



On the potential and limitations of ML forecast models

Massimo Bonavita

Principal Scientist

Data Assimilation Team Leader

ECMWF

massimo.bonavita@ecmwf.int

Motivation...

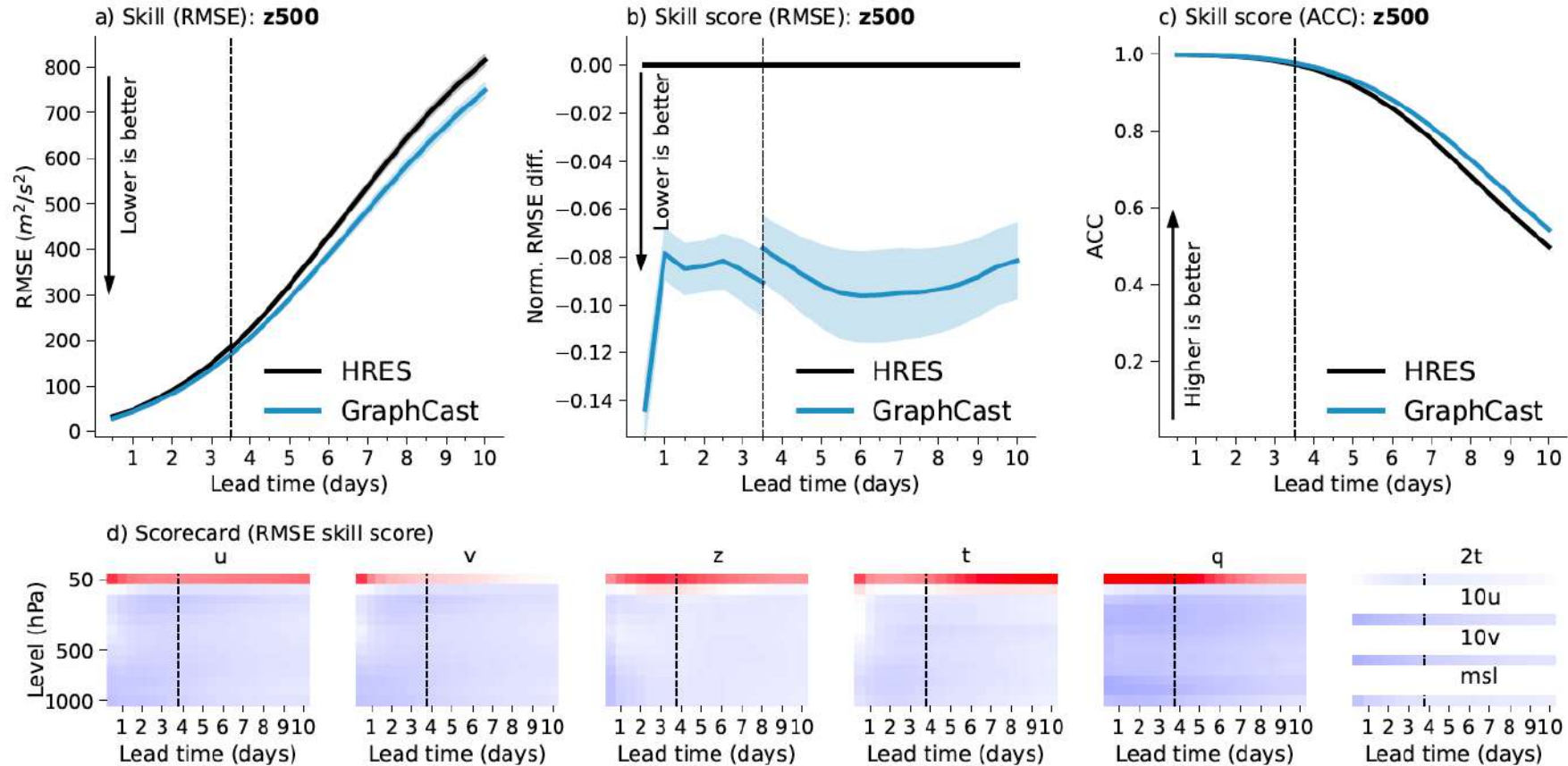


Figure 2 | Skill and skill scores for GraphCast and HRES in 2018. (a) RMSE skill (y-axis) for GraphCast (blue lines) and HRES (black lines), on z500, as a function of lead time (x-axis). Error bars represent 95%

A brave New World: Machine Learning Weather Prediction

- ML models for medium/extended-range weather prediction, trained on ERA5 reanalysis
- Starts with Dueben and Bauer, 2018, low-resolution Z500 field as an image-to-image problem, results not too exciting
- Turning point: Keisler, 2022, multiple vertical levels (13), higher resolution (1deg), Graph NN. Skill comparable to GFS, still below ECMWF
- Floodgates open: FourCastNet (NVIDIA, Pathak et al., 2022), Pangu-weather (Huawei, Bi et al., 2022), GraphCast (Google-DeepMind, Lam et al., 2022), FengWu (Academic, Chen et al., 2023)...
- Each claims to outperform all previous MLWP model and all traditional physics-based NWP systems, and at a fraction of the cost!!



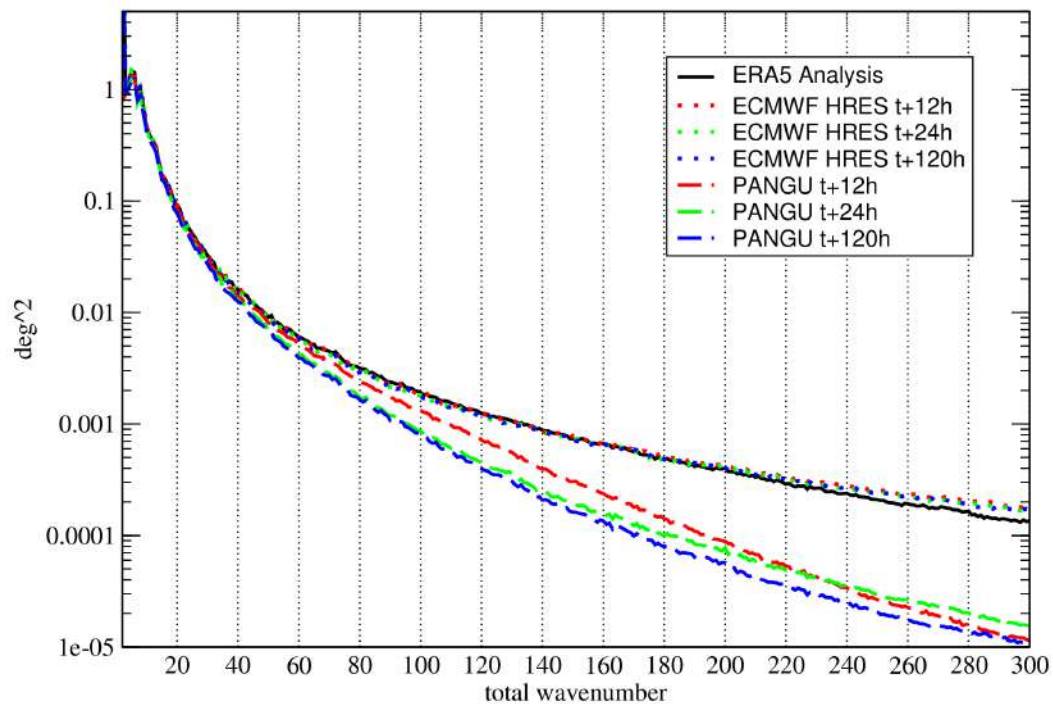
A look under the hood of MLWP models

- NN Architecture:
 1. Pangu-Weather, FourCastNet: **Vision Transformer**
 2. GraphCast and similar: **Graph Neural Network**
- Training dataset: ERA5, 0.25 deg, O(10) pressure levels + surface fields
- Timestepping:
 1. FourCastNet, GraphCast, and others: autoregressive, 6h timestep $X^t = ML(X^{t-\Delta t})$
 2. Pangu-Weather: “Hierarchical Temporal Aggregation”, i.e. train 4 separate NNs to forecast at t+1, 3, 6, 24 hours and combine them as required

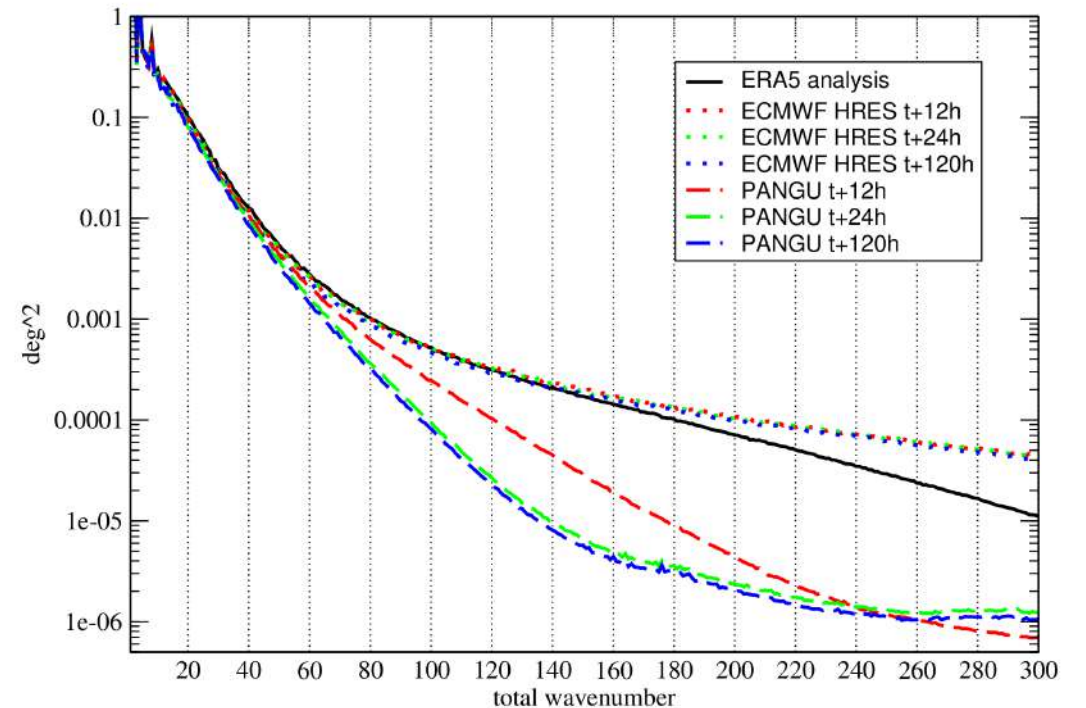
A look under the hood of MLWP models

- ERA5 Analysis and **Pangu-Weather** forecast power spectra:

