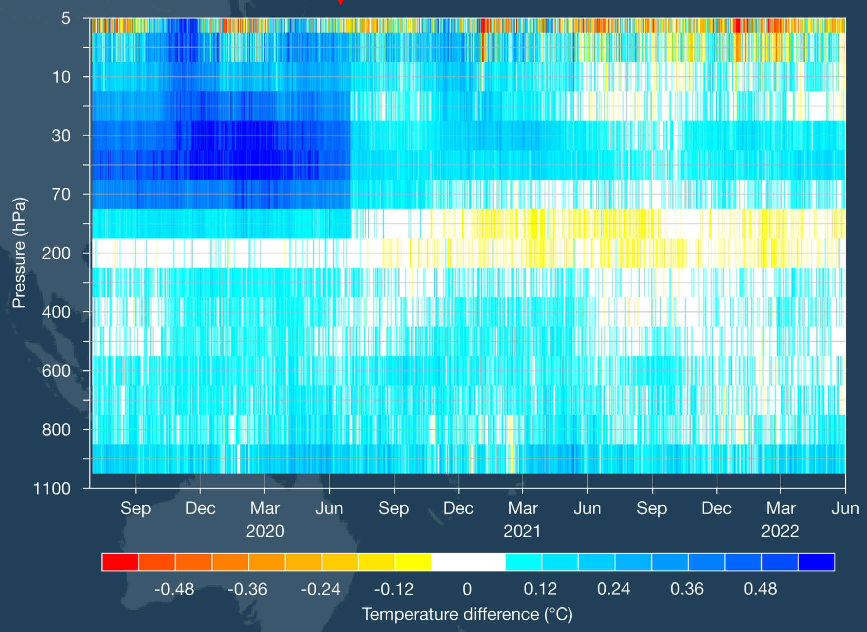
Hybrid NWP+ML – collaboration with CEREA, France

- Hybrid models augment standard physics-based models with a data-driven component:

$$x_{k+1} = M^{phys}(x_k) + F^{stat}(x_k, p)$$

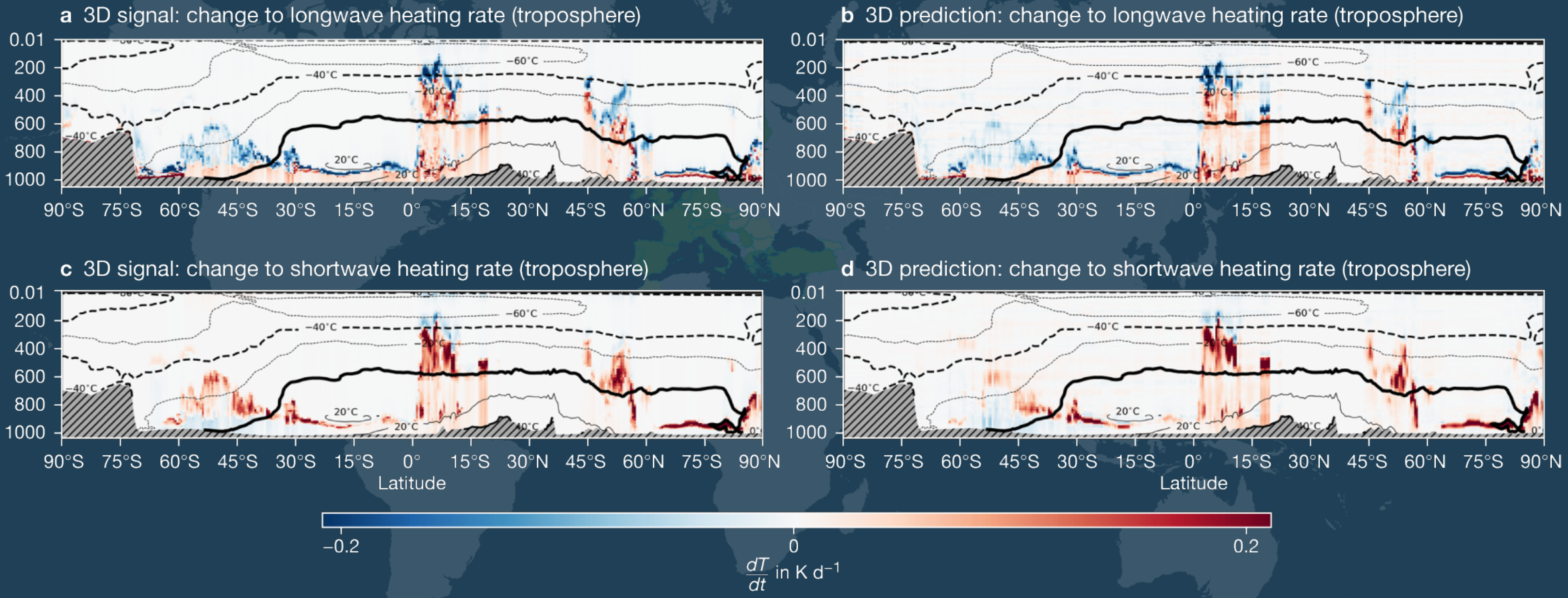
- A hybrid model is already used in the ECMWF weak constraint 4DVar analysis. Can the hybrid model approach be extended to the forecast?
- Bonavita & Laloyaux, 2020 trained offline a neural network (NN) to learn model errors, showing improved forecast skill scores in the full IFS
- Farchi et al., 2022 developed this idea introducing online training of the NN inside 4DVar: this outperforms results from offline training in simplified models.
- Current work, testing in the full IFS: results appear promising!

