AIFS

ECMWF’s data driven forecast model

Simon Lang

Simon.lang@ecmwf.int

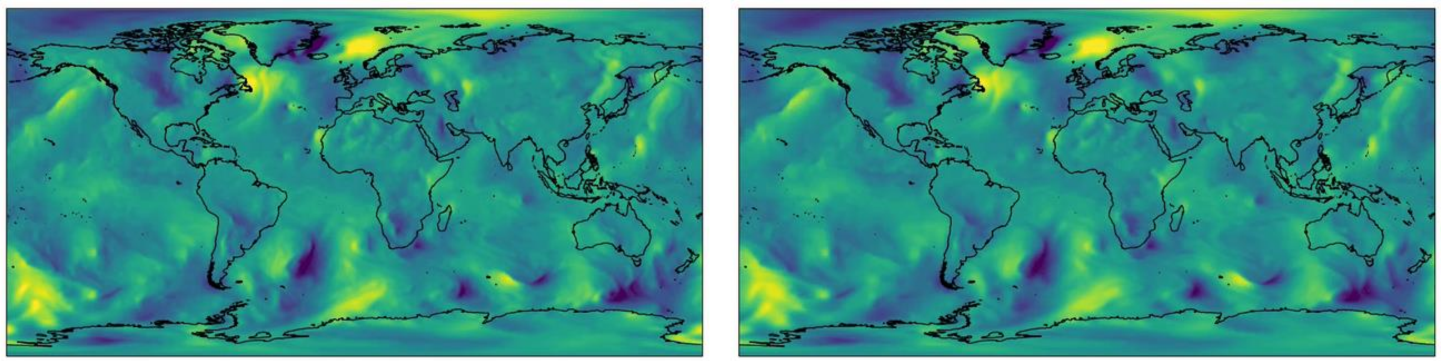
Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Zied Ben Bouallegue, Linus Magnusson, Mariana Clare, Peter Lean, Christian Lessig

AIFS

-> following Keisler 2022 and Lam et. al 2022

- GNN architecture: Interaction Networks (Battaglia et. al 2016)

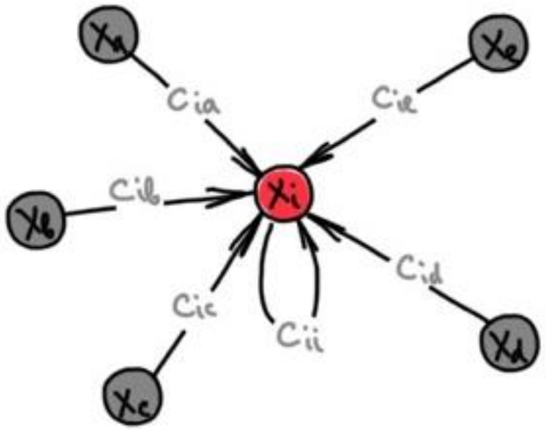
- Graph representation, hidden multi-scale mesh, edge features



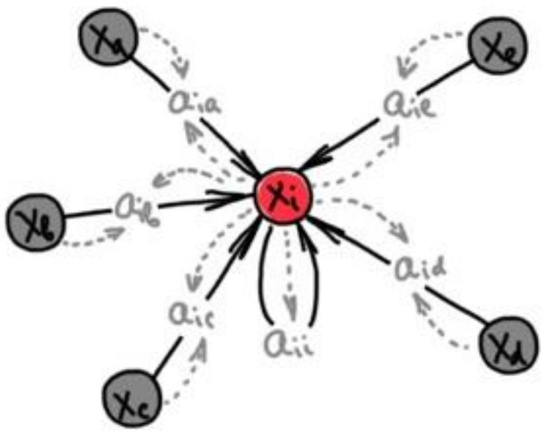
Why GNN : can handle arbitrary input / output grids, local and ad hoc grid refinement, changing grids etc. ; attractive for use in earth system science

AIFS

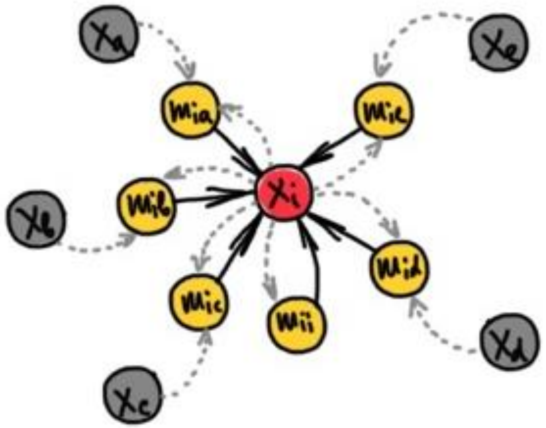
expressive power



convolutional



attentional



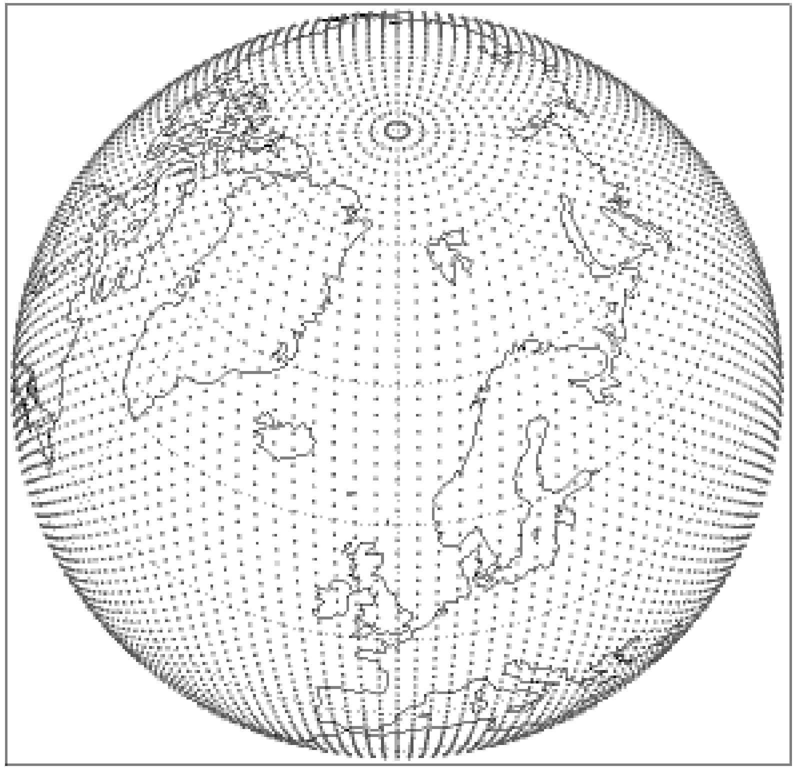
message-passing

$$\mathbf{h}_u = \phi\left(\mathbf{x}_u, \bigoplus_{v \in \mathcal{N}_u} \psi(\mathbf{x}_u, \mathbf{x}_v)\right)$$

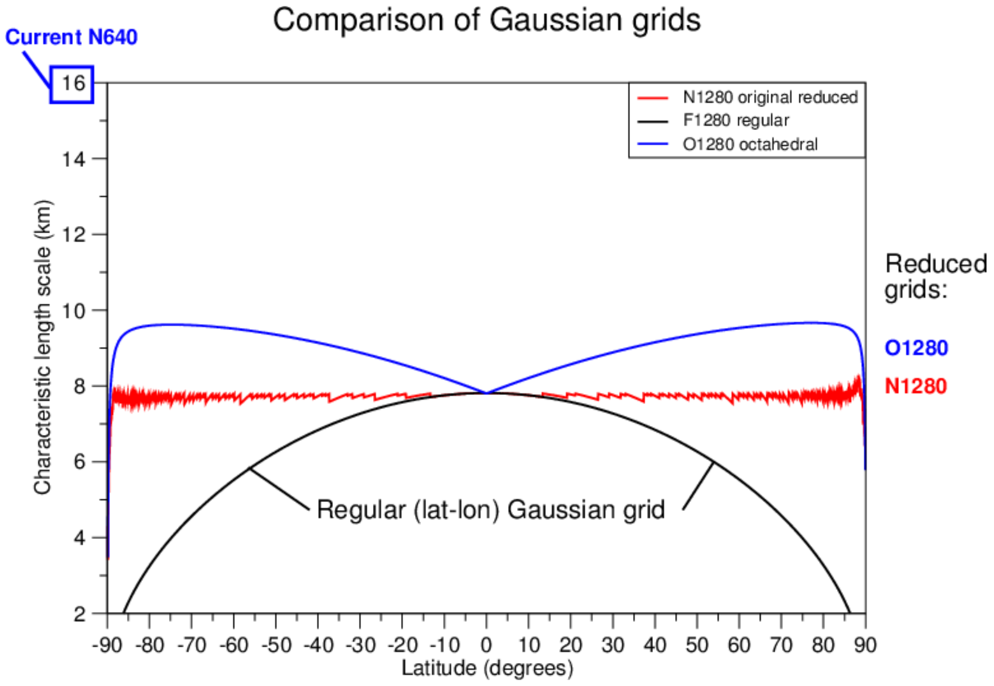
Veličković, 2023 - https://arxiv.org/abs/2301.08210

https://thegradient.pub/graph-neural-networks-beyond-message-passing-and-weisfeiler-lehman