Temperature @850hPA



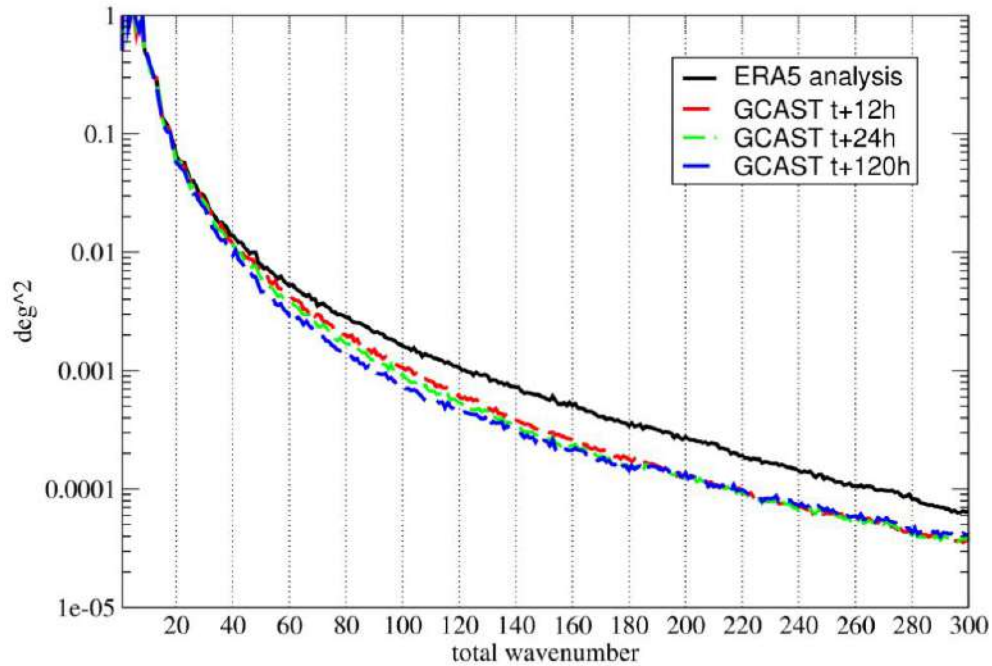
Temperature @250hPA



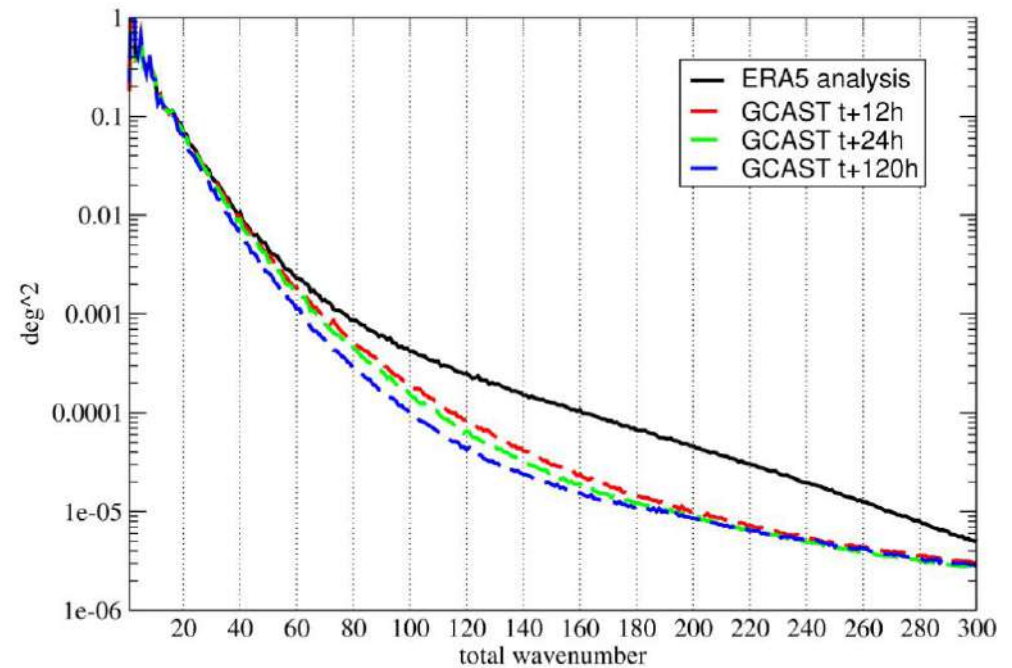
A look under the hood of MLWP models

- ERA5 Analysis and **GraphCast** forecast power spectra:

Temperature @850hPA



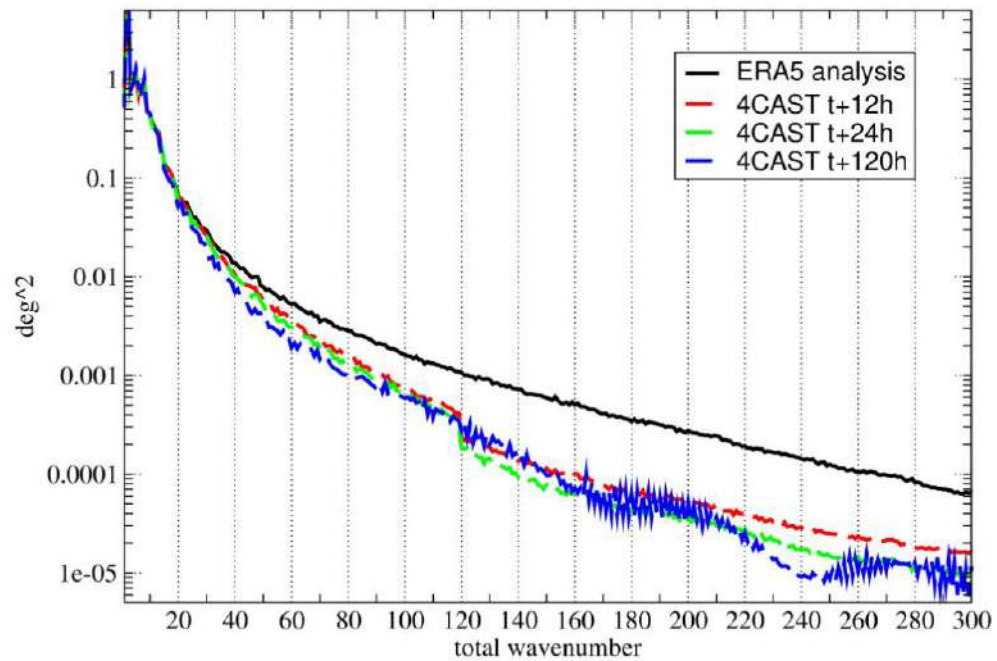
Temperature @250hPA



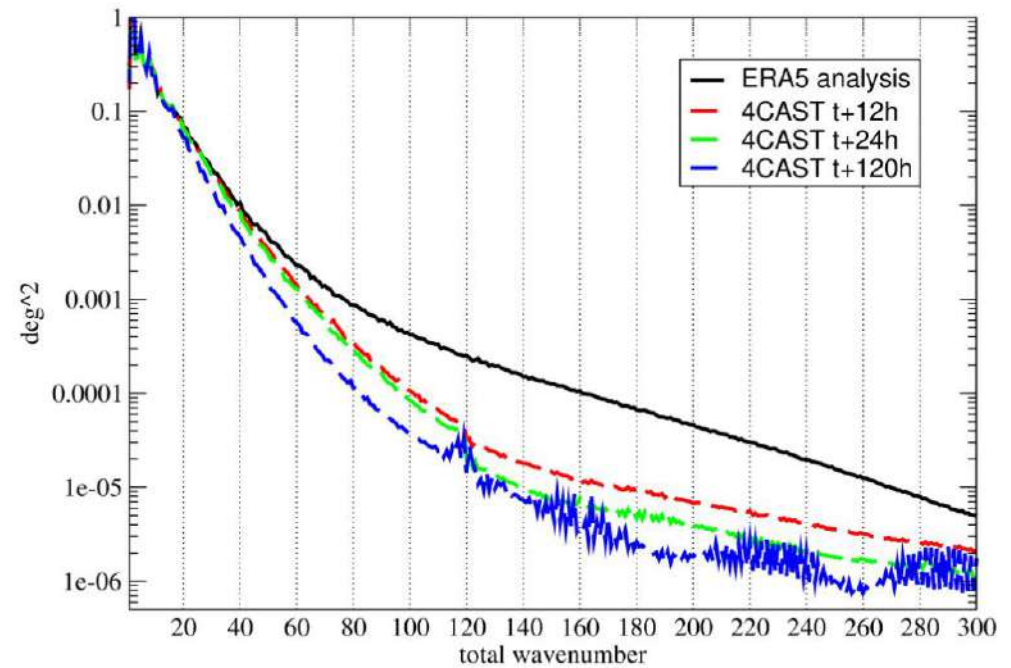
A look under the hood of MLWP models

- ERA5 Analysis and **FourCastNet** forecast power spectra:

Temperature @850hPA



Temperature @250hPA



Forecast skill

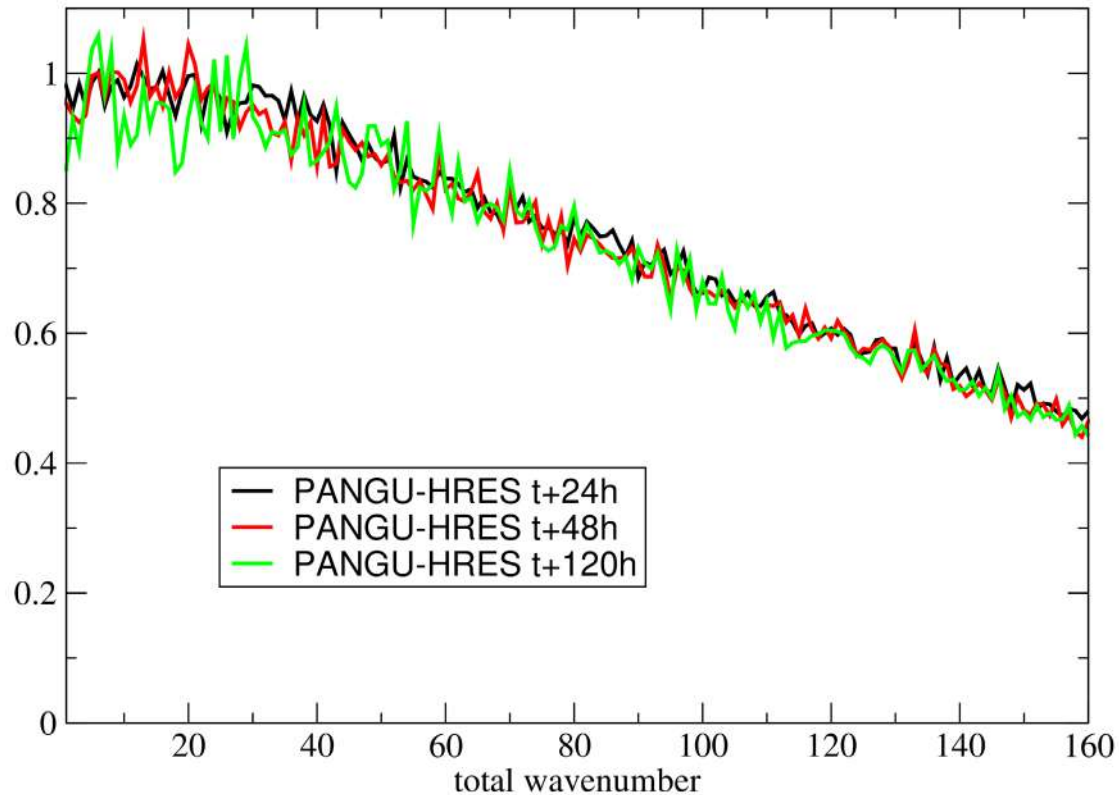
Is the shape of the ML models spectra a factor in their forecast skill?