Introduction of bias correction within 4D-var improved stratospheric representation.
Next step NN trained within 4D-var…



Alban Farchi & Marc Bocquet
@ CEREA
Bonavita, Chrust, Laloyaux
@ ECMWF

Using neural network emulators



Infero library - A lower-level API for ML Inference in Operations

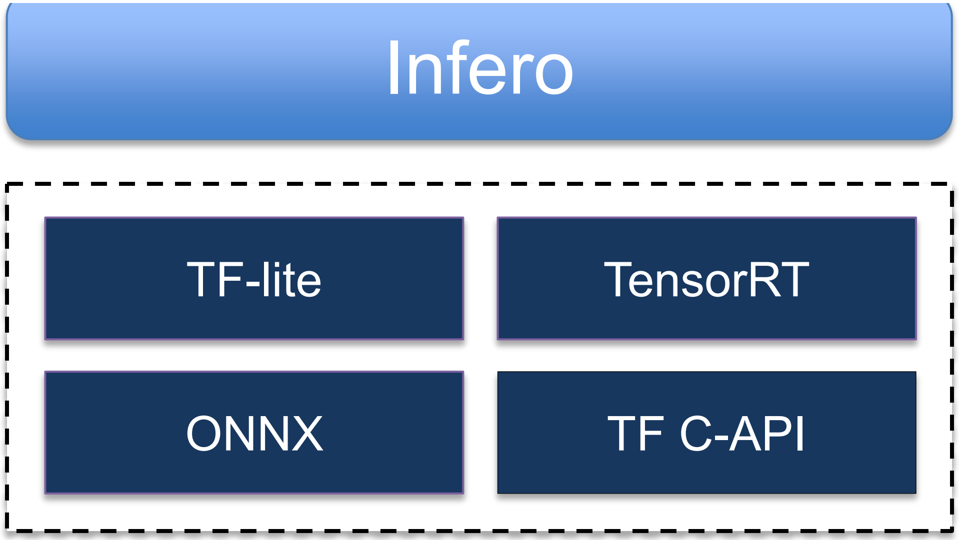
- One Interface, multiple backends
  - TF-lite
  - TensorRT
  - ONNX
  - TF C-API
- Infero provides API's:
  - C, C++, Fortran, Python
- Supports C and Fortran tensor
- Open-Source:
  - github.com/ecmwf-projects/infero

Fortran

```
model%initialise_from_yaml_file(yaml_path)
model%infer(input_tensor, output_tensor)
```

Python

```
model = pyinfero.Infero(model_path, model_type)
output = model.infer(input_tensor, output_shape)
```



More hybrid NWP+ML avenues

Learning observation operators to incorporate new observations.

Estimating fuel for fire prediction system.

Improving S2S forecasts.

Estimating sea-ice model parameters.

Bayesian optimisation for optimising uncertain model parameters.

ECMWF's ML Strategy: with **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.

Jan 2021
Machine Learning Roadmap

ECMWF

Strategy to embed Machine Learning deeply into the ECMWF operational chain

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

Oct 2022
1km^2 global

Jua.ai

1x1km global
48 hours lead time

5 minute timesteps

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

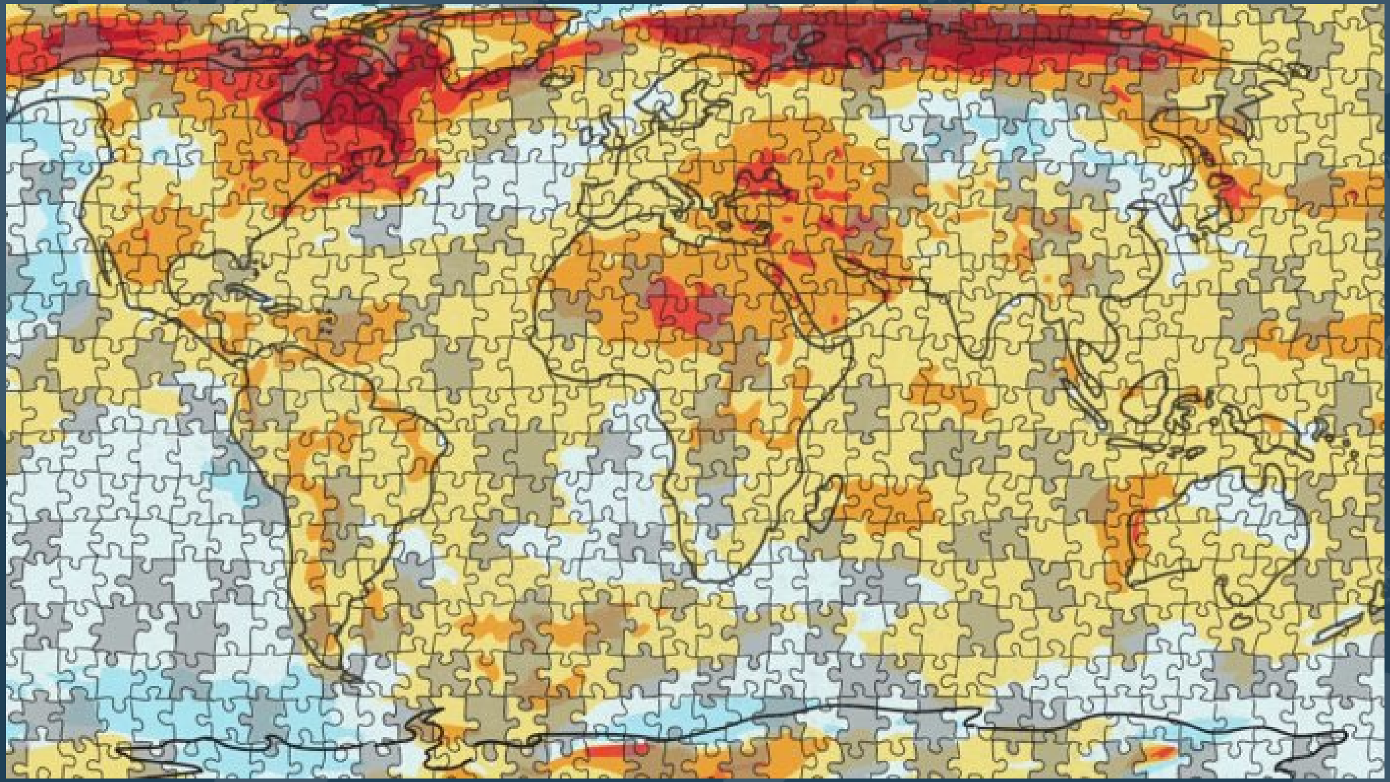
Sharp spatial features

The rise of data-driven forecasting models

- Modern machine learning architectures.
- Training on reanalysis (ERA5 ~40years)
- Given a state of the atmosphere q, u, v, t, z (~13 pressure levels).
- Predict the state of the atmosphere in 6 hours time.
- Iterate this forward to make 10-day forecasts.

