Evaluation of ML forecasts

Zied Ben-Bouallègue

Aug. 2025, Bonn

introduction

Computer simulations

Traditionally:
- Verification of forecasts
- Diagnostic of models

while now:
- Verification of forecasts
- Diagnostic of models
- Falsification of ML models

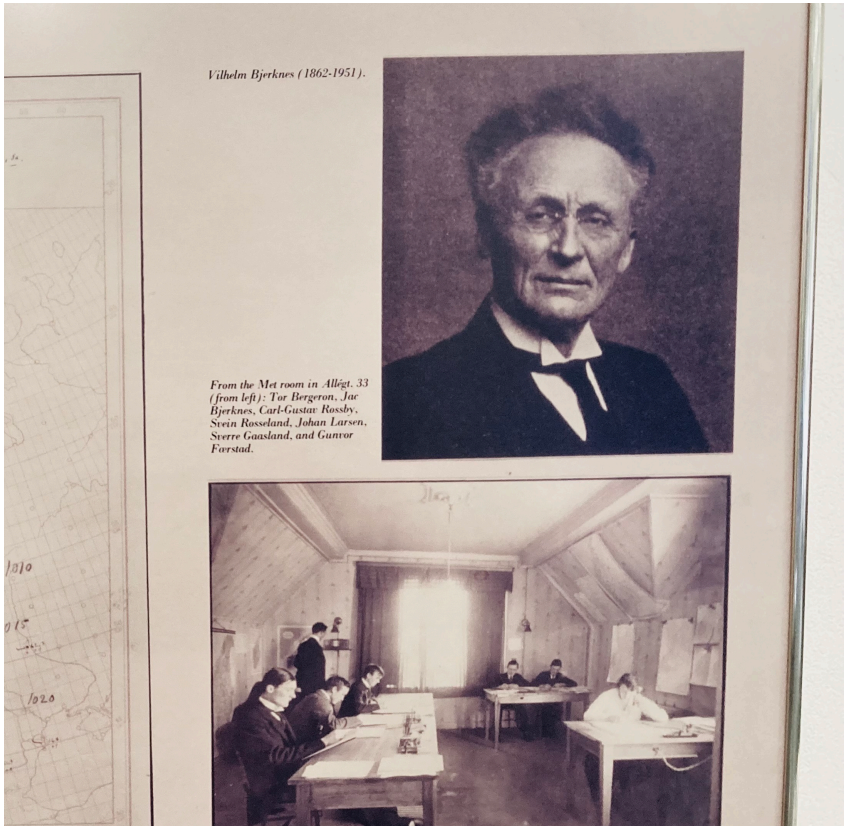
introduction

Computer simulations

Traditionally:
- Verification of forecasts
- Diagnostic of models

while now:
- Verification of forecasts
- Diagnostic of models
- Falsification of ML models