- Method:
 - 1) Compute the ratio of spectral variances of ML models' forecasts to IFS forecasts at all forecast ranges
 - 2) Use 1) to spectrally filter IFS forecast output
 - 3) Verify filtered IFS forecast output

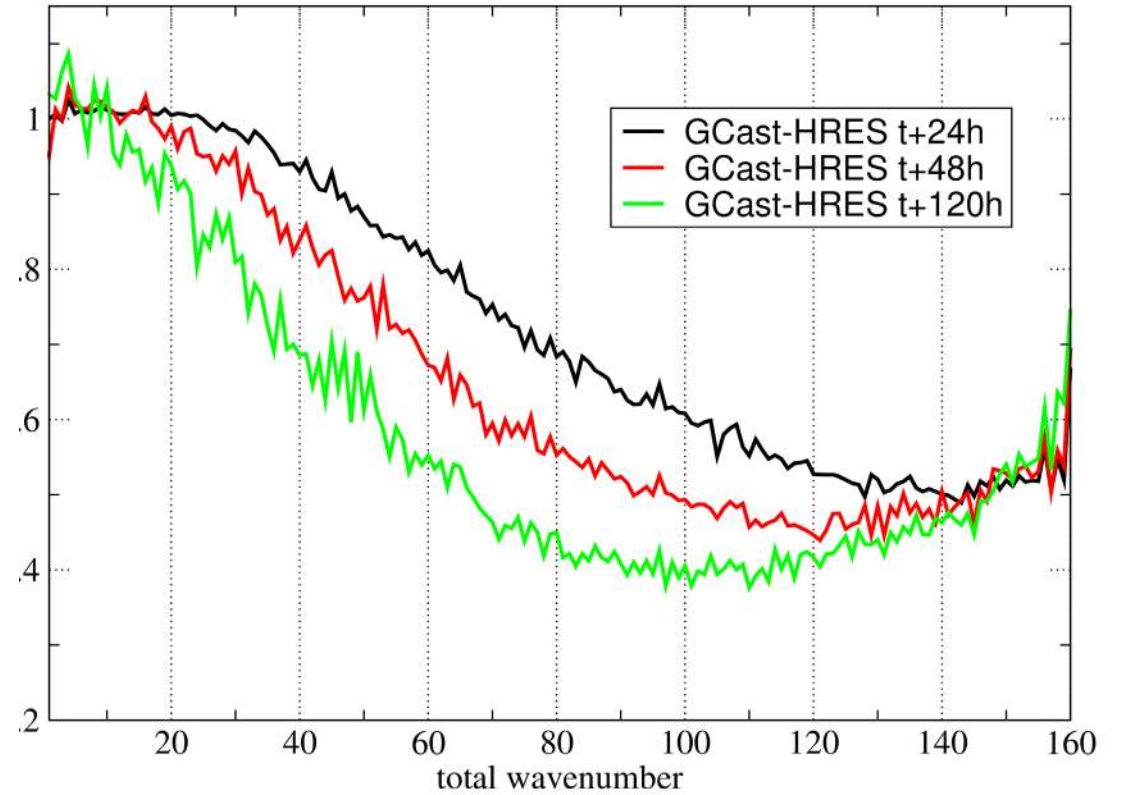
Forecast skill

Is the shape of the ML models spectra a factor in their forecast skill?

Pangu_W – IFS ratio @U10

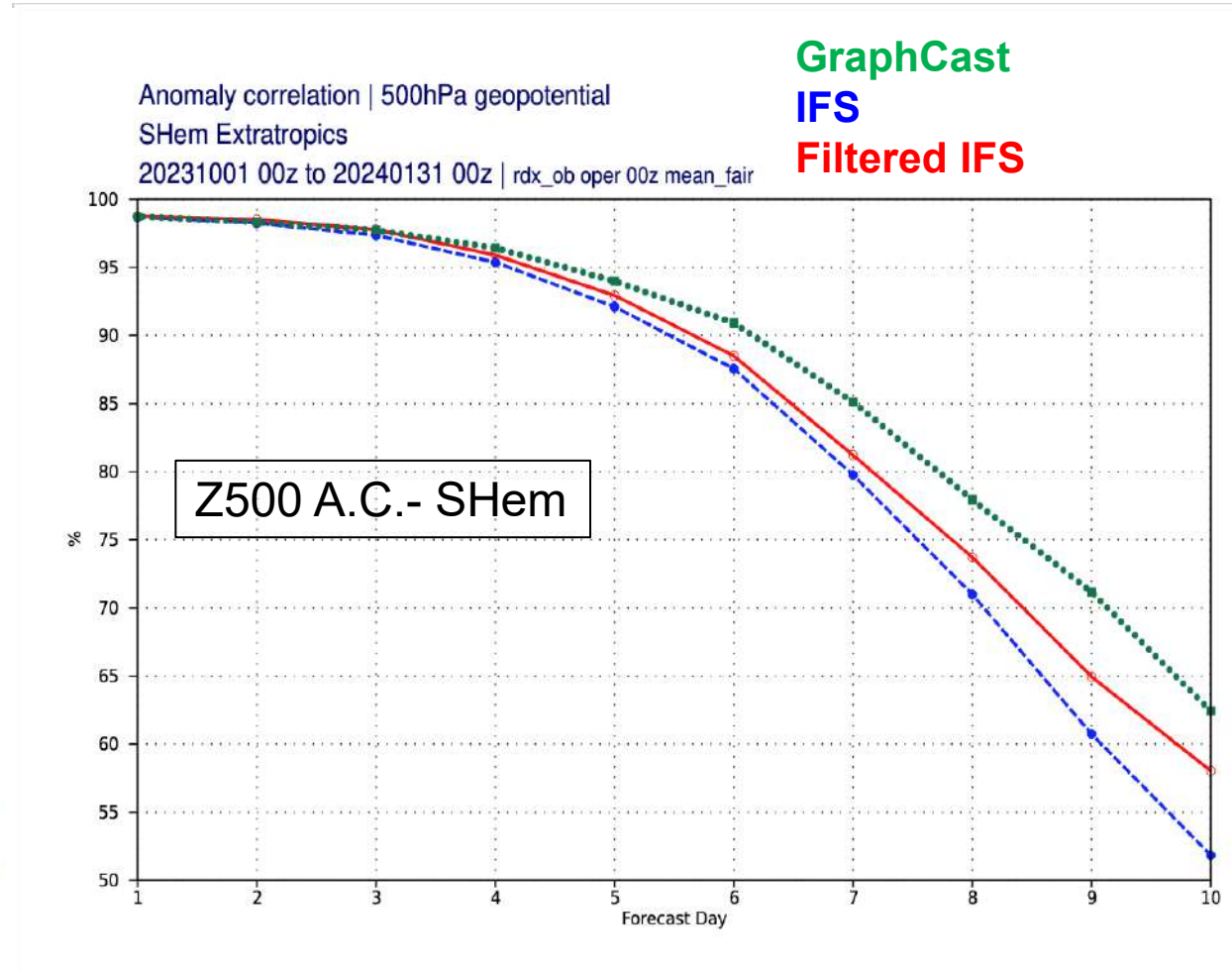
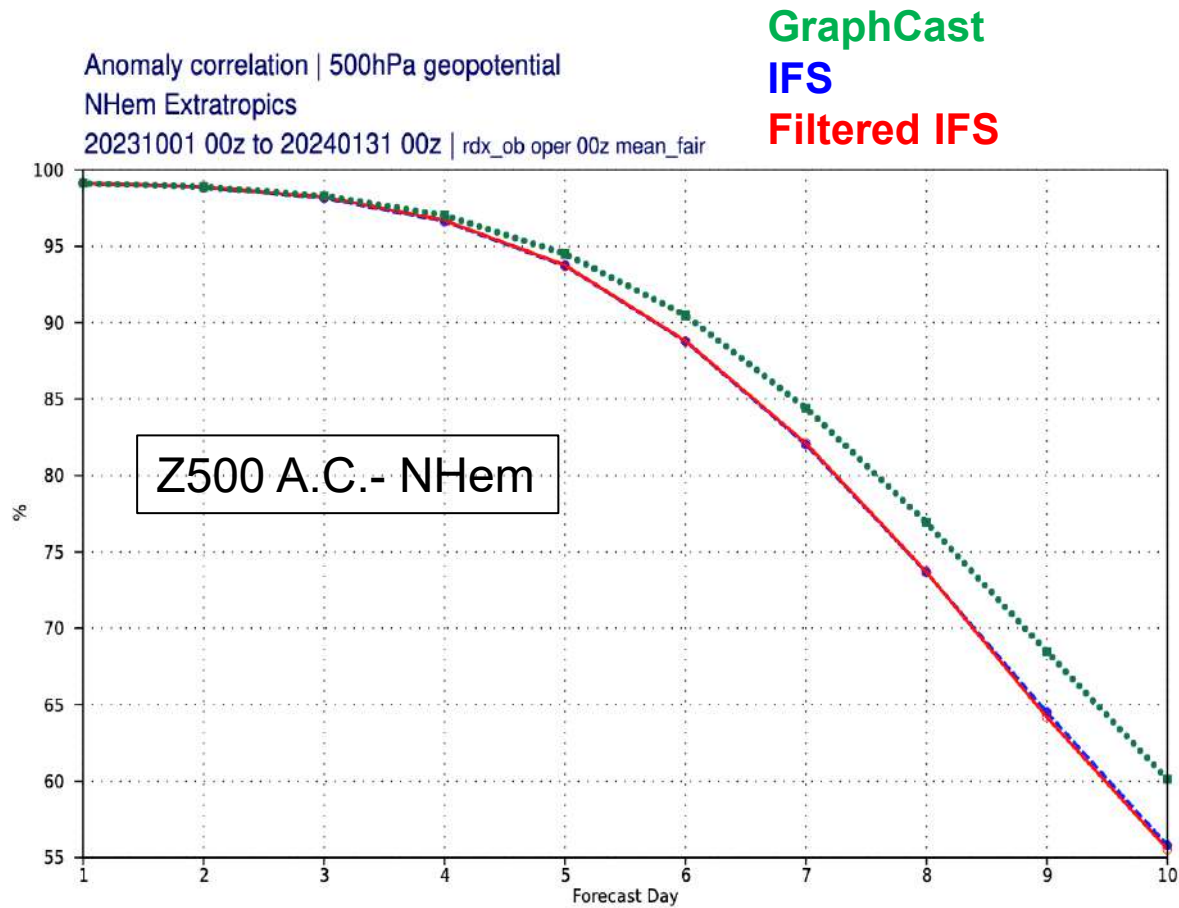


GraphCast – IFS ratio @Z500



Forecast skill

Is the shape of the ML models spectra a factor in their forecast skill?