AIFS grid



o96 / o160 / n320

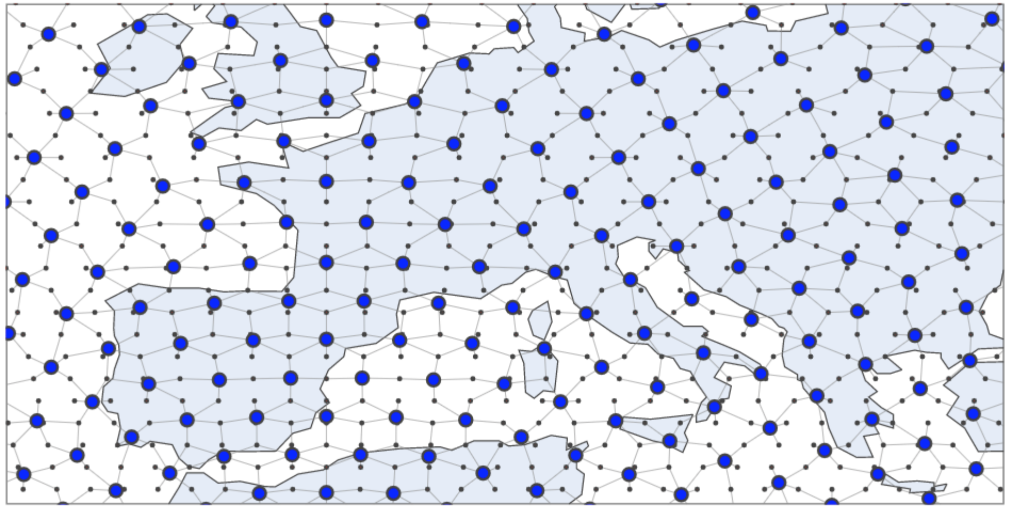


+ (equi-)area weights
+ weighting along plevs (vertical)
+ per-variable weights in the loss

o96 ~ 1 deg
o160 ~ 0.5 deg
n320 ~ 0.25 deg

https://confluence.ecmwf.int/display/FCST/Introducing+the+octahedral+reduced+Gaussian+grid

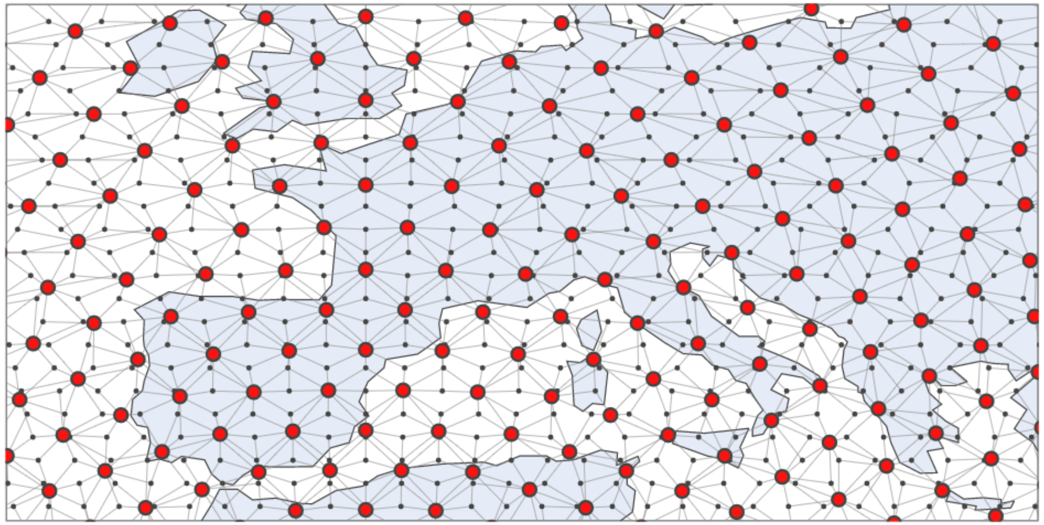
1 x Encoder



Era5

Skip-connection
(residual)