in the lobby of ECMWF headquarters



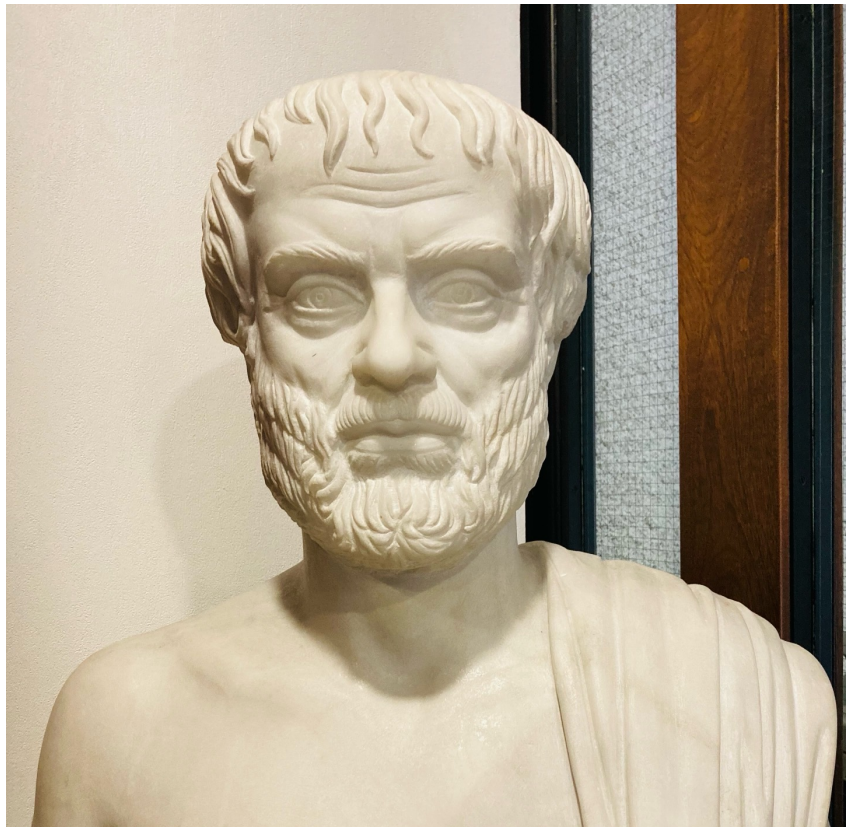
introduction

Computer simulations

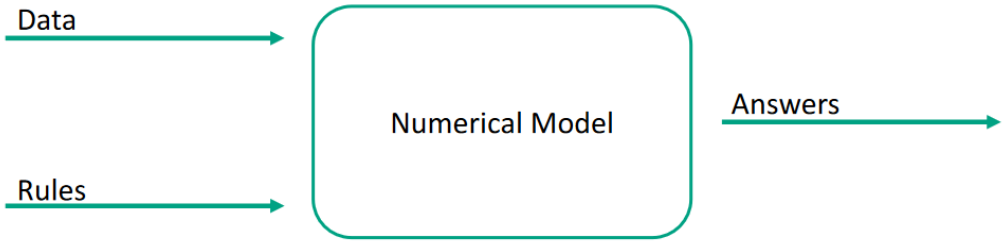
Traditionally:
- Verification of forecasts
- Diagnostic of models

while now:
- Verification of forecasts
- Diagnostic of models
- Falsification of ML models