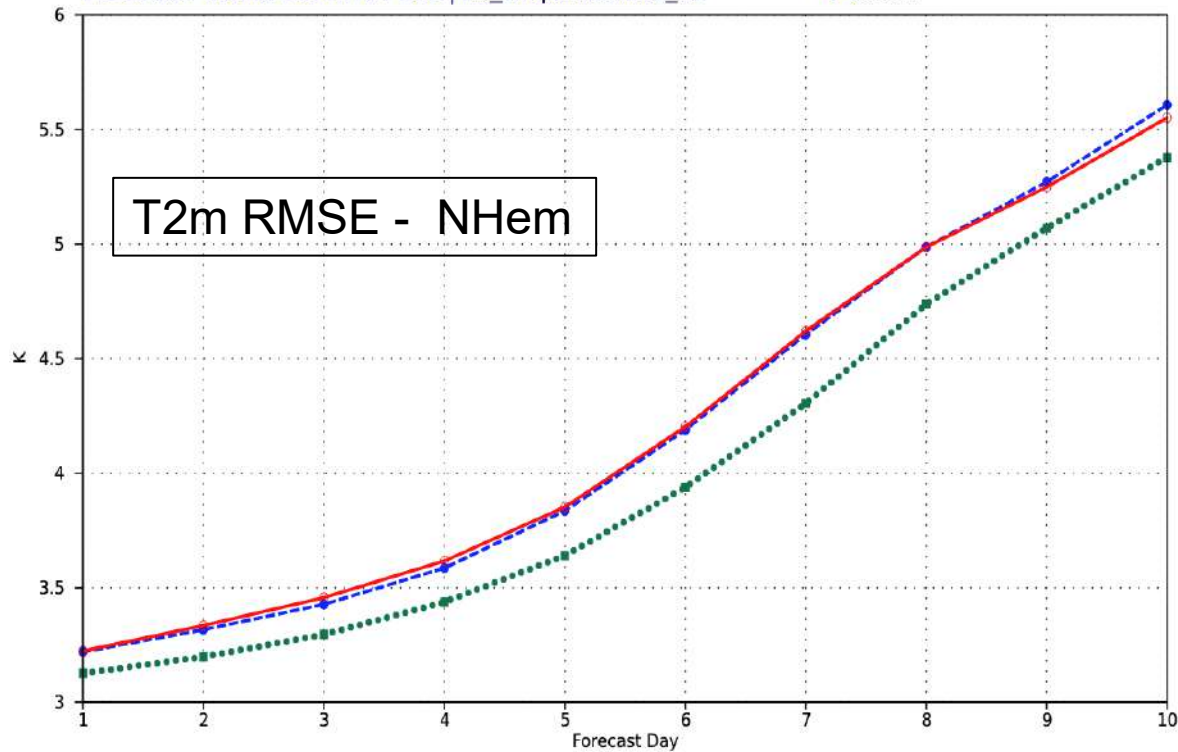


Forecast skill

Is the shape of the ML models spectra a factor in their forecast skill?

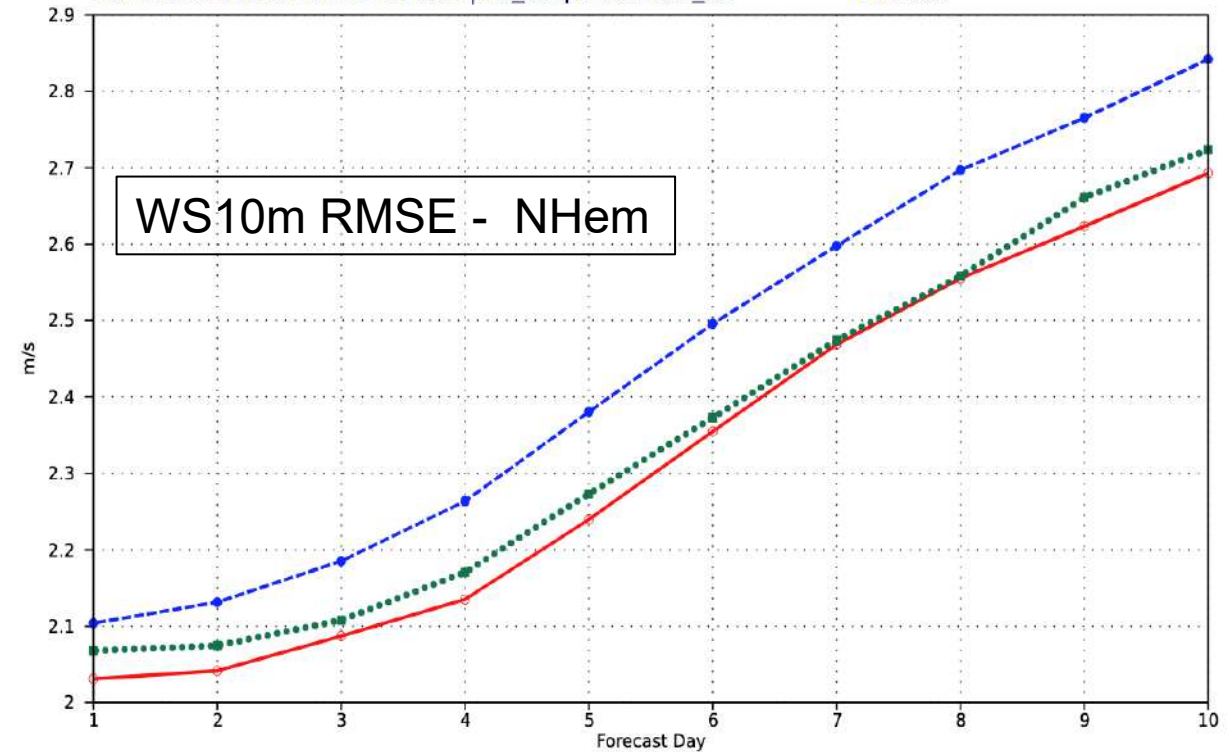
GraphCast
IFS
Filtered IFS

Root mean square error | 2 meter temperature
NHem Extratropics
20231001 00z to 20240131 00z | rdx_ob oper 00z mean_fair



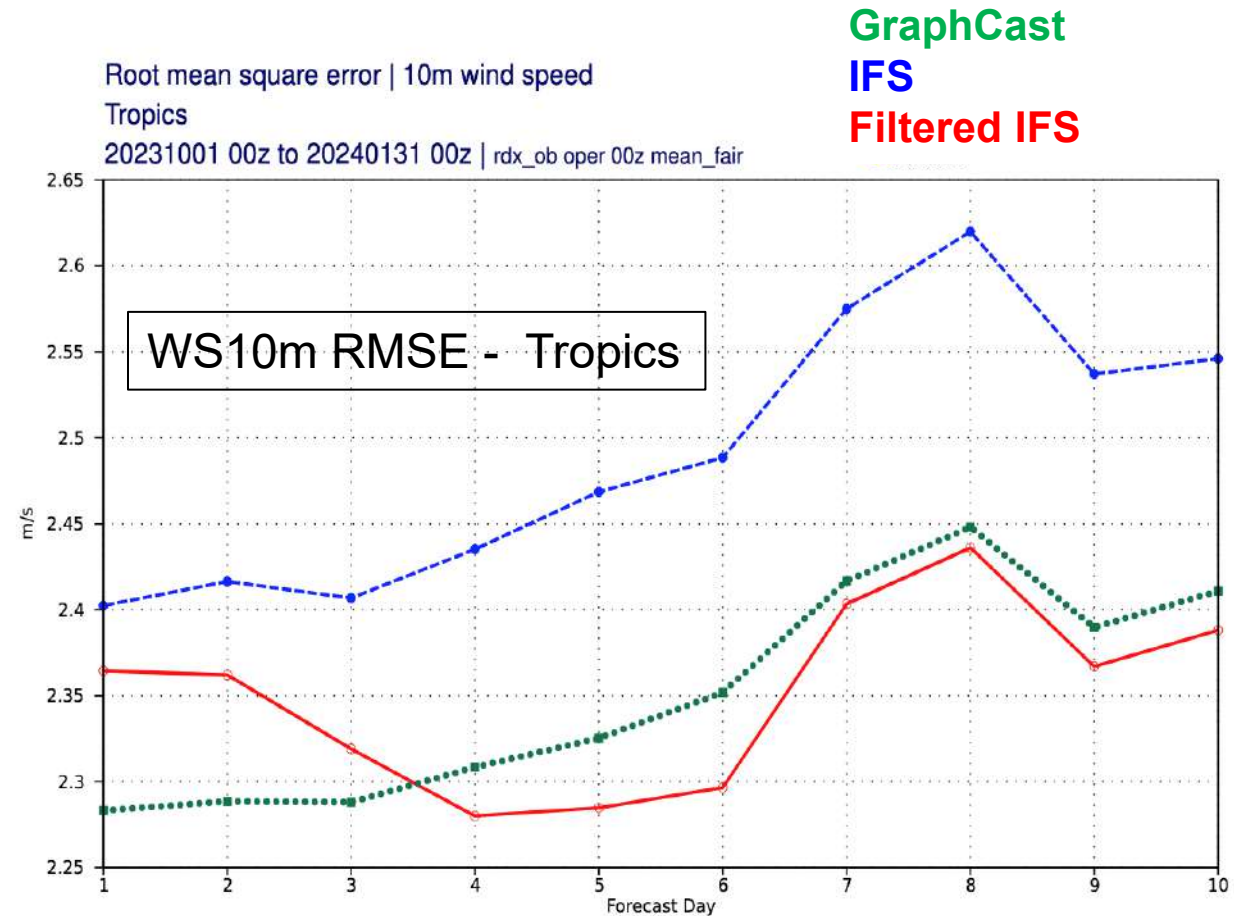
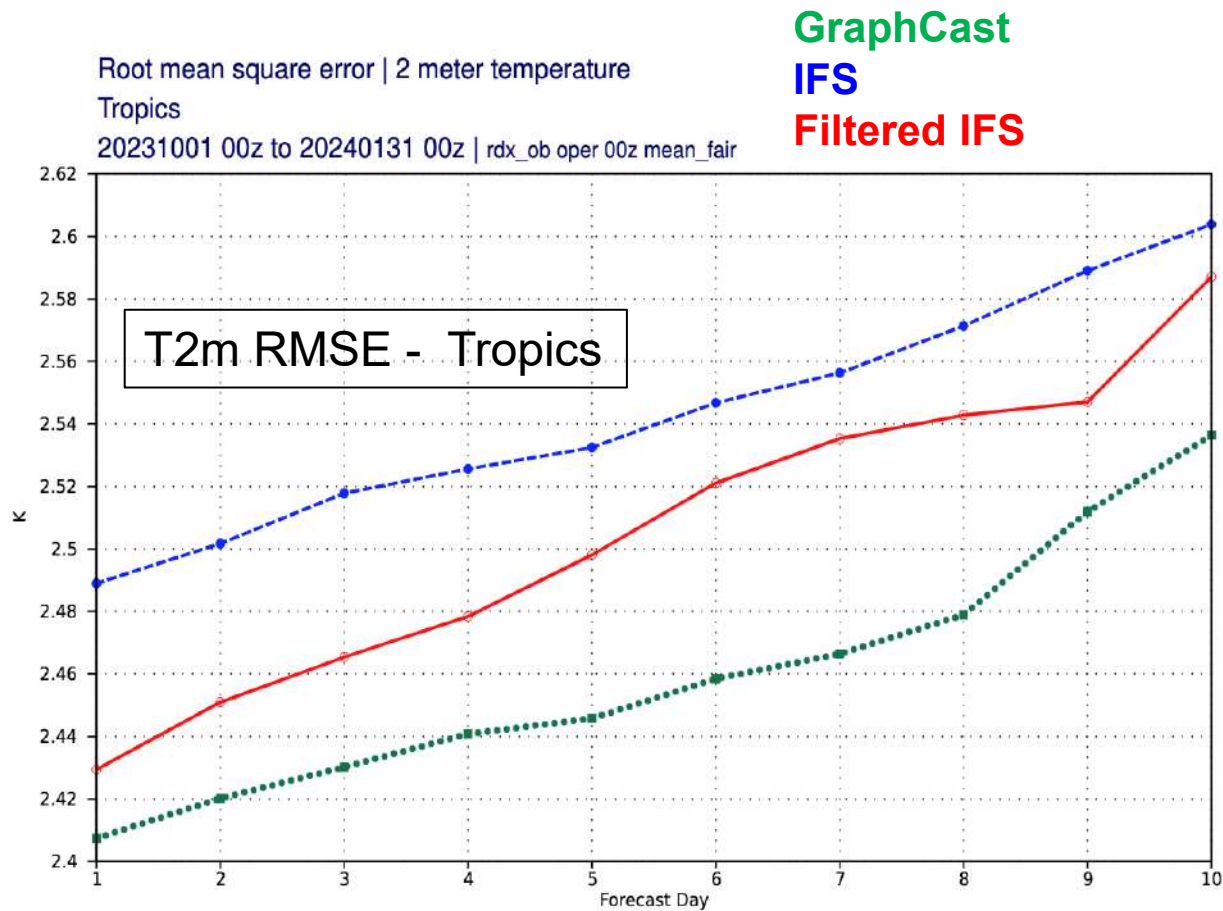
GraphCast
IFS
Filtered IFS

Root mean square error | 10m wind speed
NHem Extratropics
20231001 00z to 20240131 00z | rdx_ob oper 00z mean_fair



Forecast skill

Is the shape of the ML models spectra a factor in their forecast skill?