Why is Reanalysis used to train ML models?

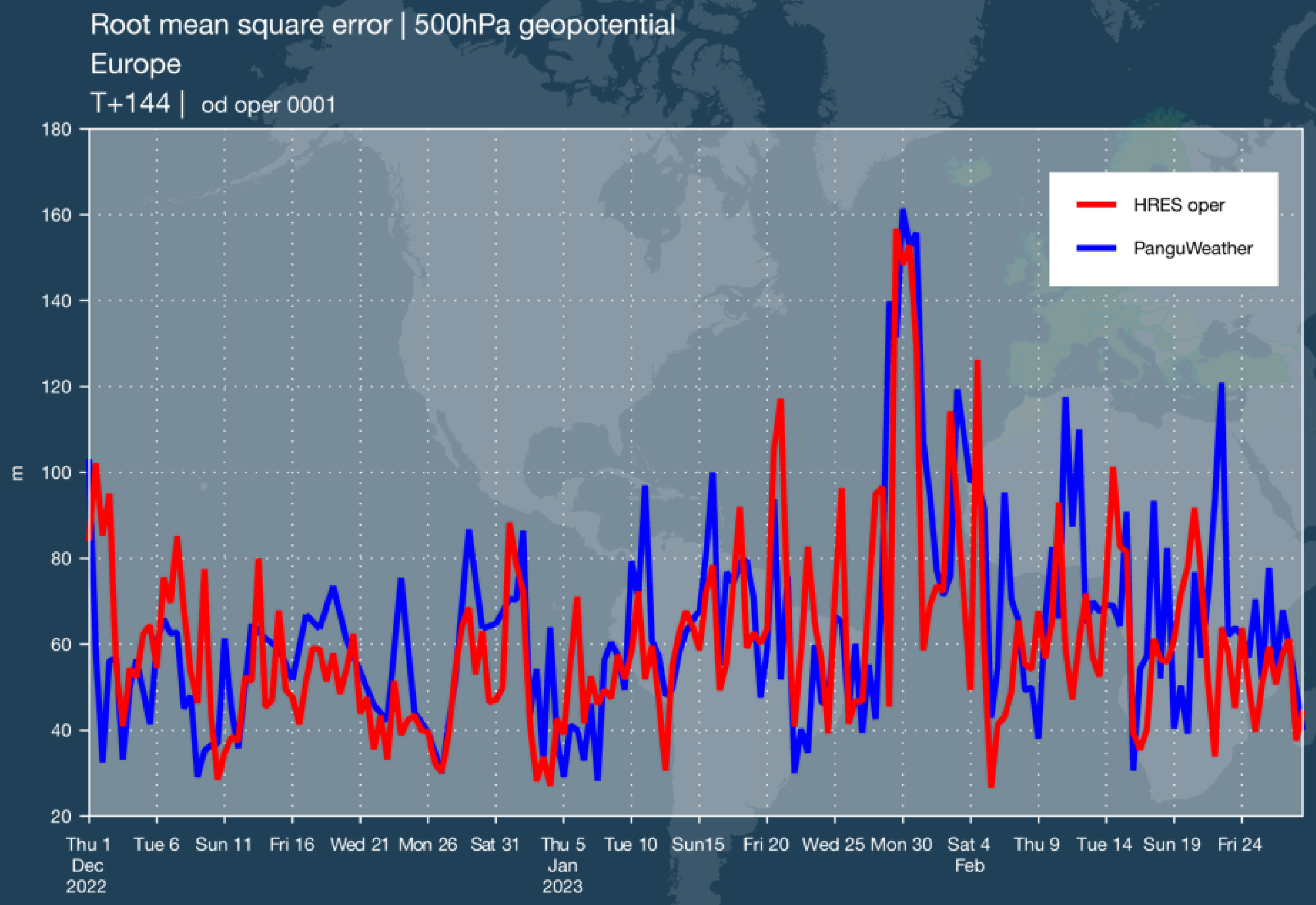
Reanalysis combines observations with cutting-edge weather models, to provide maps without gaps.



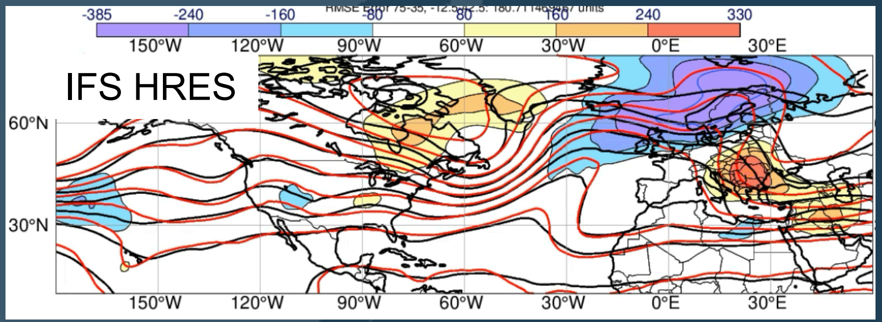
ERA6 coming out in 2026...