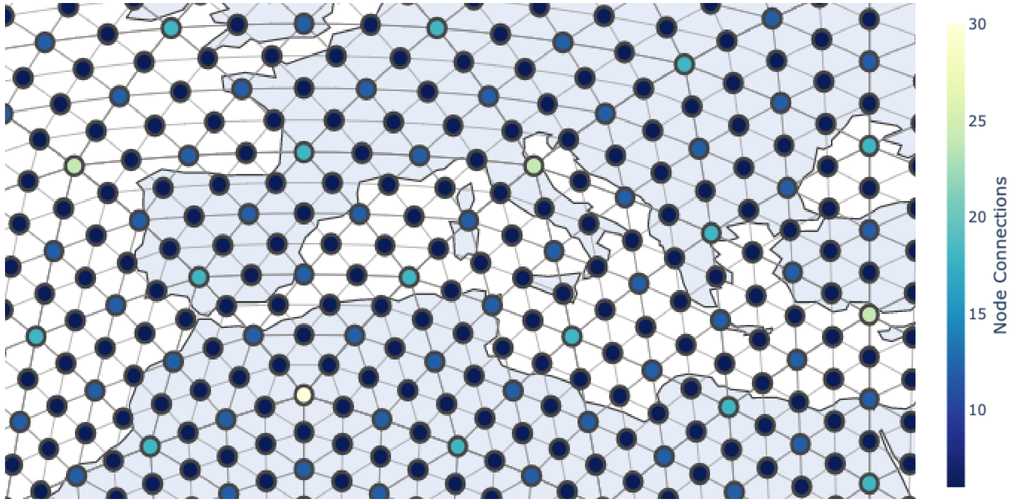
1 x Decoder



Era5

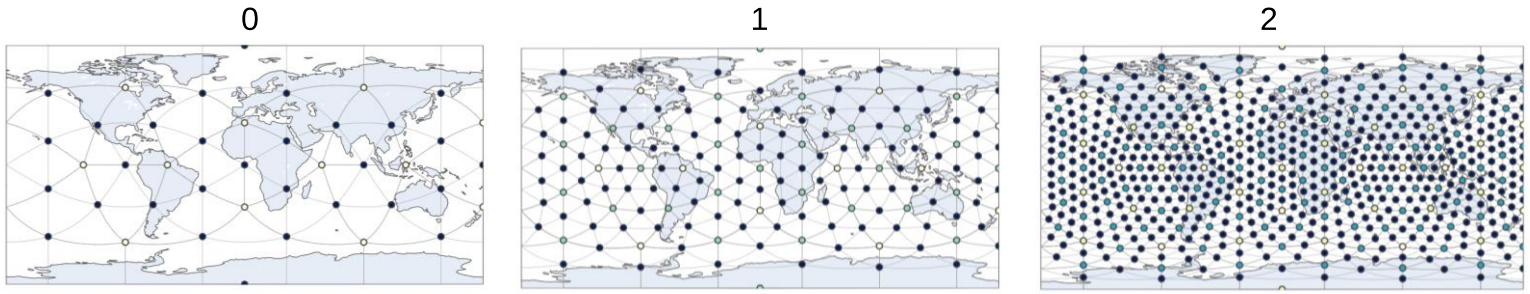
16 x Processor

with skip-connections

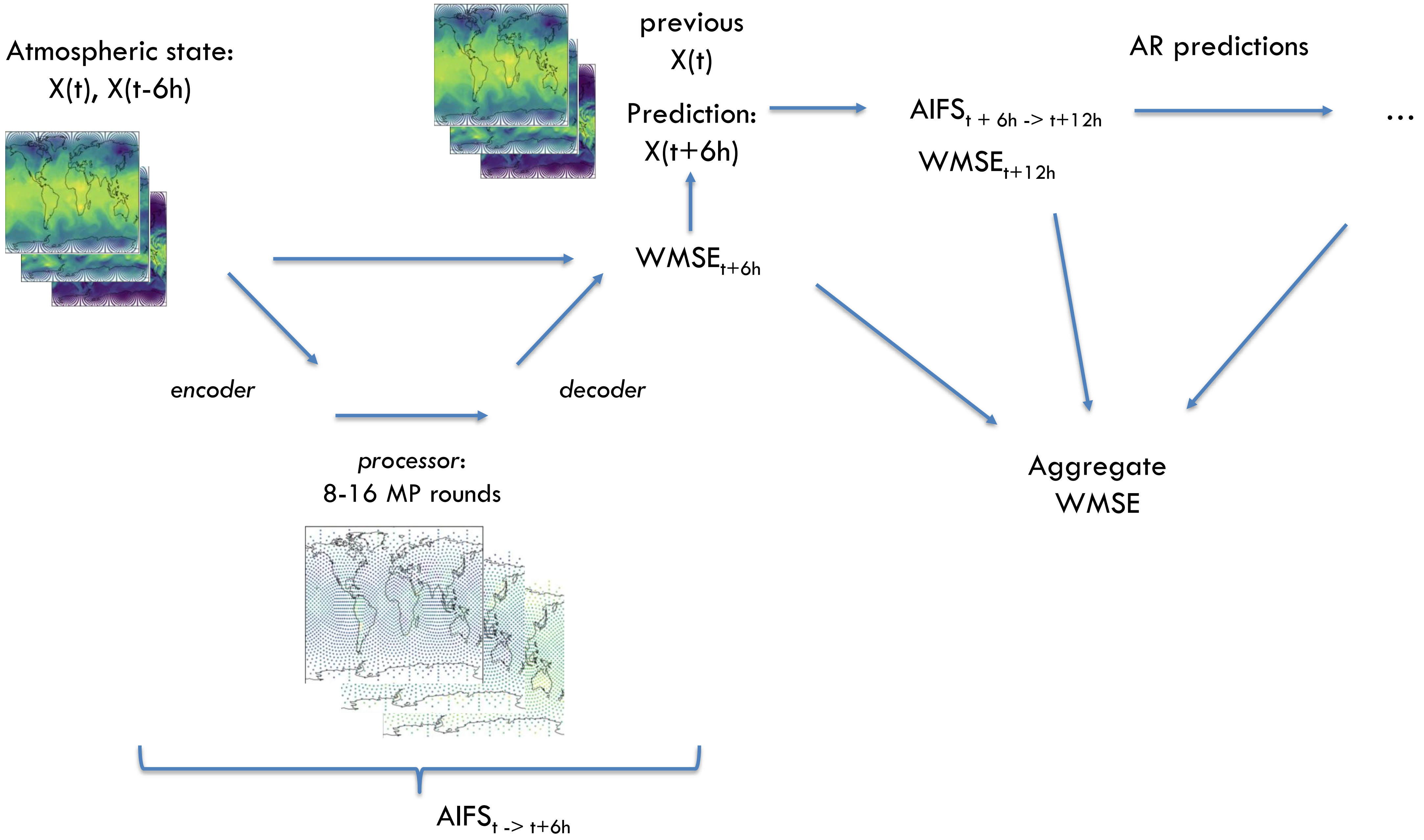


e.g.
Hidden mesh "l6" ~ 40 000 Nodes
Multi-scale interconnectivity
~ 320 000 edges

Simultaneous multi-level message passing



...



Atmospheric state:
X(t), X(t-6h)
previous
X(t)
Prediction:
X(t+6h)
AR predictions
$\mathrm{AIFS}_{t+6h \to t+12h}$
WMSE_{t+12h}
…
WMSE_{t+6h}
encoder
decoder
processor:
8-16 MP rounds
Aggregate
WMSE
$\mathrm{AIFS}_{t \to t+6h}$

AIFS current reference model

Model:

- O96 ERA5 grid, ~1-degree
- “Level 5” hidden grid, ~2-degree

Variables:

13 pressure levels – u, v, w, q, t, z
surface: 2t, 10u, 10v, 2d, sp, msl, sst

Training:

Step 1: 4 days on 16 GPUs to minimise errors for single 6h step
Step 2: 34 hours on 16 GPUs to minimise errors up to 3 days
Step 3: 4 hours on 16 GPUs minimising errors up to 3 days on operational analysis

Total ~6 days on 16 GPUs

Scaling up AIFS

*transformer = fully connected GNN

- Modest number of parameters, model size currently not much of an issue
- Data size is large, lots of grid points -> lots of nodes and edges (similar to very large sequence length in transformer based* models)

GPU memory is limited => 40 GB on ATOS, 64 GB on LUMI / Leonardo

For large (parameters) models there exist quasi out of the box libraries, this is not the case for big input data...

Memory saving options:

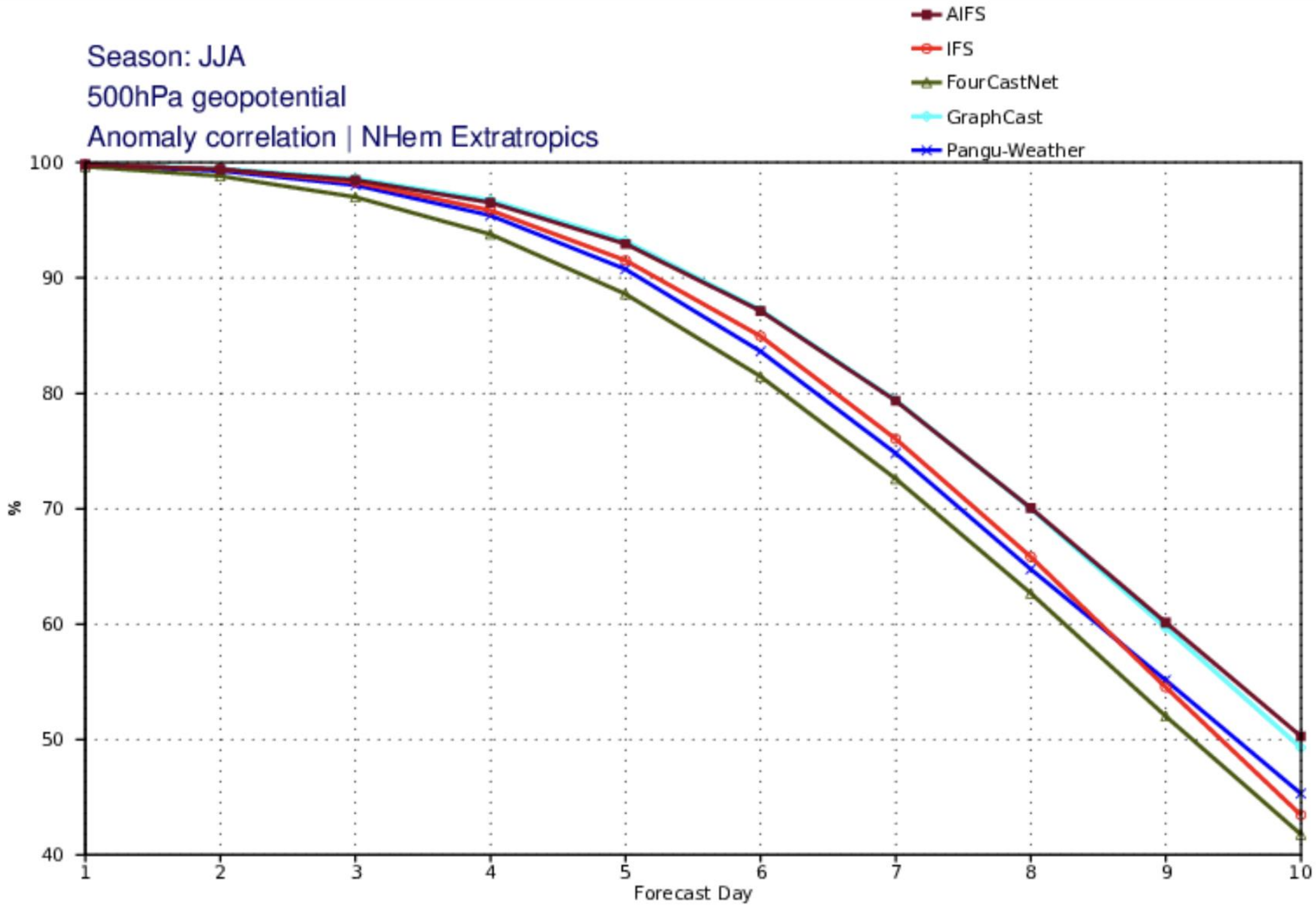
Do not keep everything in memory -> use re-computation in backward : large memory savings. We trade compute for memory as much as possible.