in the lobby of ECMWF headquarters



Deduction



Induction

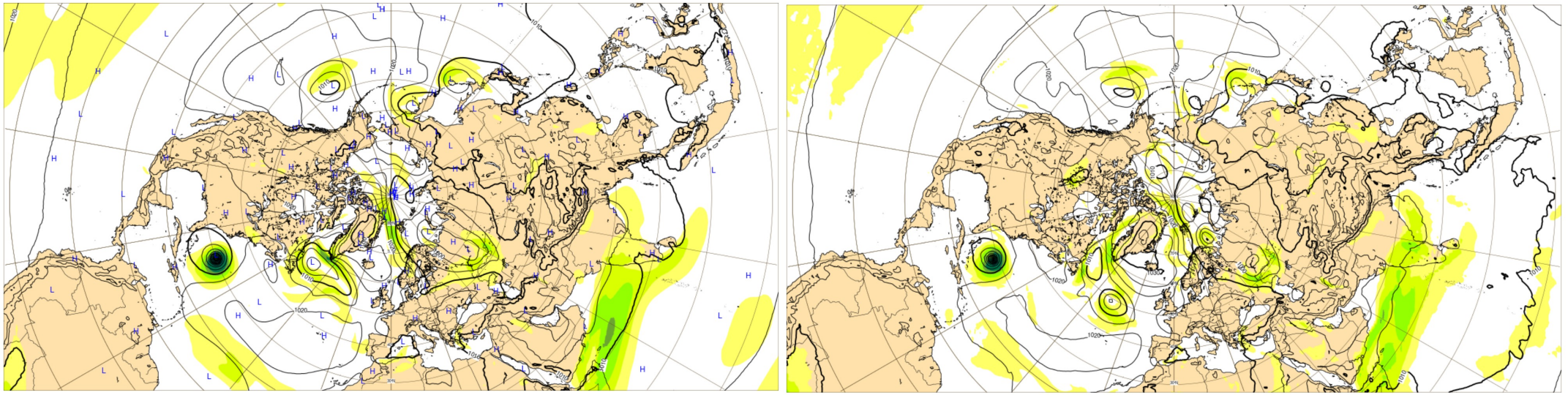
Day +6 forecasts

Mean sea level pressure and 850 hPa wind speed

.

Base time: Thu 14 Aug 2025 06 UTC Valid time: Wed 20 Aug 2025 06 UTC (+144h) Area : North Pole

Base time: Thu 14 Aug 2025 00 UTC Valid time: Wed 20 Aug 2025 00 UTC (+144h) Area : North Pole



850 hPa wind speed (m/s)

10 15 20 25 30 40 50 60 80 100 (m/s)

Mean sea level pressure (hPa)

ECMWF

What is a realistic forecast*?

Definition

Realism: the quality [..] of representing […] a thing in a way that is accurate and true to life.

Source: Oxford Languages

* inspired by “What is a good forecast? An Essay on the Nature of Goodness in Weather Forecasting”, Murphy, 1993.

What is a realistic forecast*?

Let’s discuss 3 types of realism:

- Type 1 or functional realism
- Type 2 or structural realism
- Type 3 or physical realism

* inspired by “What is a good forecast? An Essay on the Nature of Goodness in Weather Forecasting”, Murphy, 1993.

Type 1 of realism (functional)

Q: is the forecast skilful? How to measure performance?

T: forecast verification metrics, scoring functions,…

Example 1 (Jin et al, 2024)

- “Comparing different models **on "real" observations** from near-surface direct measurements, as a starting point, underscores the growing emphasis on the **practical value** of these models to people’s daily lives. “
- They “propose a few **possible tasks with specific scoring methods** that can be ranked across research” models.

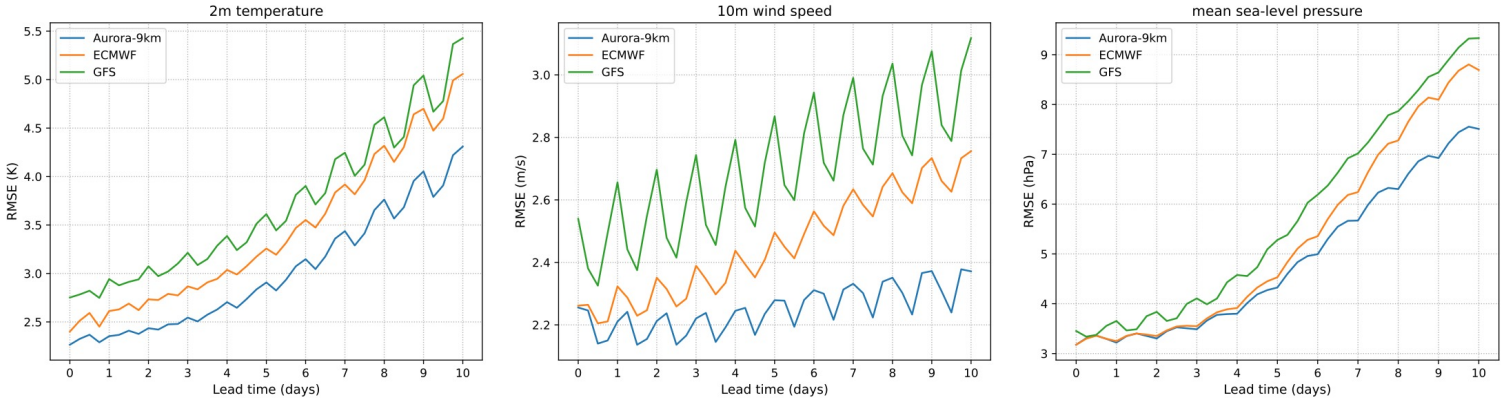


Figure 17: As in Fig. 14, except the forecasts are evaluated against the WeatherReal-Synoptic observations, and omitting the cloud variable.