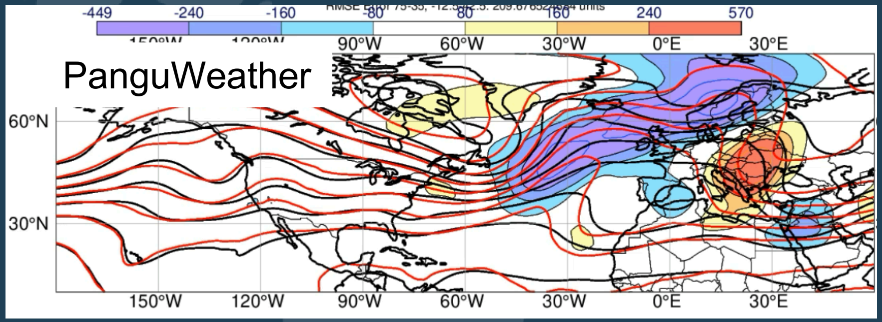
Time-series of day 6, RMSE over Europe

Same starting point....similar results

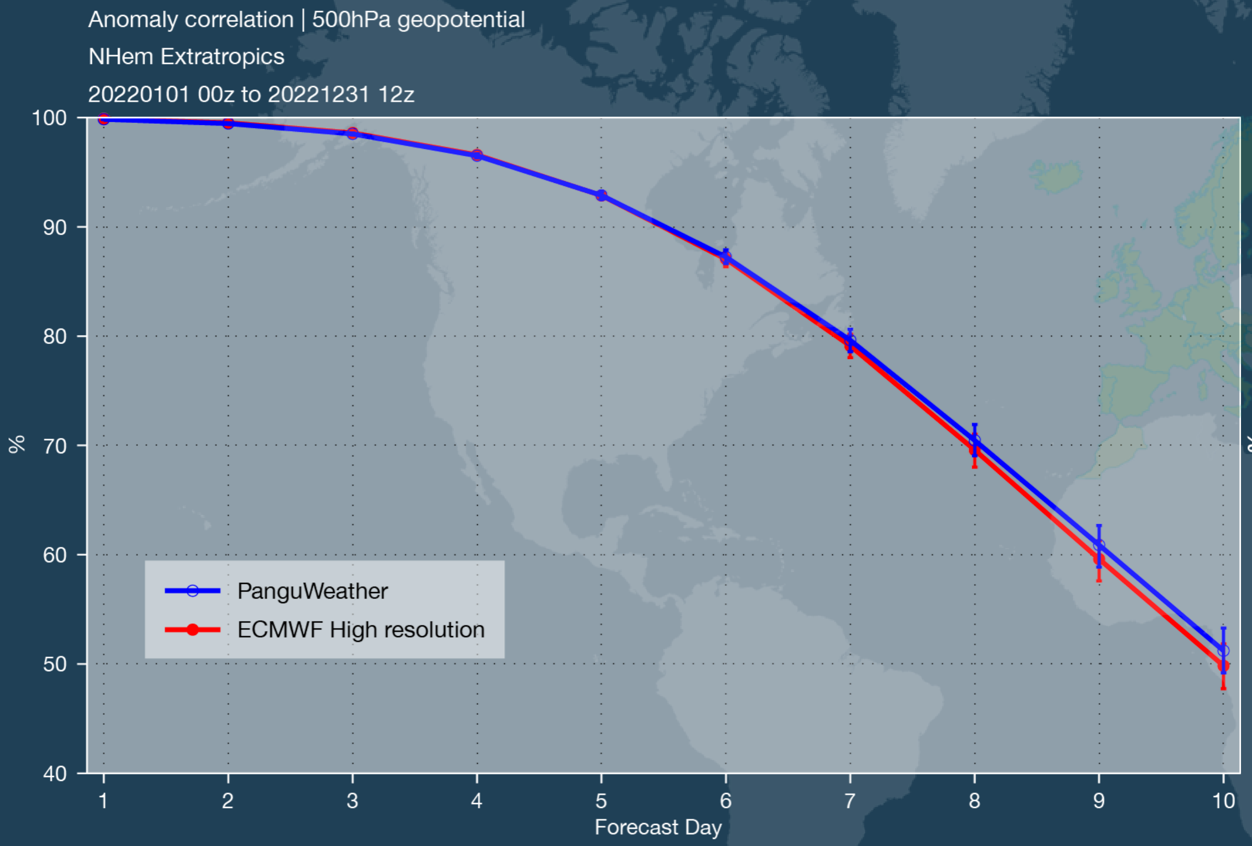


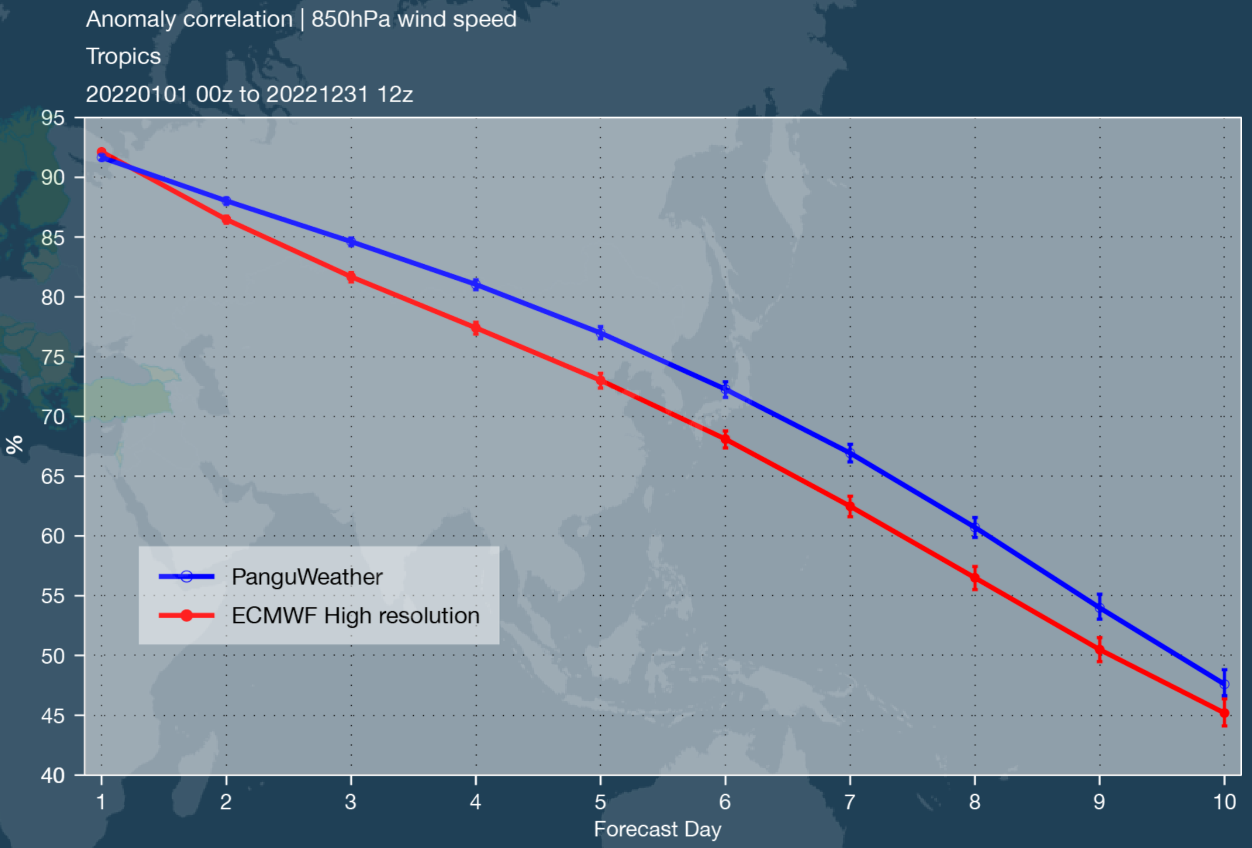