Multi GPU model : shard one model instance across multiple GPUs ; forward and backward pass different in terms of required communication (e.g. Kurth et. al, 2022)

=> increases instantaneous available memory for more edges / nodes and also possible to split activations across GPUs

- Aggressive re-computing activations allows for rollouts of ~ 1. degree model on a single GPU
- Model parallel set-up makes it possible to go to native ERA5 resolution (and beyond?) and faster time to solution ; it will also allow to build larger models in the future

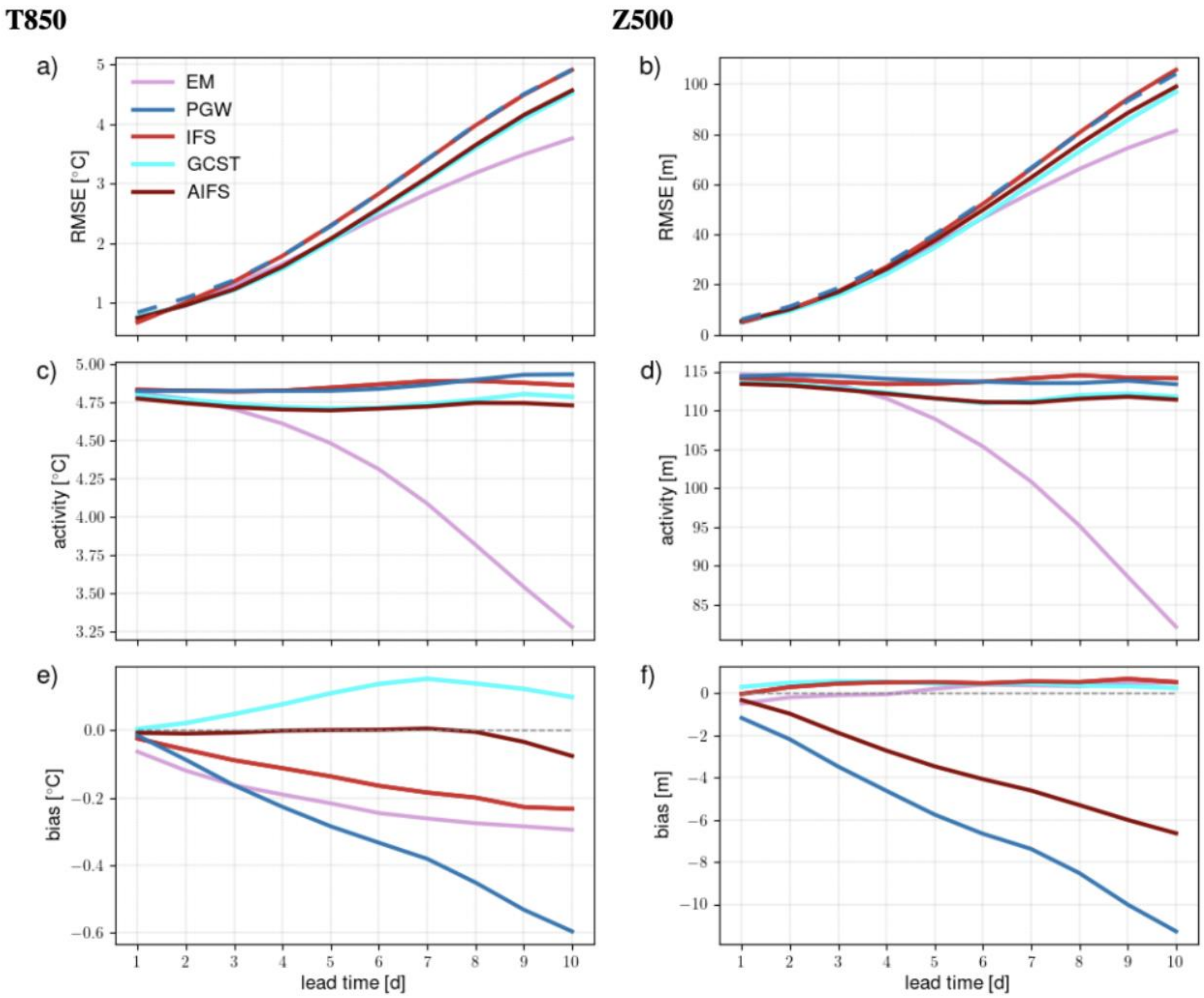
Headline score: anomaly correlation for Z500, Summer 2023