Forecast skill

Is the shape of the ML forecast spectra a factor in their forecast skill?

- **Definitely! But not the only one...**
- ML skill for variables with longer error correlation length scales (e.g. geopot., temperature) does not benefit much from ML spectra filtering
- Skill of variables with redder error spectra (wind, humidity) is largely driven by intelligent smoothing of ML forecasts
- What is/are the other ingredients of ML skill?
 - Ability to do online correction of flow-dependent model errors
 - Lack of upscale error growth from convection and moist processes in the forecast (e.g., Zhang et al., 2002; Selz and Craig, 2023)
 - to be continued...

A look under the hood of MLWP models

- The peculiarities of the forecast spectra of ML models have other consequences beyond interpretation of forecast skill measures
- One of them is the **lack of physical consistency**

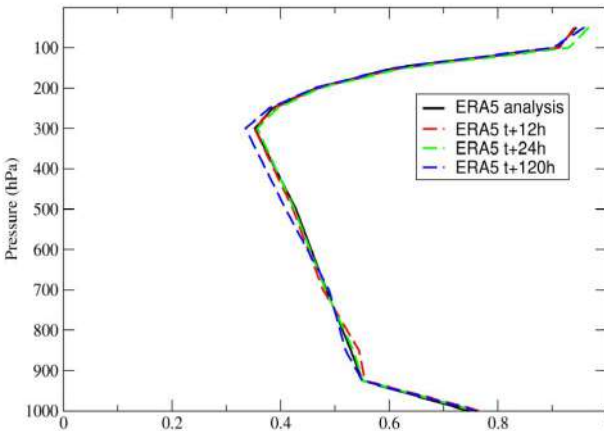
ML models dynamics

Vorticity and divergence decomposition of the flow:

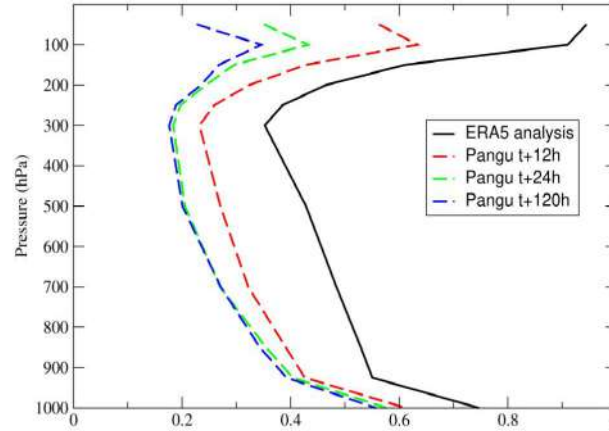
$$\mathbf{u} = \mathbf{u}_d + \mathbf{u}_v = -\nabla\chi + \mathbf{k} \times \nabla\psi$$

$$\nabla^2\chi = \delta, \quad \nabla^2\psi = \zeta$$

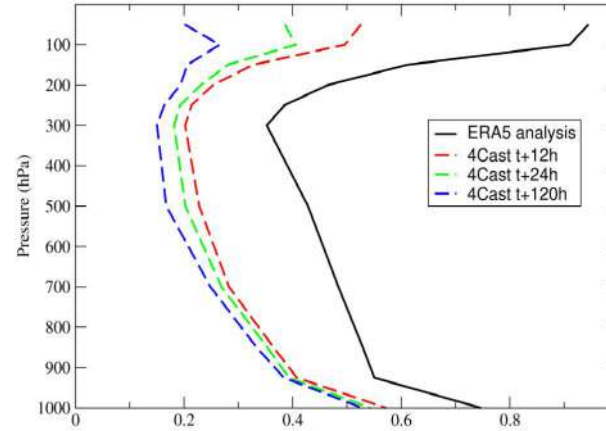
ERA5



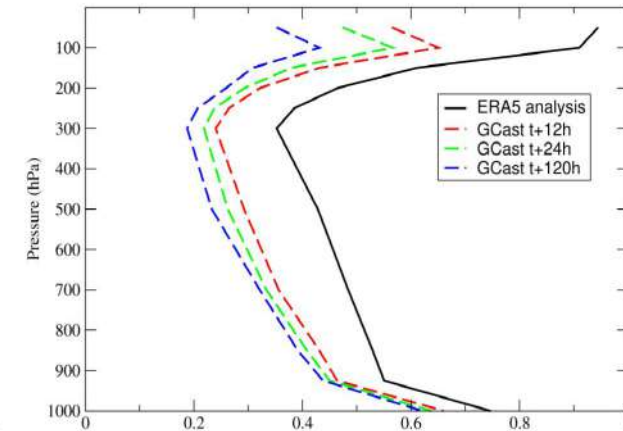
Pangu-W



4CastNet



GraphCast



$$|\delta|/|\zeta|$$

ML models dynamics

What does this mean?

Divergent motions dynamically drive vertical motions.

Vertical velocity is not typically predicted by ML models but can be diagnosed by integrating the continuity equation on forecasted pressure-level fields (Holton and Hakim, 2012):

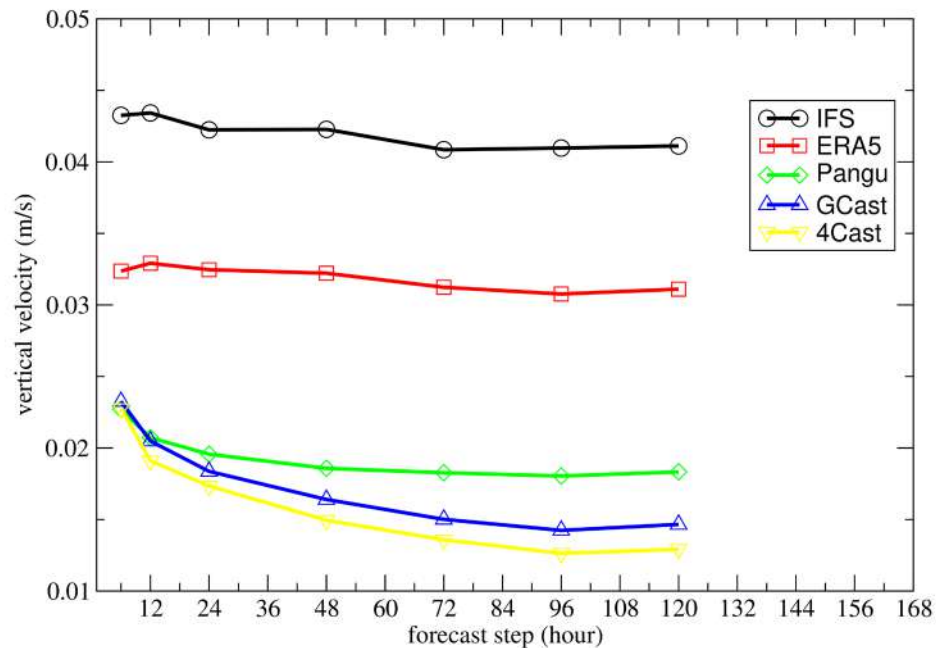
$$\omega(p) = \omega(p_s) - \int_{p_s}^p \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right)_p dp$$

This is a (surprisingly!) good proxy over Ocean and low topography areas

ML models dynamics

$$\omega(p) = \omega(p_s) - \int_{p_s}^p \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right)_p dp$$

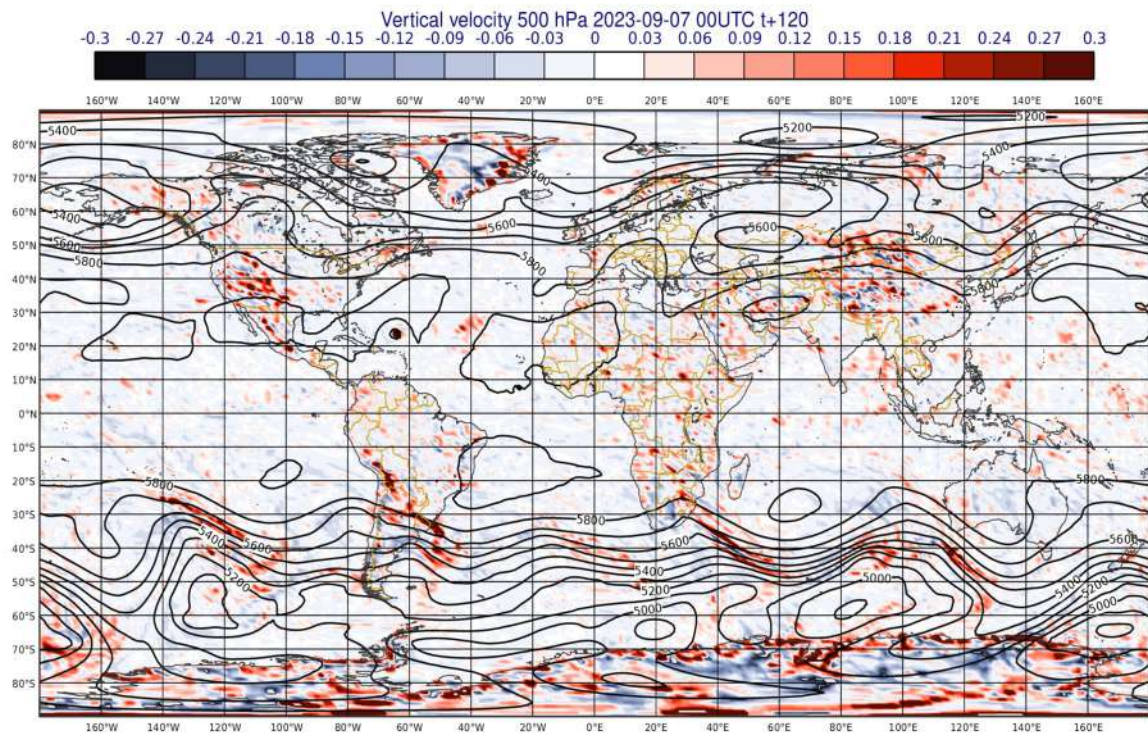
The progressive reduction in the magnitude of the ML models forecasted divergence leads to **increasingly weak vertical velocity** forecasts:



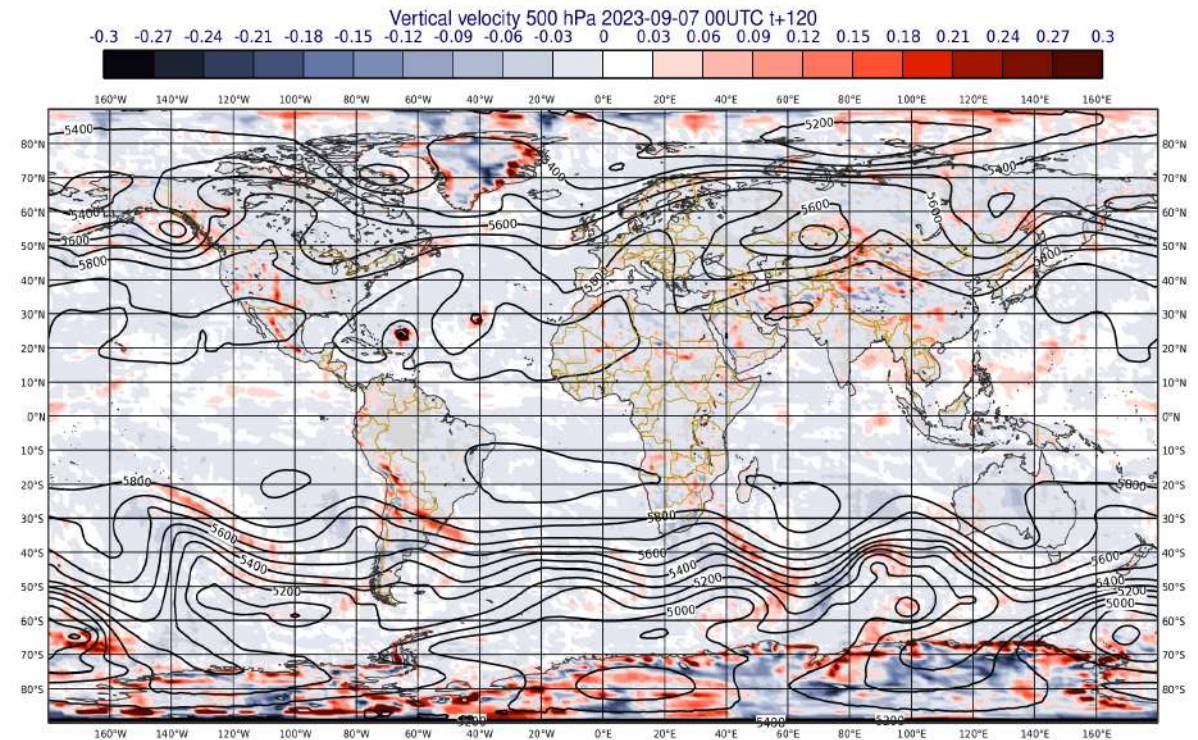
Evolution of mean abs value
of fcst vert velocity
IFS, ERA5
Pangu-W
GraphCast
FourCastNet

ML models dynamics

ERA5 fcst vert. vel.
2023-09-07 00UTC t+120h

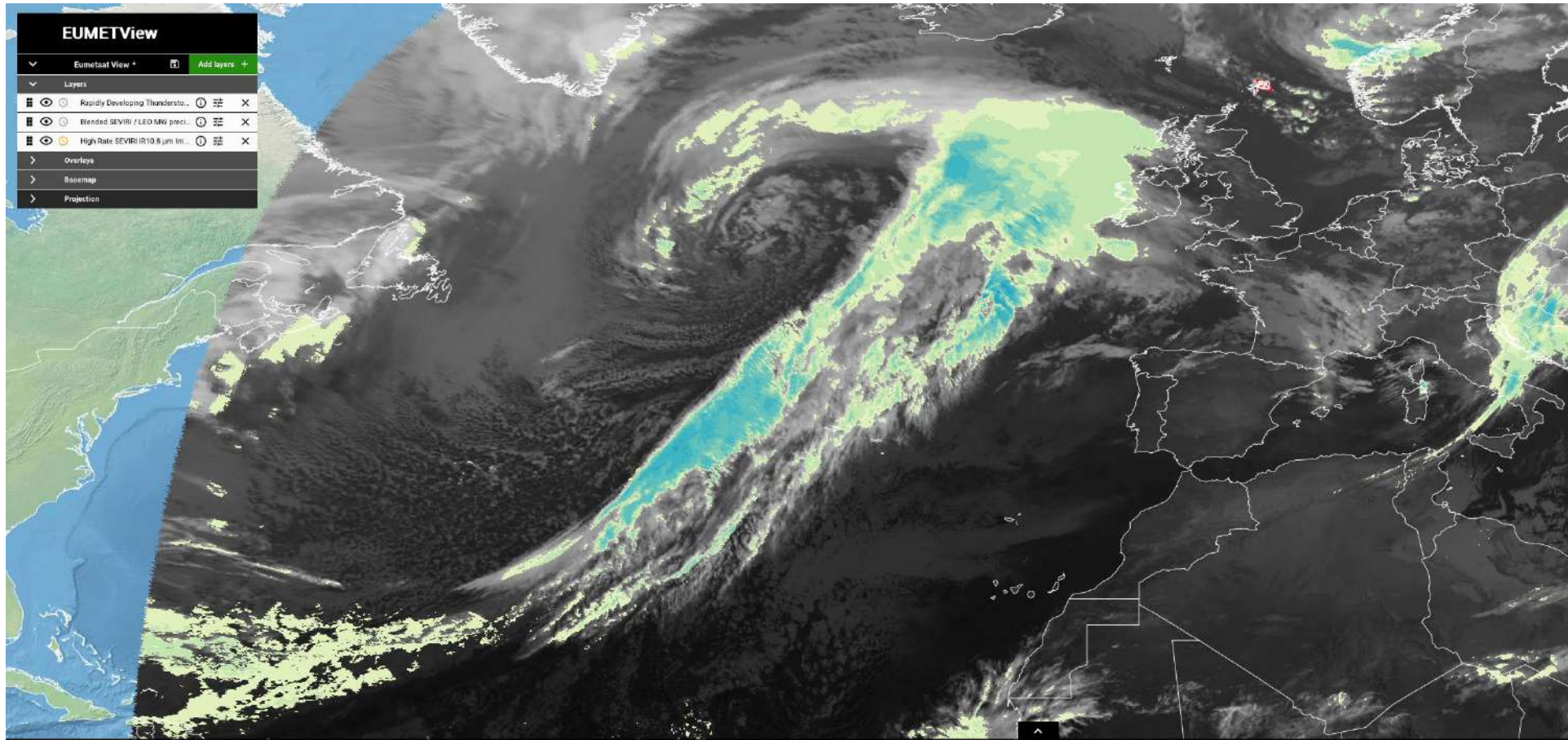


Pangu-Weather fcst vert. vel.
2023-09-07 00UTC t+120h



ML models dynamics

Extra-tropical cyclonic development

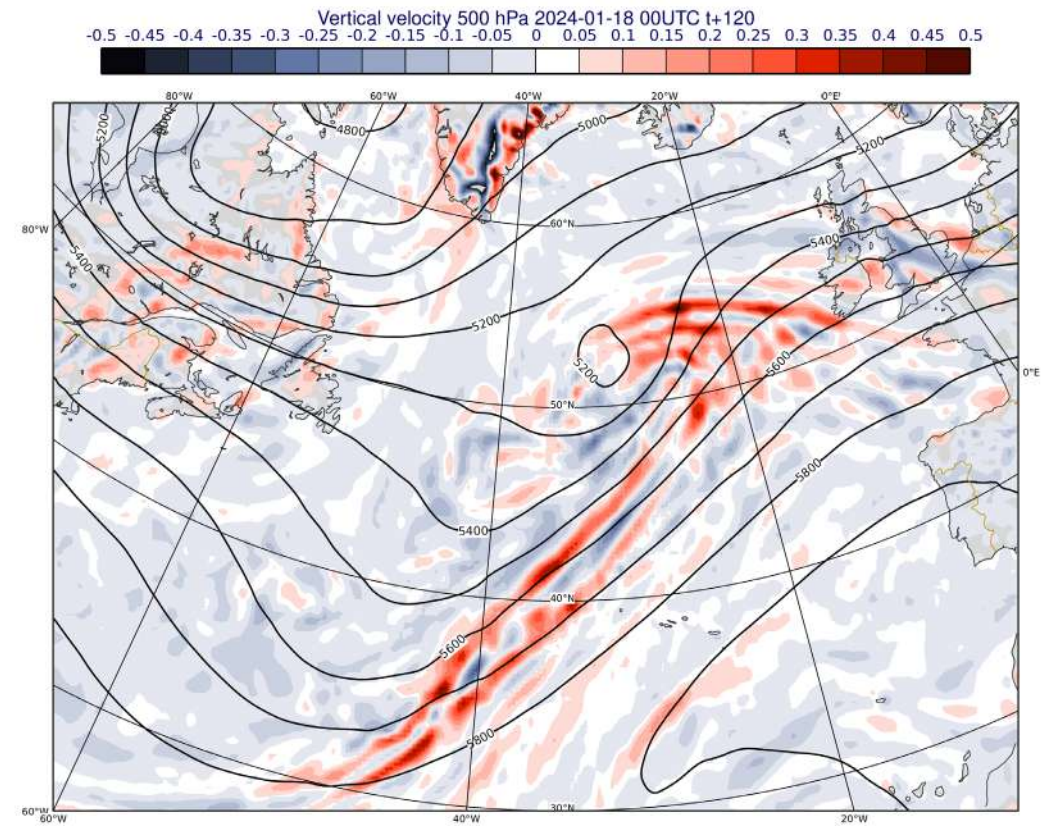
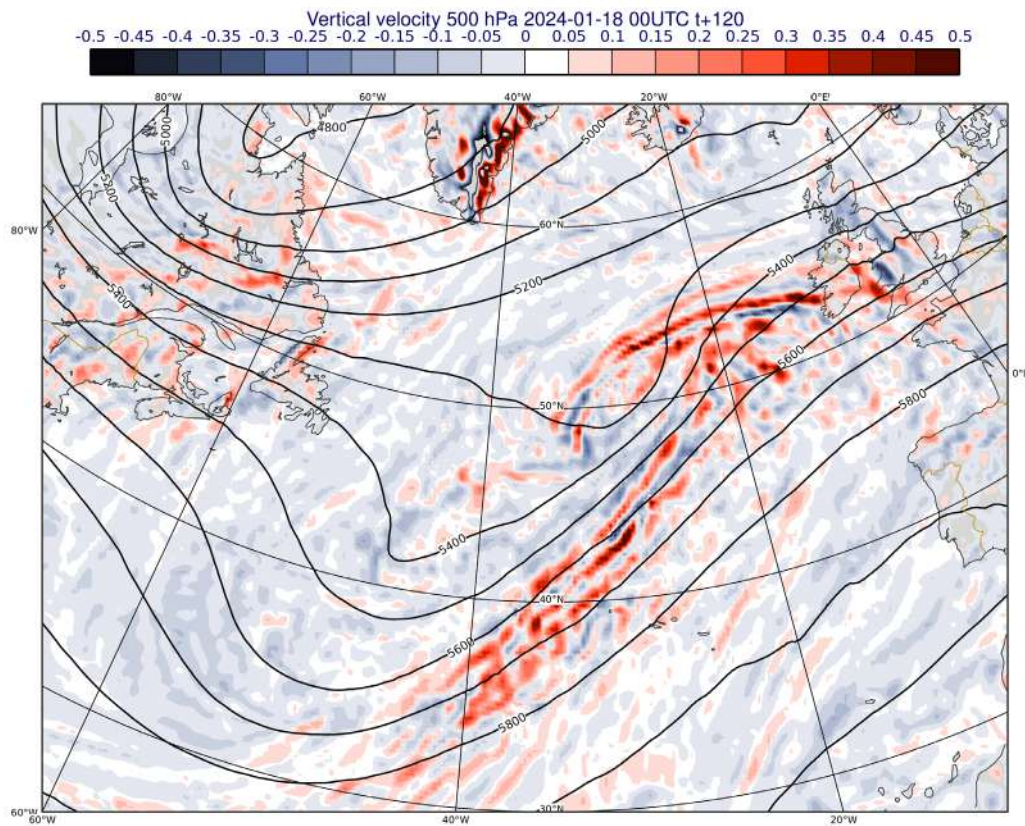