+144h forecast errors 30 January 00UTC



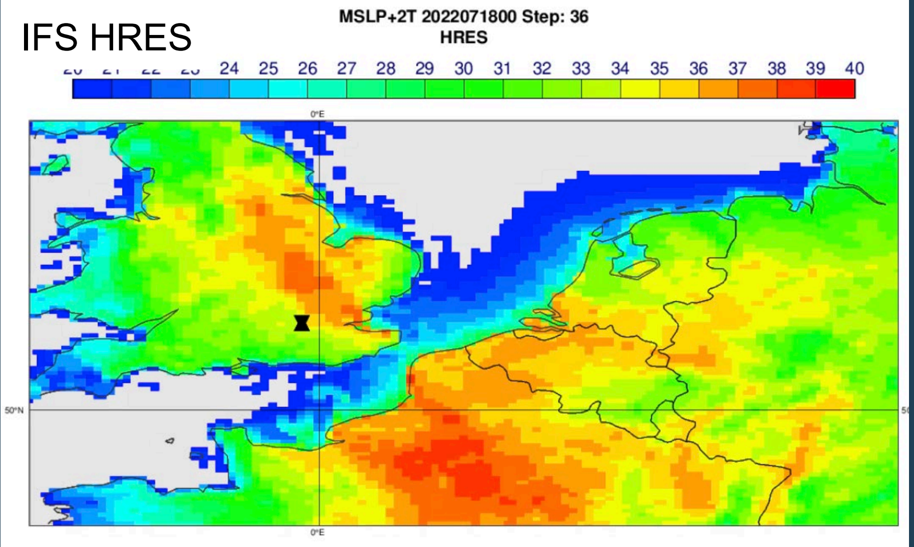


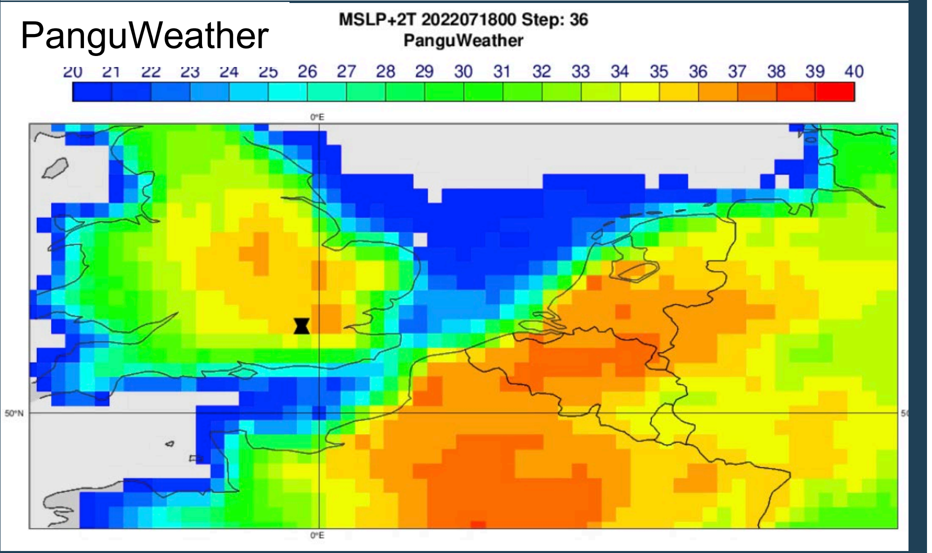
What the analysis is showing: **an undeniable skill**