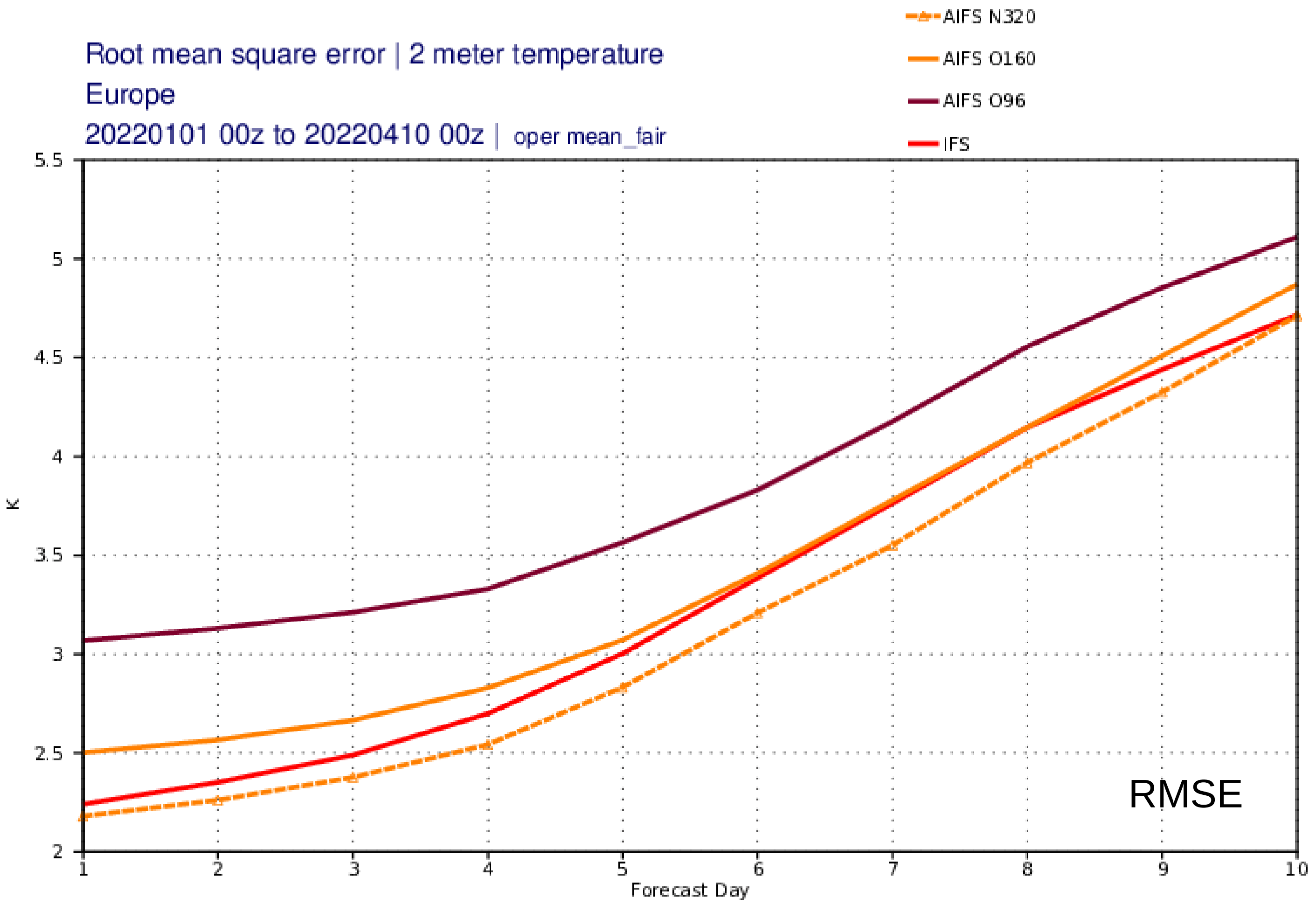
ECMWF

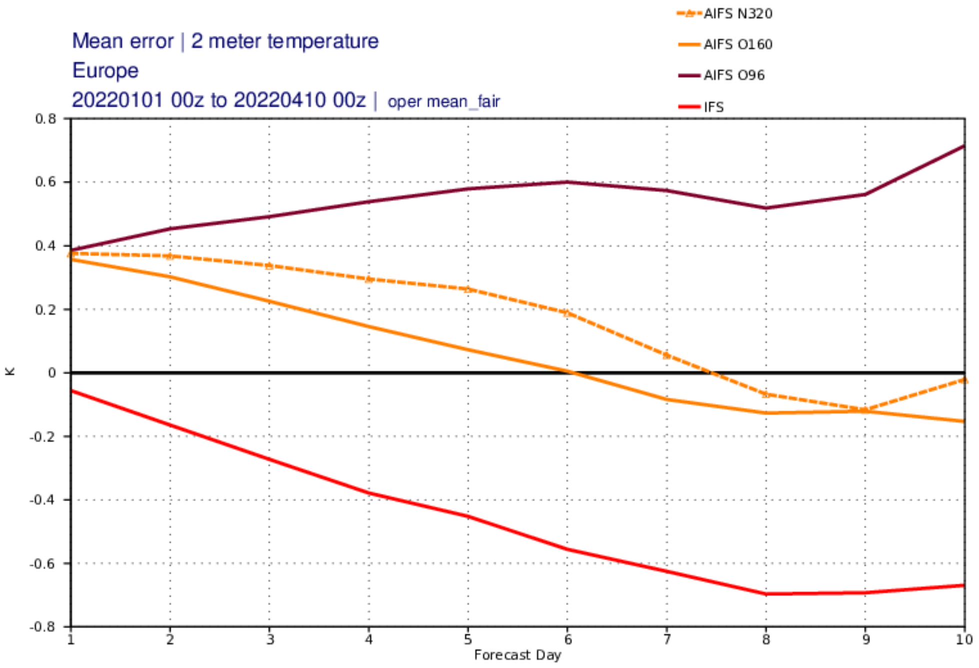
RMSE, activity, and bias

DJF 2022/2023
NHem Extratropics



Verification against SYNOP observations





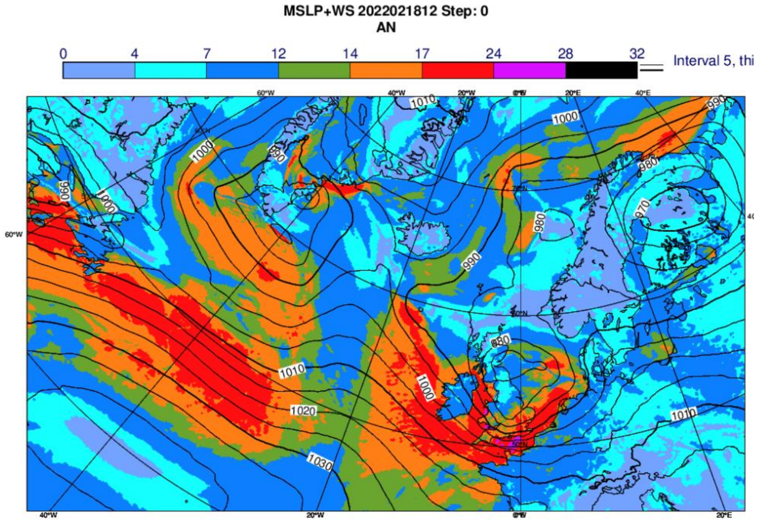
BIAS

ECMWF

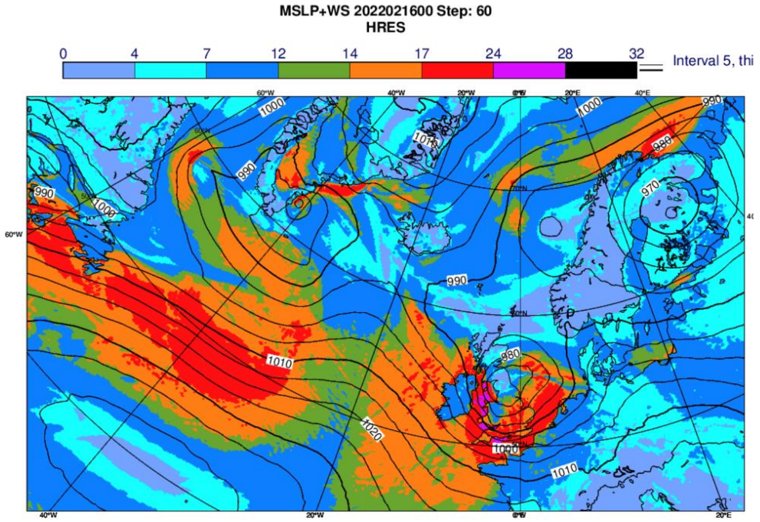
2022-02-16 00z + 60h (Storm Eunice over UK)