Example 2 (Price et al, 2024)

- "**Extreme** heat, cold, wind, and other severe surface weather **pose hazards to lives and property**, **but can be anticipated** "
- "the value of a forecast to **a particular decision-maker** depends on the situation's specific cost/loss ratio"

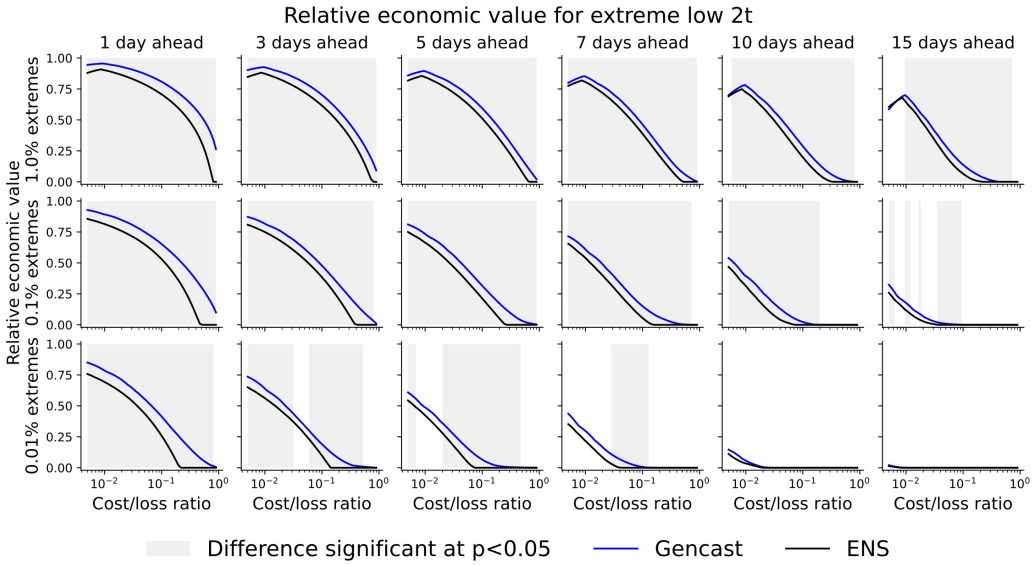


Figure F14 | Relative economic value plots for extreme low temperatures. Regions for which GenCast is better then ENS with statistical significance are shaded in grey.

the concept of elicitation

(a forecast, a scoring function)

The concepts of *forecast elicitation* and *consistency with a scoring rule* answer the following:

- what is the optimal consensus forecast for a given measure of accuracy?
- what is the appropriate measure accuracy to assess a given consensus forecast ?

Example: the **ensemble mean** is consistent with the root mean squared error (**RMSE**). The RMSE is optimised when the ensemble mean is issued.

Type 2 of realism (structural)

Q: is the forecast (statistically) consistent with the observations? How to measure reliability?

T: diagnostic tools, feature-based analysis,…

Example 1 (Bonavita 2024)

- Ml-based forecasts “appear to become increasingly **“blurry” with increasing forecast lead times**”.
- ML models “consistently show **reduced forecast variability** at smaller (sub-synoptic, mesoscale) spatial scales”.