ML models dynamics

Extra-tropical cyclonic development

ECMWF IFS

ERA5

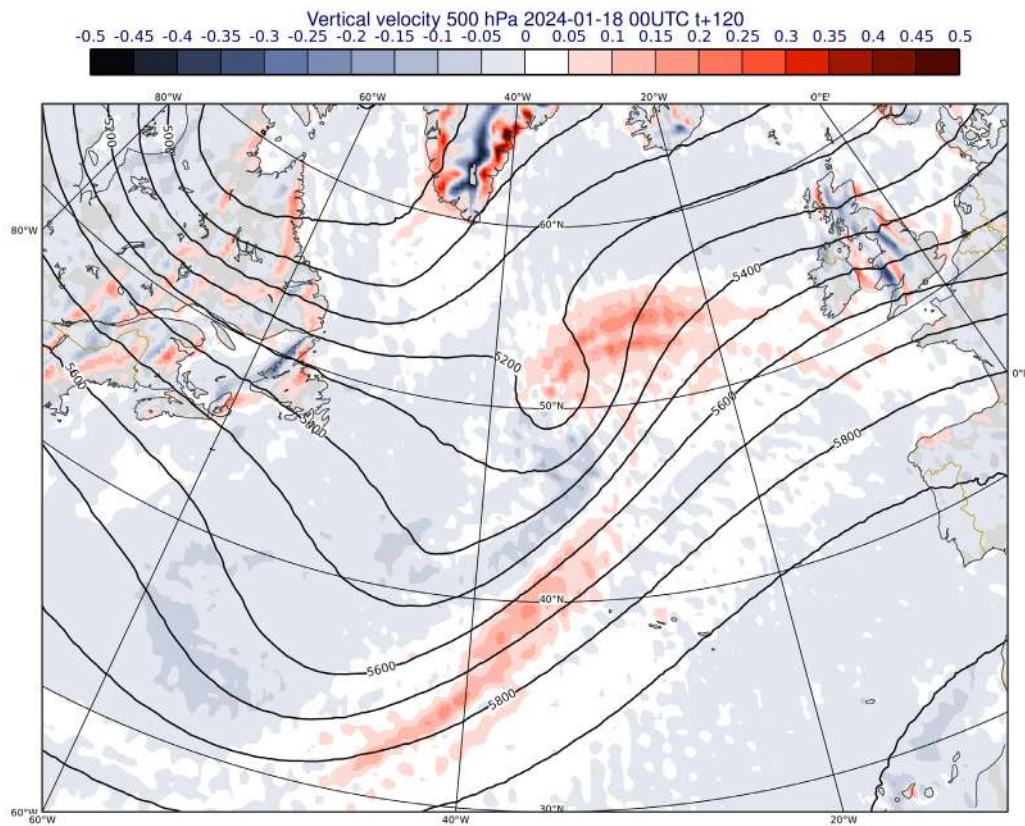


Z500 + vert. vel. (m/s, shaded)
2024-01-18 00UTC t+120h

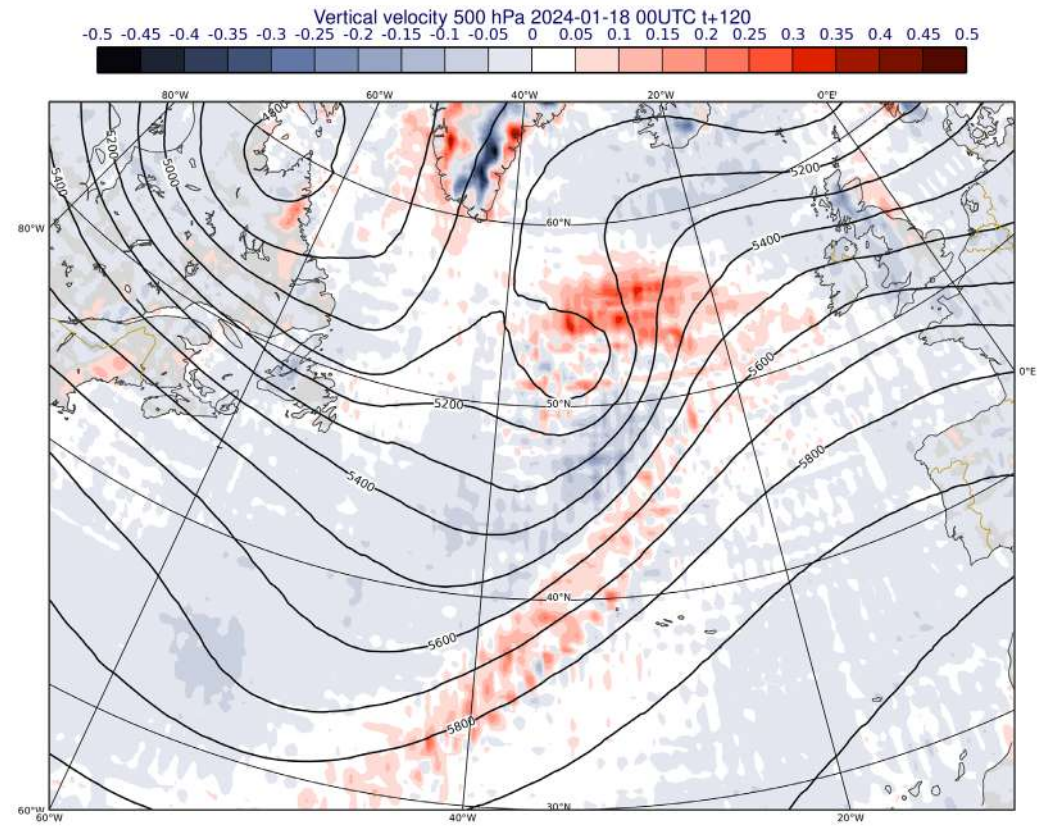
ML models dynamics

Extra-tropical cyclonic development

GraphCast



PANGU

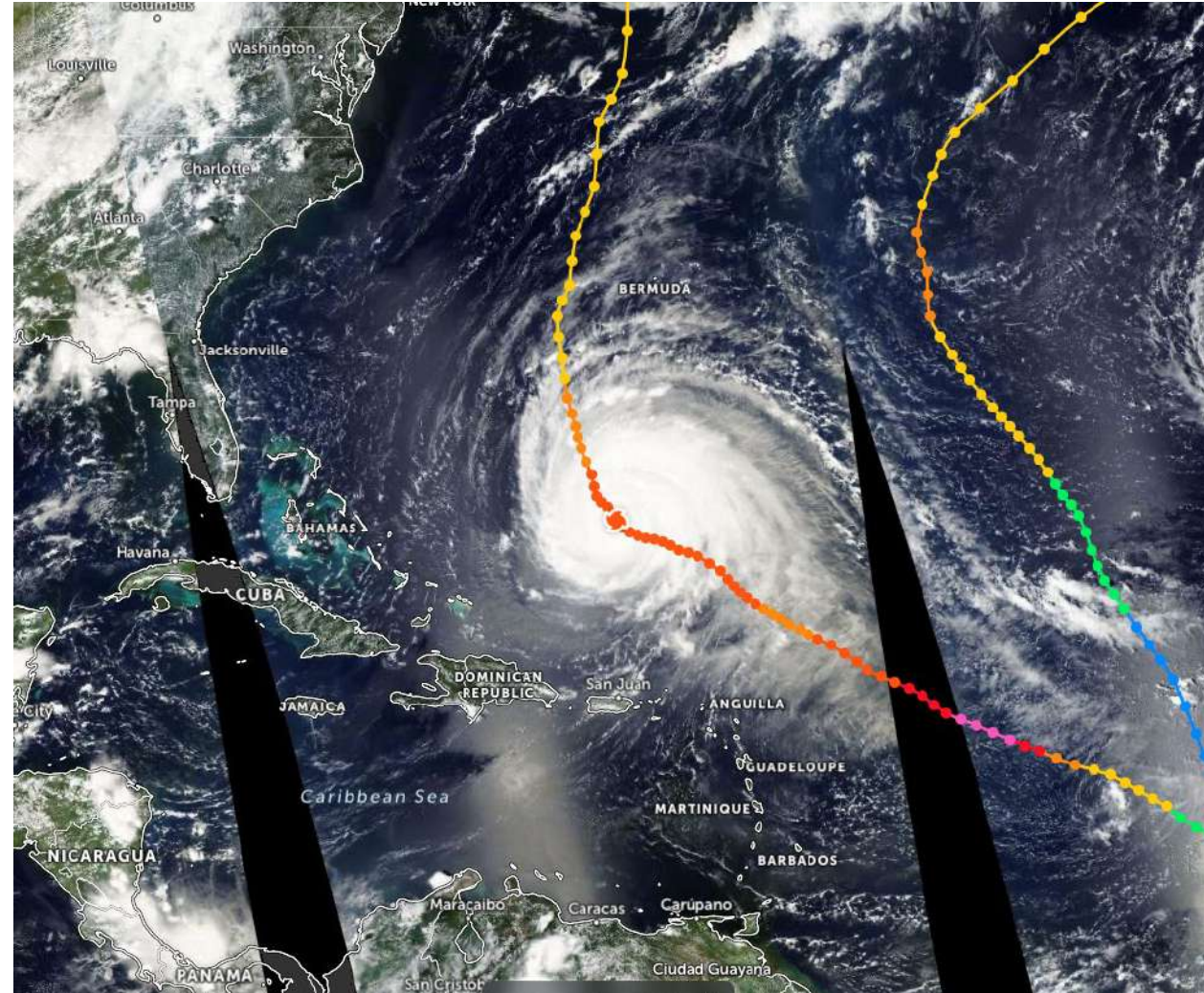


Z500 + vert. vel. (m/s, shaded)
2024-01-18 00UTC t+120h

ML models dynamics

Hurricane Lee, 12 September 2023
01UTC

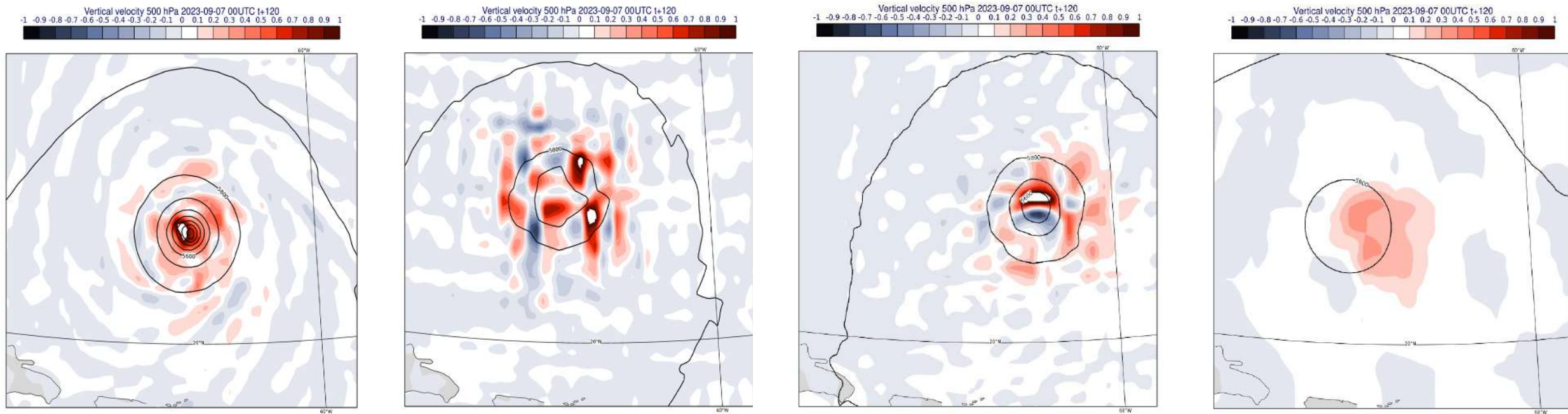
Strongest TC of the 2023 Atlantic
Season, Category 3 at the time