See ECMWF Newsletter 176

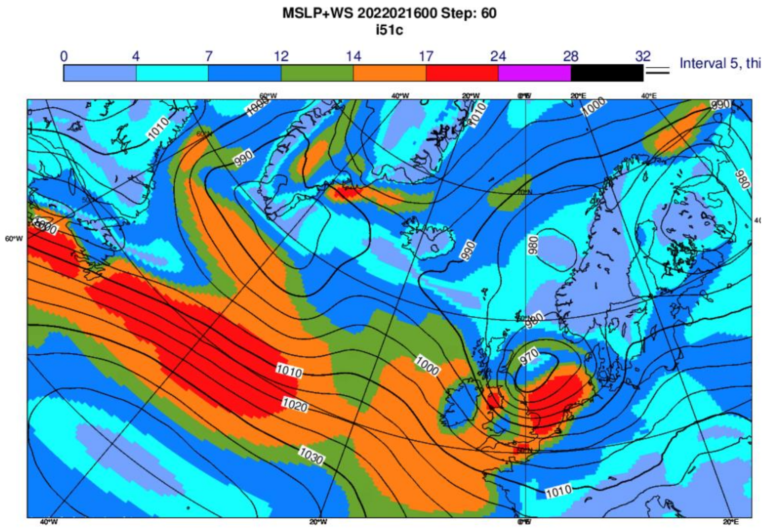
Analysis



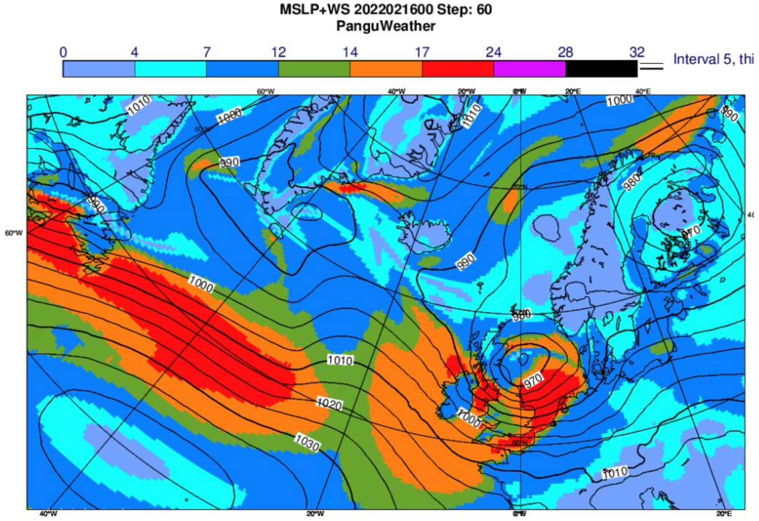
HRES



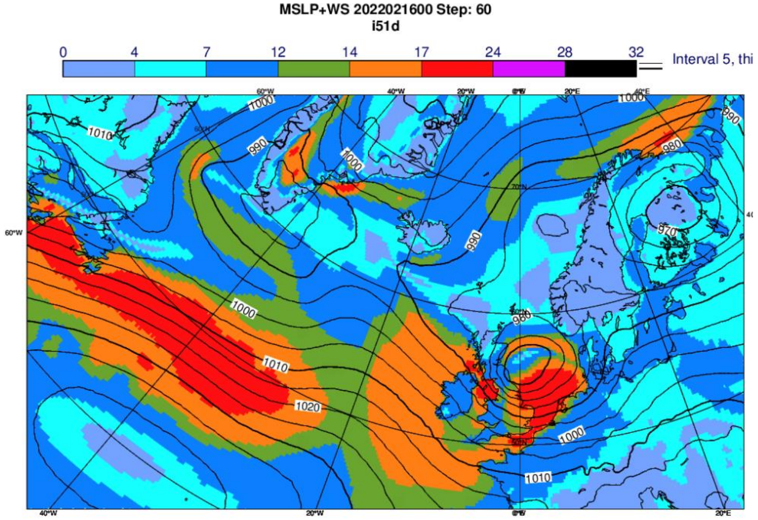
Fourcastnet



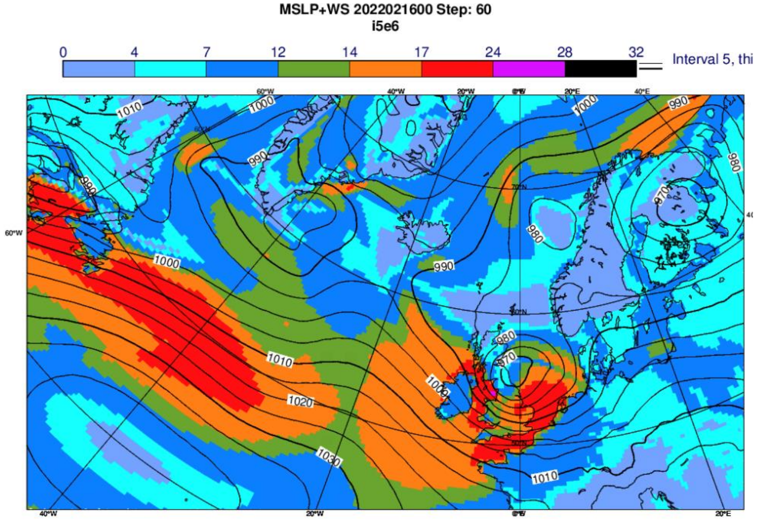
PanguWeather



Graphcast

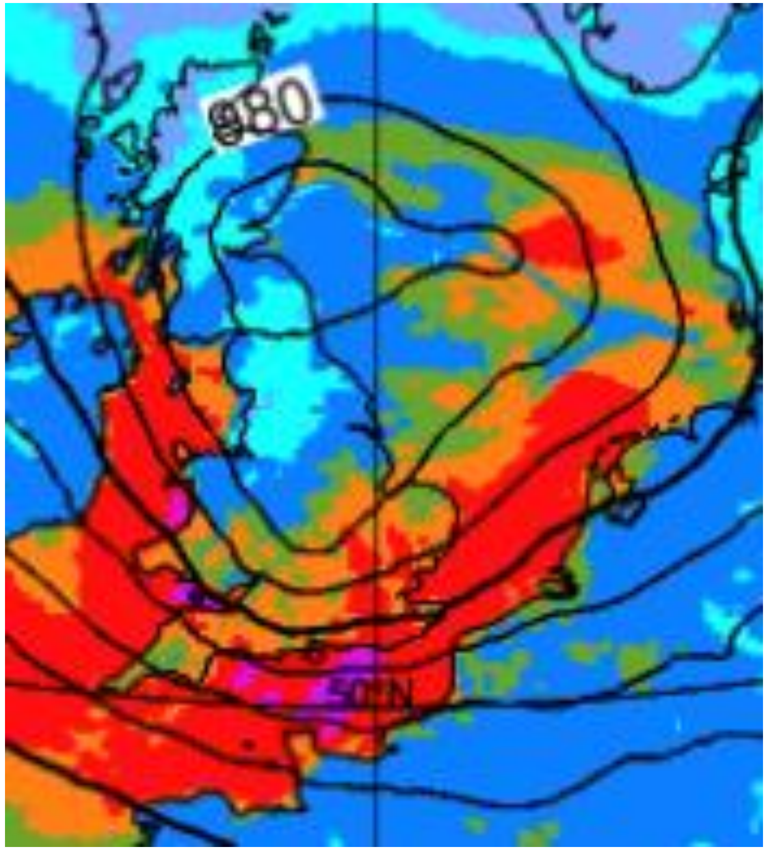


AIFS n320

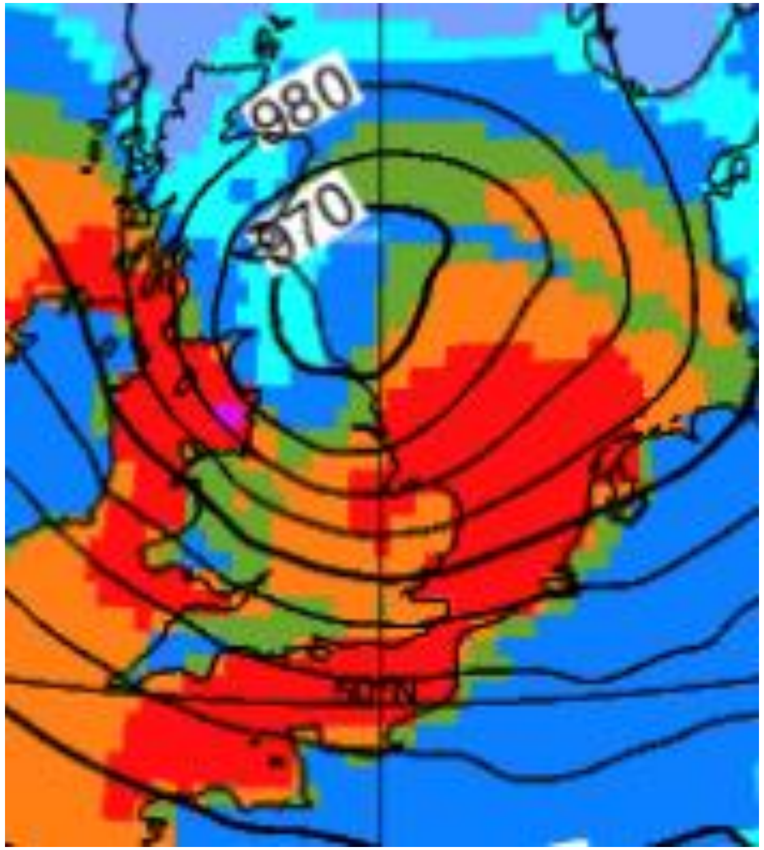


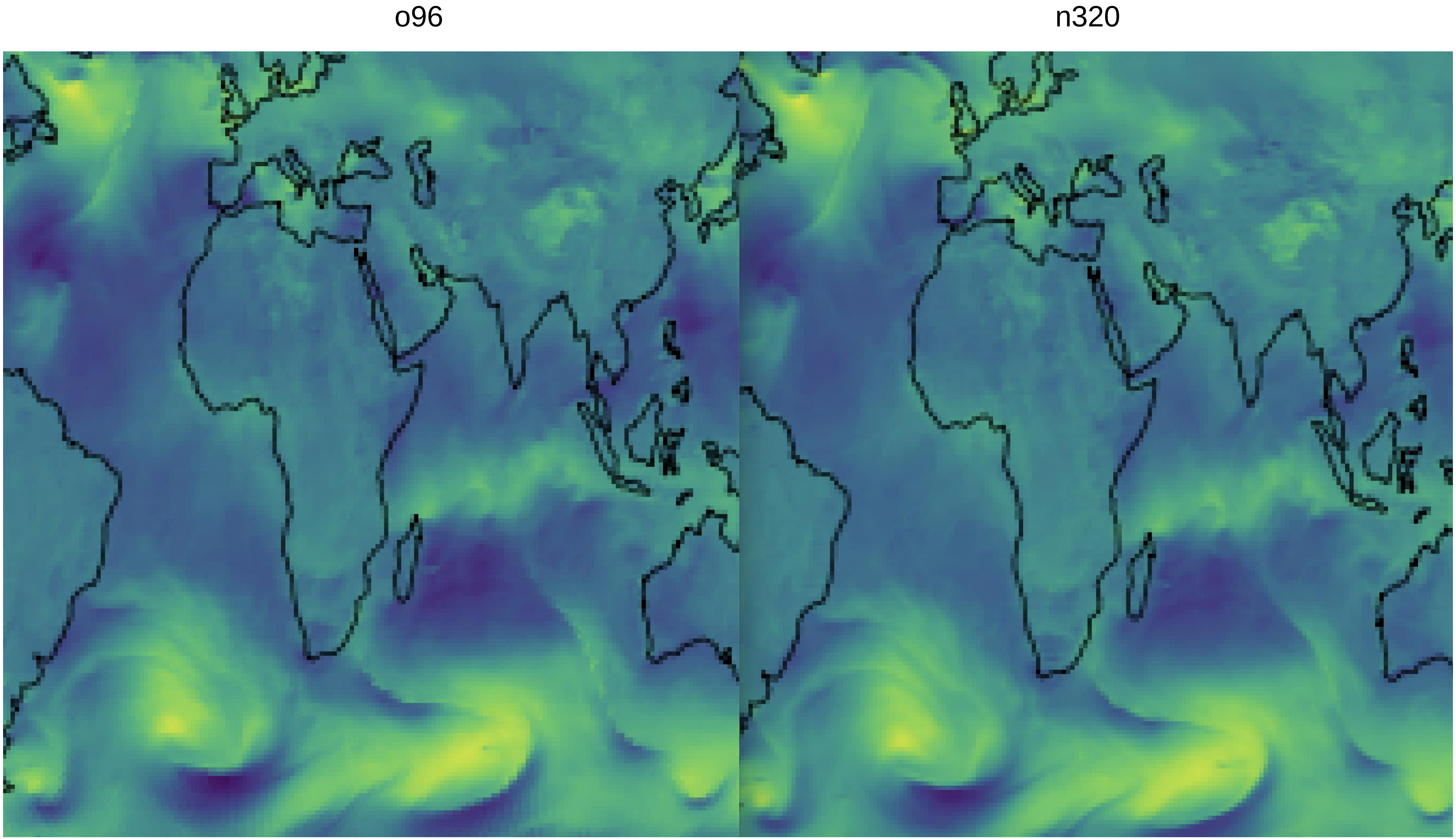
2022-02-16 00z + 60h (Storm Eunice over UK)