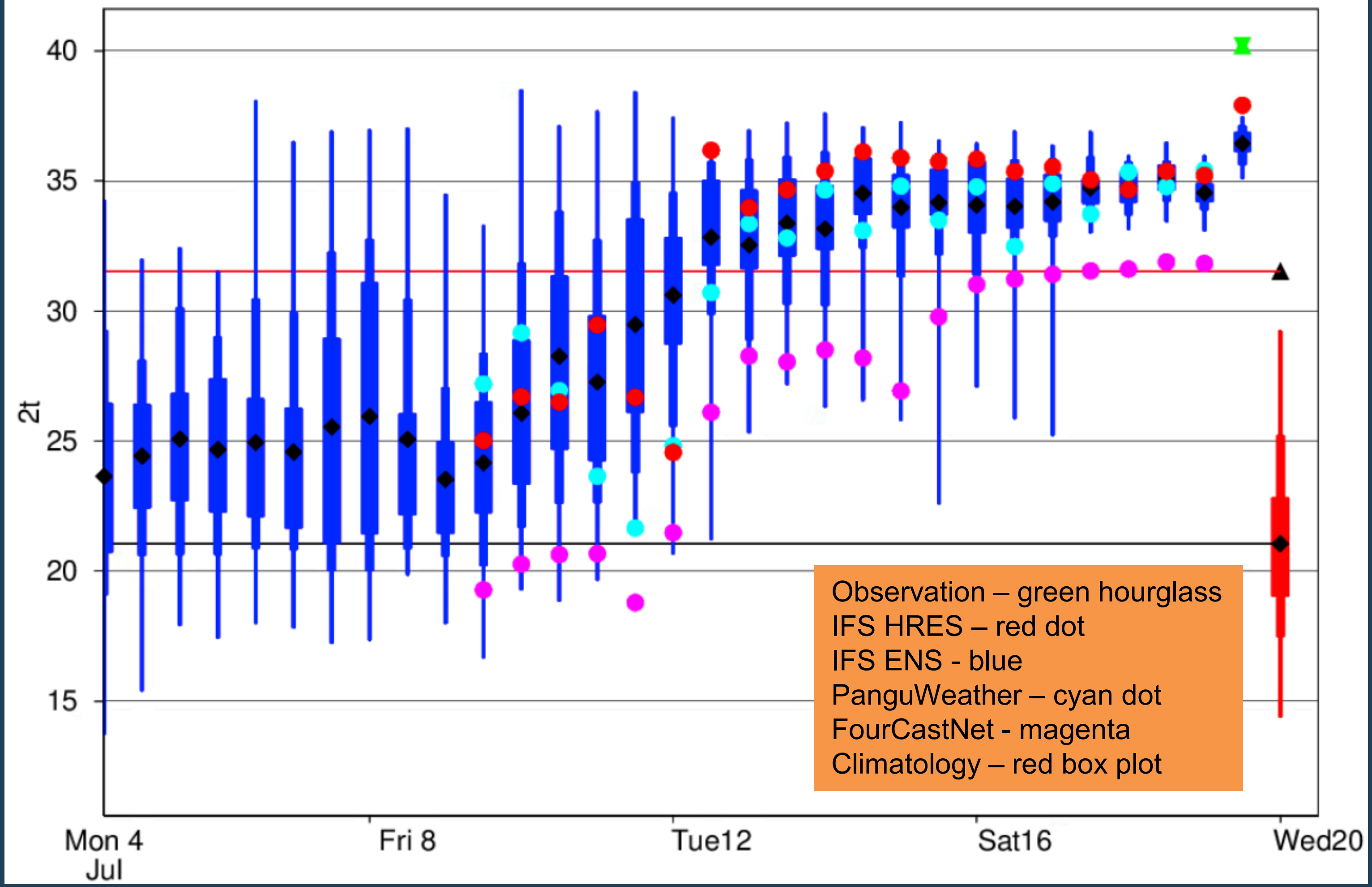


UK heatwave 2022

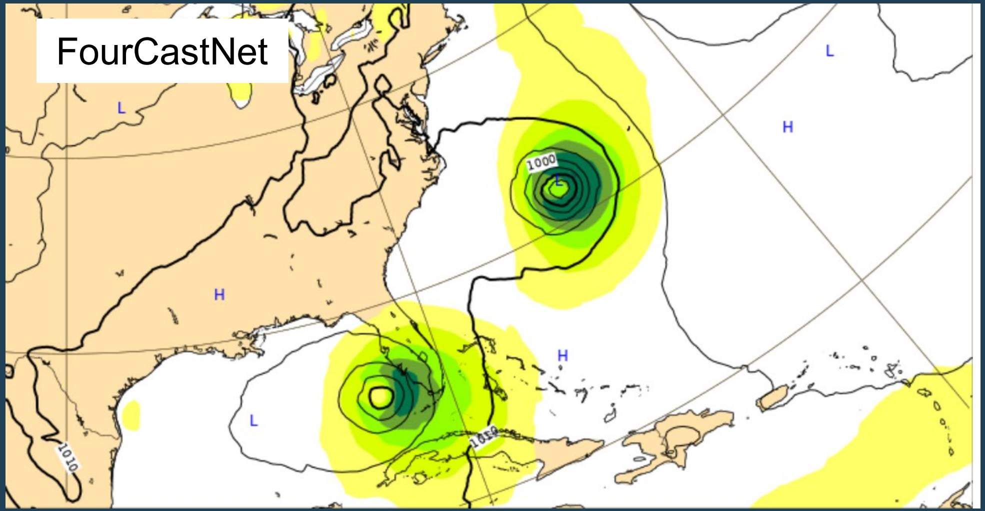


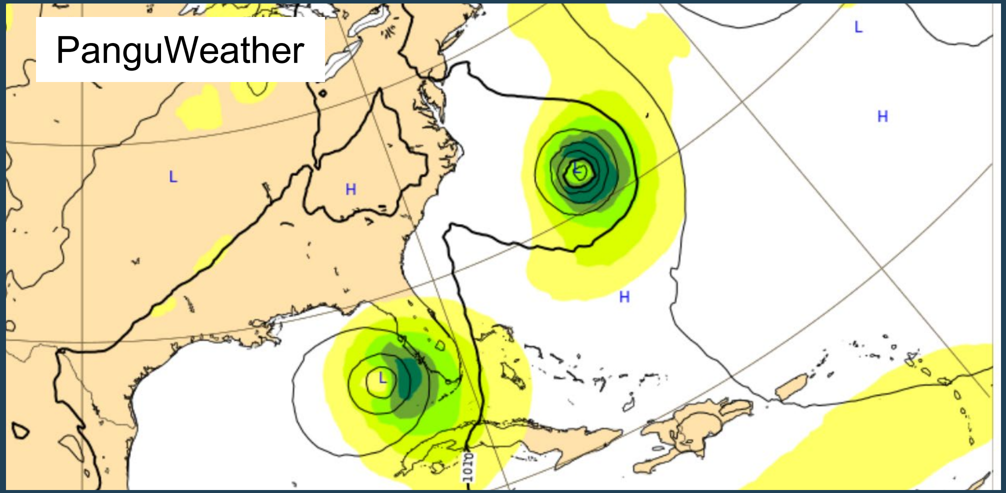


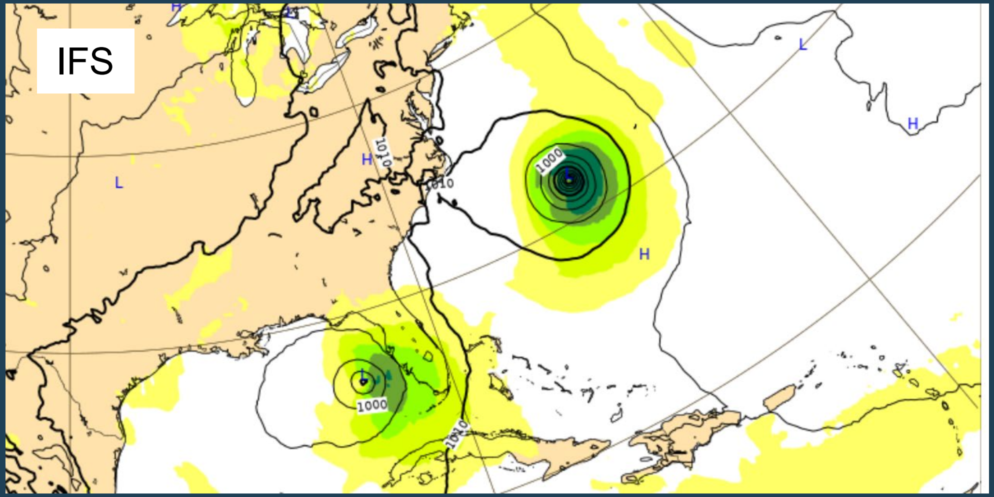