<https://zoom.earth/storms/lee-2023/#map=satellite-hd>

ML models dynamics

Z500 + vert. vel. (m/s, shaded)
2023-09-07 00UTC t+120h



ML models dynamics

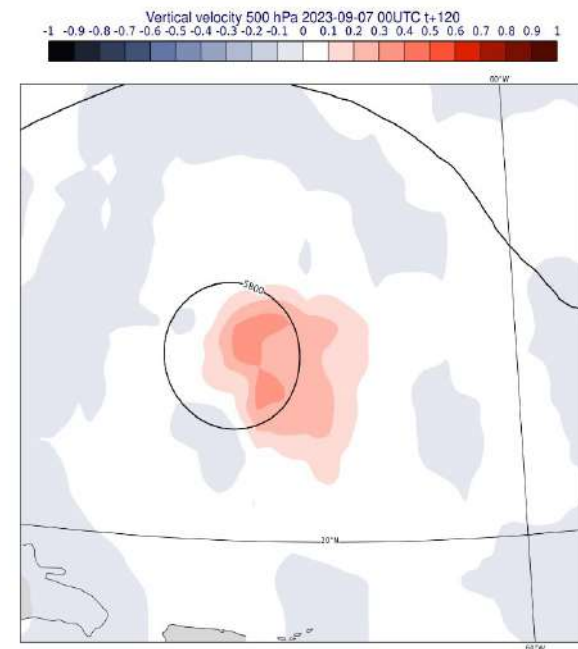
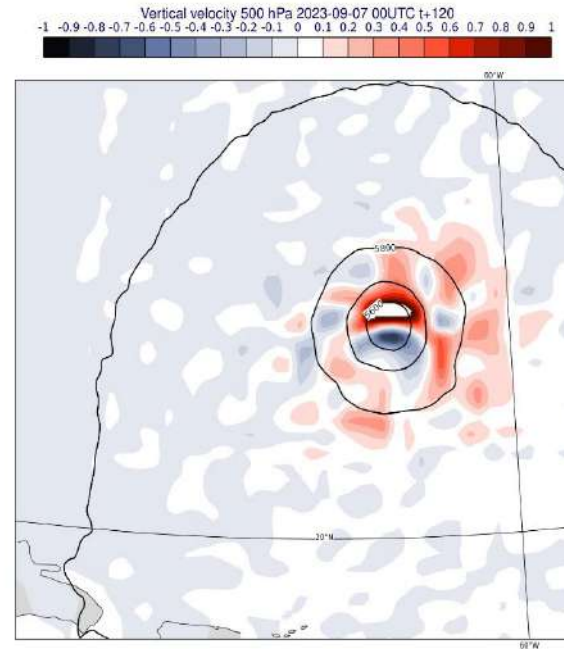
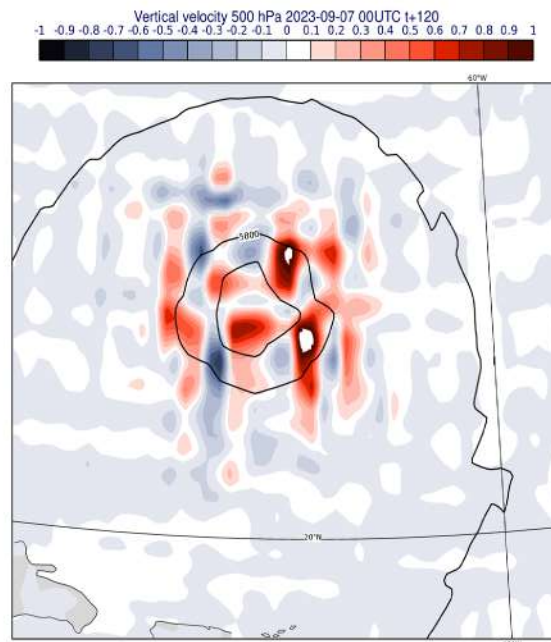
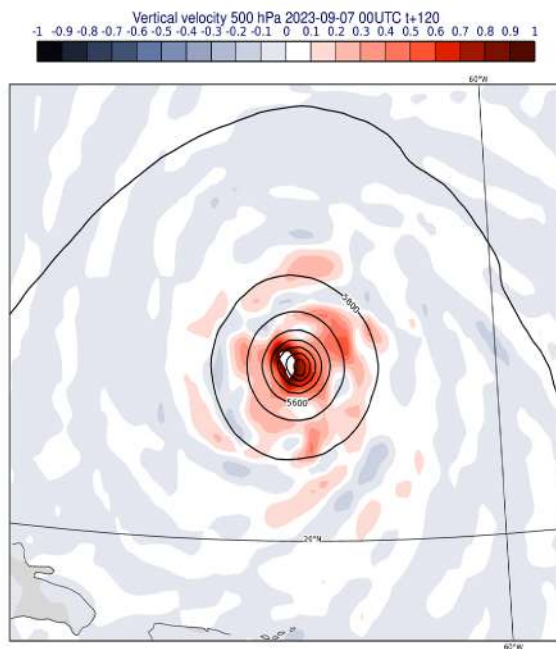
Z500 + vert. vel. (m/s, shaded)
2023-09-07 00UTC t+120h

IFS

Pangu-Weather

GraphCast

FourCastNet



Discussion (1)

1. The “**blurriness**” of predictions is a feature of MLWP models’ forecasts
2. This is because the output of any statistical regression (linear or nonlinear, eg NN) derived using a **L2/L1 norm*** converges to the conditional **mean/median** of the target data in the training distribution in the limit of large sample size (Bishop, 1995):

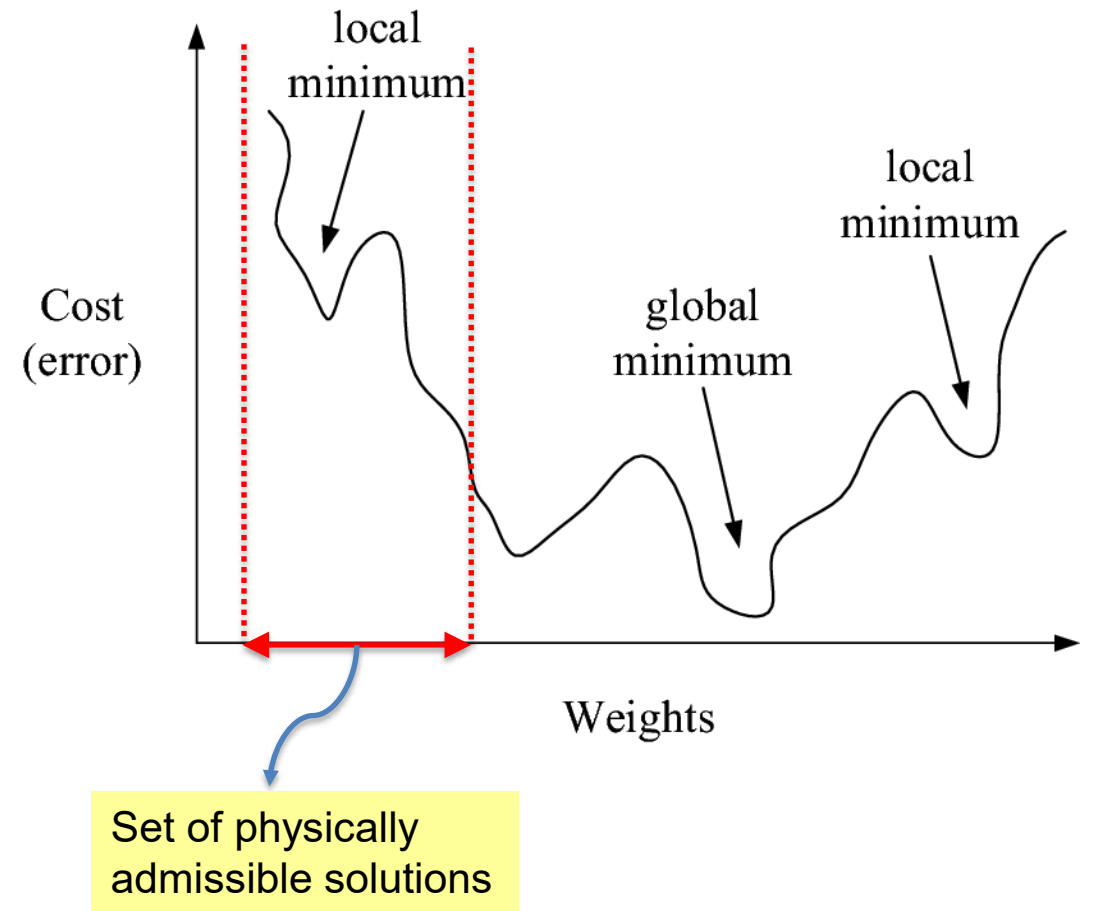
$$ML(\mathbf{x}, \mathbf{w}) \xrightarrow{N \rightarrow \infty} \langle \mathbf{y}_k | \mathbf{x} \rangle$$

3. One can try different loss functions and/or enforce constraints in spectral space, but RMSE/ACC types of forecast skill scores will likely suffer

*Mean Squared Error / Mean Absolute Error

Discussion (2)

1. Another issue is that **different atmospheric motions/scales are not equally predictable** (eg, divergence is less predictable than vorticity, mesoscales are less predictable than synoptic scales, etc.)
2. Current ML forecast models are trained with an **unconstrained minimisation** of a L2/L1 loss function: this will give you the best RMSE/ACC but not a physical state
3. Constraints from **conservation laws/physical balances** will need to be enforced in the loss function to guarantee the ML produces a physically consistent output
4. But enforcing conservation laws/physical balances will have an impact on RMSE/ACC forecast skill scores



Discussion (3)

1. ML models are effective and uber-efficient **forecast tools**. We are gradually understanding where their skill is and what drives it
2. ML are **not physical emulators**, by construction!
3. There are ways to enforce physical realism in the ML solution, at the cost of perceived forecast skill: **physics-constrained ML**
4. Alternatively, one can take the road of embedding ML tools in the standard DA/NWP machinery: **Hybrid ML-DA/Model**
5. To be continued...