IFS

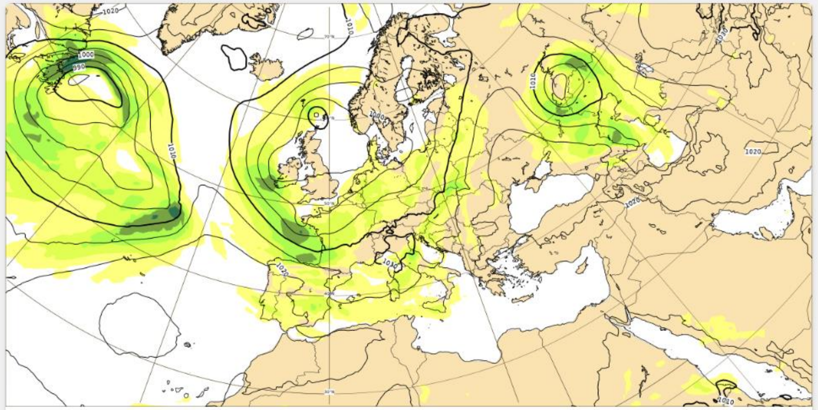


AIFS

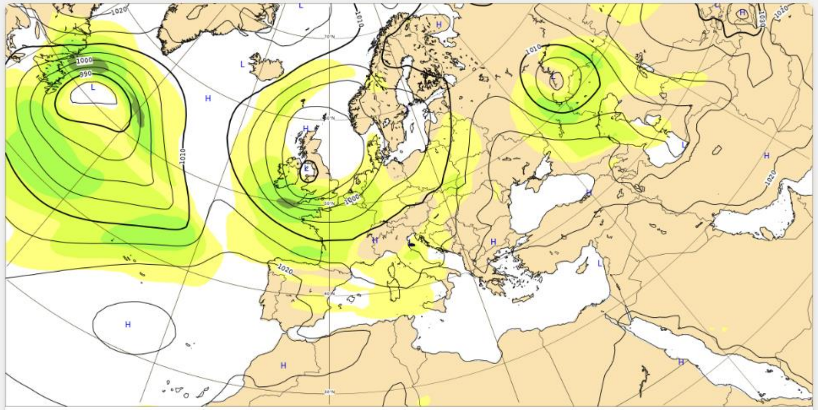



o96
n320

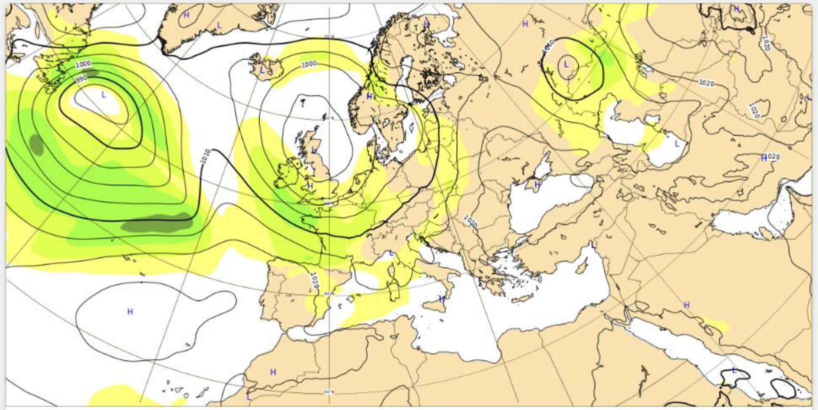
Forecast from 06/11/2023 – 96h ; https://charts.ecmwf.int



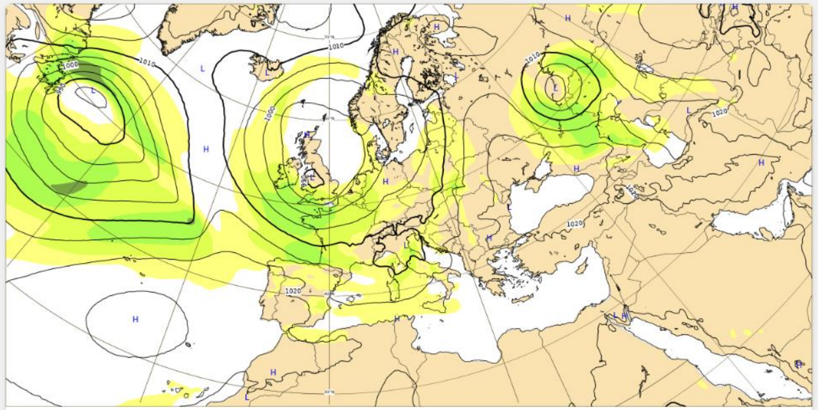
Mean sea level pressure and 850 hPa wind speed



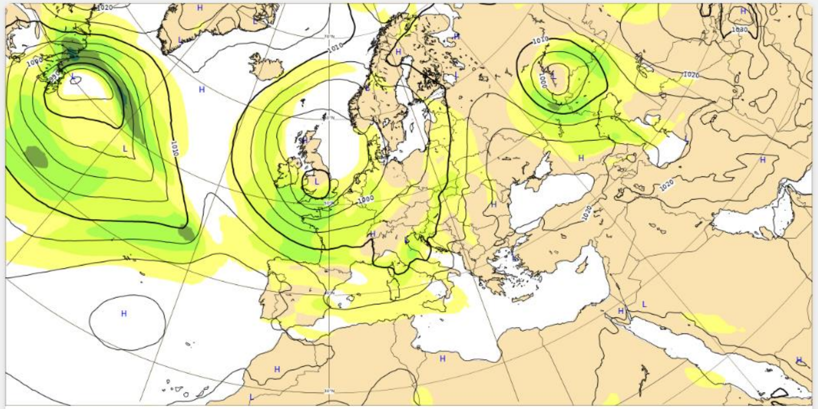
Experimental: AIFS (ECMWF) ML model: Mean sea level pressure and 850 hPa wind speed



Experimental: FourCastNet ML model: Mean sea level pressure and 850 hPa wind speed



Experimental: GraphCast ML model: Mean sea level pressure and 850 hPa wind speed

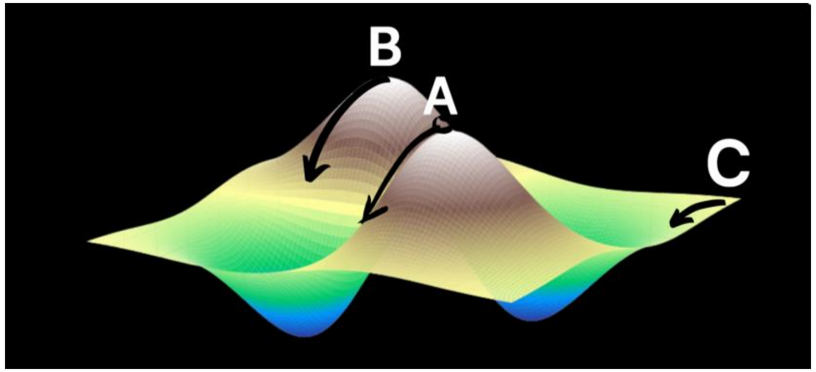


Experimental: Pangu-Weather ML model: Mean sea level pressure and 850 hPa wind speed

Towards Ensembles -> Preliminary ideas to capture uncertainty

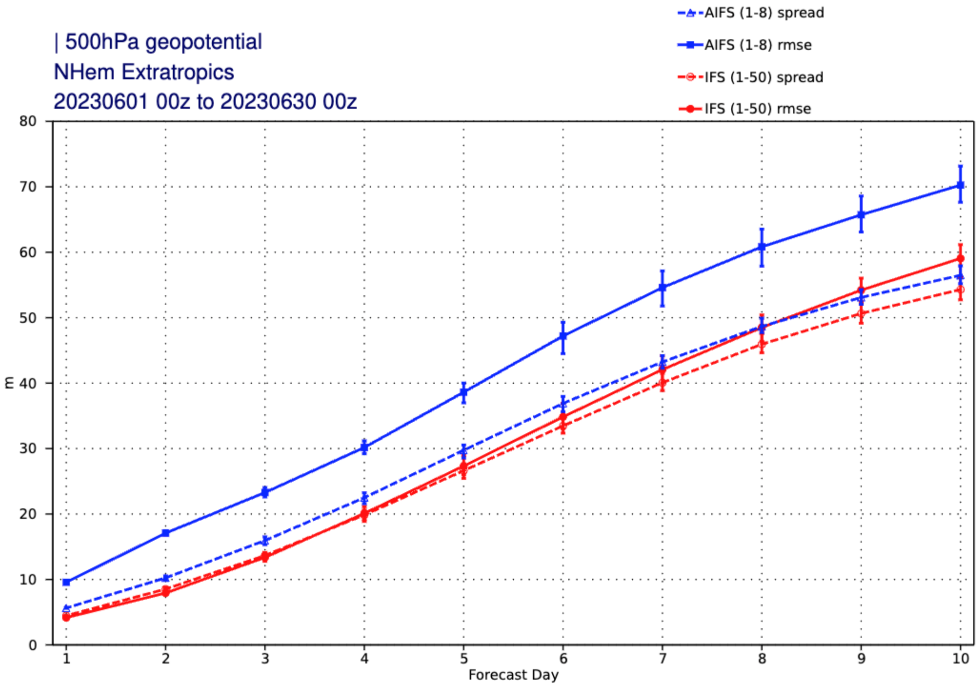
1. Use the **ensemble initial conditions** to initialise the AI model

2. Using the randomness of the optimisation process to find different quasi-optimum solutions of the NN weights that can be used to build a ***multi-model ensemble***