2m temperature Heathrow 19 July 12UTC

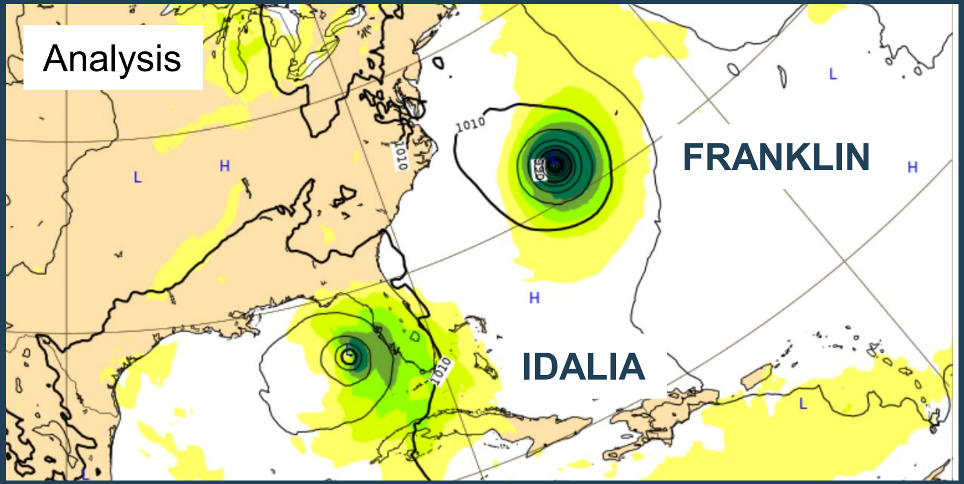


Tropical cyclones Idalia and Franklin (day 2 forecasts, valid on 30 Aug 2023 00UTC)









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

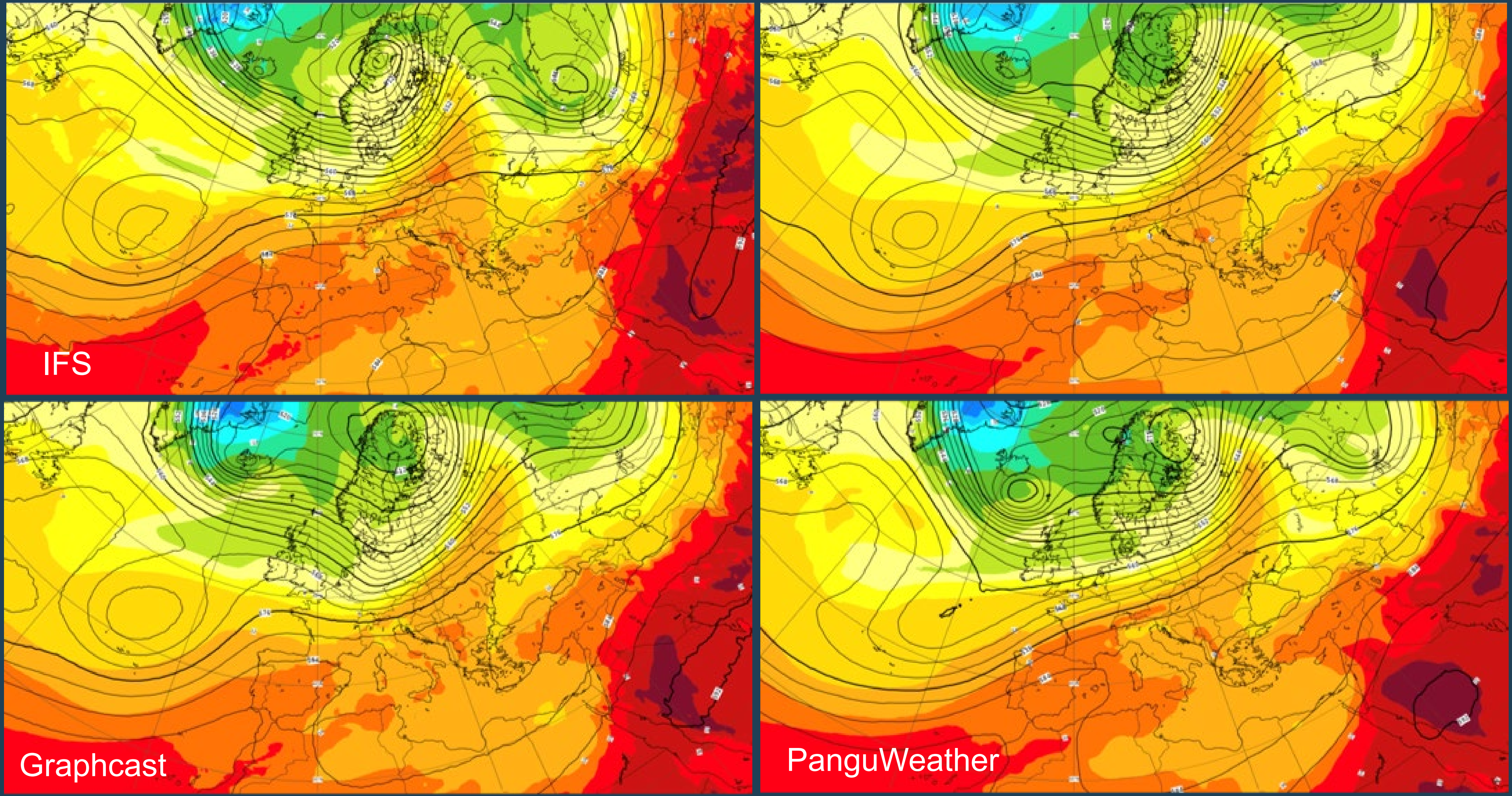
Observations-driven ML system

A whole system reinventing the path from observations to predictions.

A scientific challenge

stay tuned…

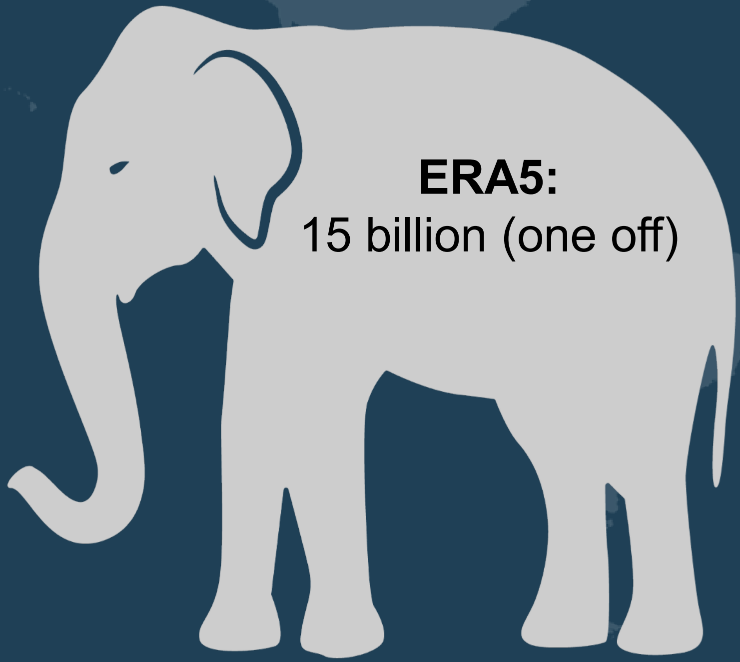
Thursday’s weather (T+156h)



Dichotomy between training & inference costs

Training:
O(32) GPUs, O(weeks)
NVIDIA have shown scaling out to 2k GPUs

Inference:
1 GPU, O(minute)
compare with 50nodes, O(hour)



ECMWF HRES:
180 000
per forecast

Pangu:
0.3
per forecast

Future HPC systems?

Current models, 25km, 6-hourly output
destination IFS 9km, 1-hourly output?

~3x smaller spacing

Scaling of ML models depends on spatial architecture.

FourCastNet, Spectral transforms, $O(n^2 \log n^3)$
Pangu/FuXi/FenWu, Vision transformers, $O(n^2\text{-}n^4)$ (algorithm dependent)
GraphCast, GNN, $O(n^2)$

Estimated 9km requirements
100-1000 GPUs, O(weeks)

Summary

- Hybrid NWP+ML offers avenues for improvement & acceleration.
  - HPC challenges around coupling.
- Data-driven forecasting models offer new paradigm.
  - Comparable accuracy with IFS
  - Outstanding challenges around ensembles/TC intensity/precipitation.
  - Significant acceleration of predictions.
  - Significant GPU resources required for training.
- Future systems will need consider both NWP & ML.
  - What role for exotic hardware?