3. Train to minimise ***probabilistic scores*** like kernel CRPS or optimise distributions

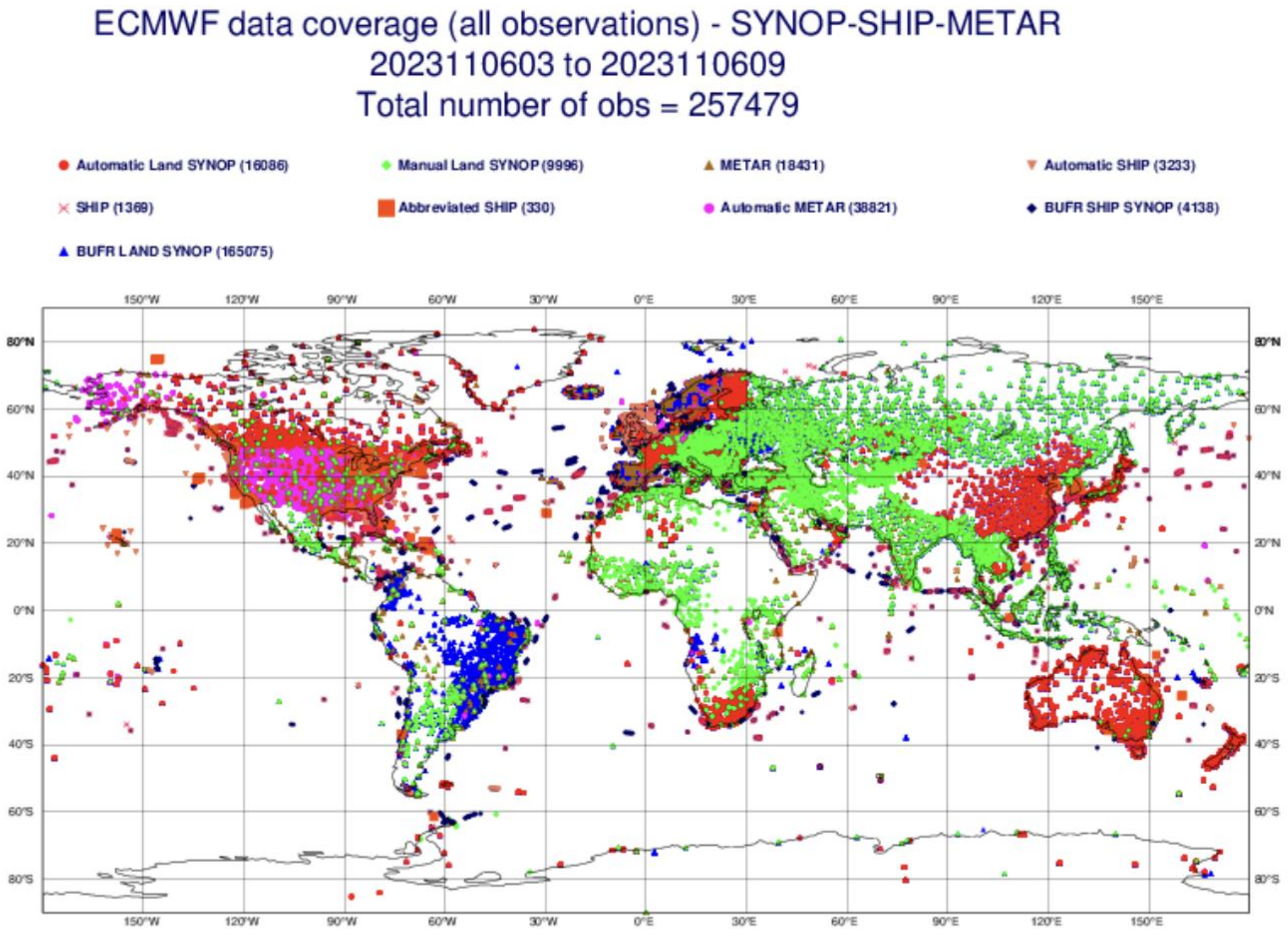
4. ***Generative models*** where NN is trained on real samples and then generates new samples. Examples include Generative Adversarial Networks and Diffusion models



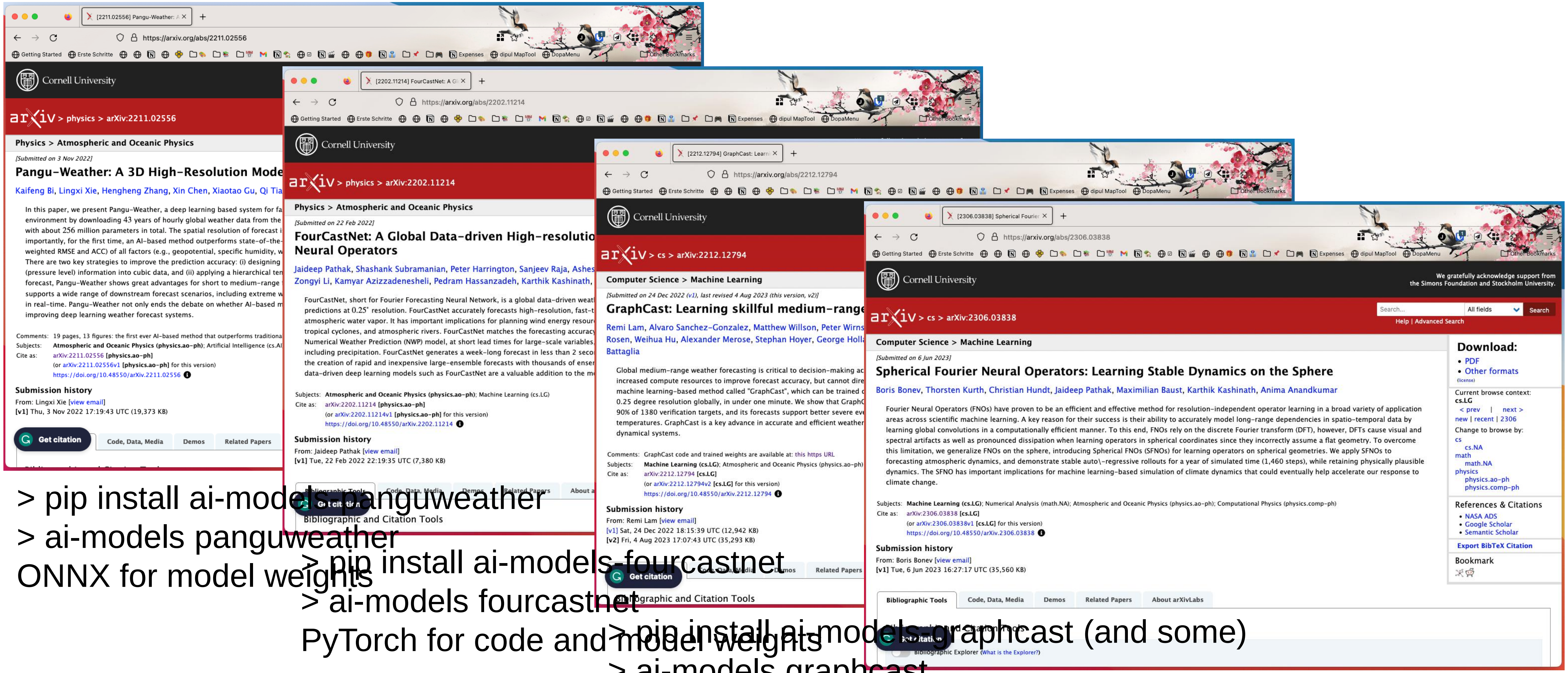
Learning from observations: challenges

Multitude of scattered and noisy data needs to be fused into coherent representation.

- Data is sparse in space and time
- Heterogeneity of data sources



AI-Models Plugins for FOSS Data-Driven NWP



> pip install ai-models-panguweather
> ai-models panguweather
ONNX for model weights

> pip install ai-models-fourcastnet
> ai-models fourcastnet
PyTorch for code and model weights

> pip install ai-models-graphcast (and some)
> ai-models graphcast
Jax for code and model weights

> pip install ai-models-fourcastnetv2
> ai-models fourcastnetv2
PyTorch for code and model weights

prepml

- **prepml** is the companion tool to **ai-models**
- Uses ecFlow
- It allows to run inferences over many years
- Archives all outputs in the MARS archive in research mode
- It feeds into ECMWF's scores database so that models can be evaluated
- It allows users to run development code as well
- It can create ensembles using various combinations of models, inputs, ...

Summary

- We have a baseline model
- Model can be scaled up to high-resolution; good scaling for at least O(100) GPUs
- Extensive supporting software infrastructure
- GNN can support arbitrary grids, including high resolution over only parts of the globe
- Next …
  - Different methods for constructing reliable ensemble forecasts
  - Make use of observations …
  - Further improve model, more output parameters, …

ECMWF

Questions?