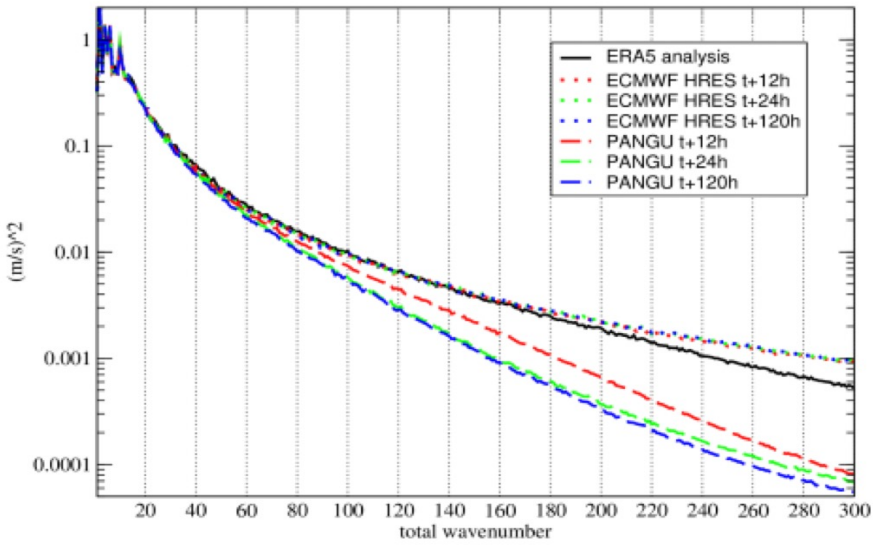


Figure 1: *Power spectral density as a function of total wavenumber of ERA5 analysis (continuous line), ECMWF IFS operational forecasts (dotted lines) and Pangu-Weather forecasts (dashed lines) at lead times t+12h, t+24h and t+120h.*

Example 2 (Pasche et al, 2024)

- Case study: 2023 South Asian **humid heatwave**
- "Model accuracy might degrade due to mis-represented **dependencies between variables**".
- **"Non-linear combinations** of predicted output variables (e.g., wind chill, see Equation 1) have the potential **to reveal weaknesses of ML models**"

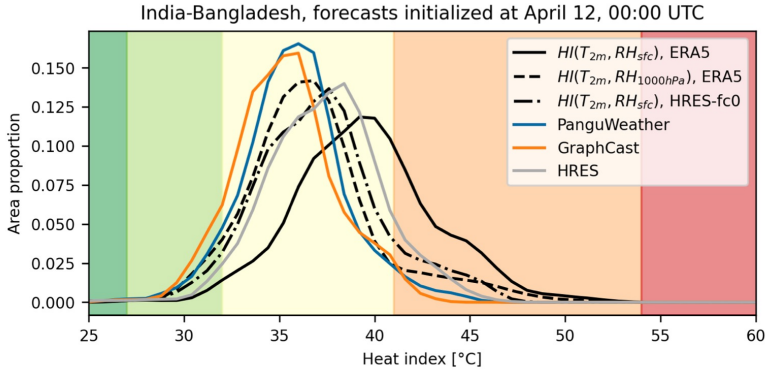


Figure 5: Proportion of area in study region with given mean daily maximum heat index during April 17–20, 2023, computed using area-weighted kernel density estimation. Shaded areas in the background indicate threat levels (see Section BB.3). Dark green: low risk, light green: caution, yellow: extreme caution, orange: danger, red: extreme danger. Compared are distributions resulting from forecasts initialized 6 days prior to the start of the event and different ground truths: HI computed from RH_{sfc} using ERA5 and HRES-fc0, and the version of HI, in which we substitute $RH_{1000\,\mathrm{hPa}}$ for RH_{sfc}

the concept of reliability

- reliability implies that the forecast is statistically consistent with the observation
- a forecast is reliable when drawn from the same (conditional) probability distribution as the observation
- reliability is measured in terms of expectations

Type 3 of realism (physical)

Q: is the forecast trustworthy? How to measure (physical) consistency of the forecast?

T: physical test, case-study analysis, …

Example 1 (Hakim and Masana, 2024)

- “Physical tests that examine the evolution of **spatially localized disturbances** are particularly effective in analysing model physics, since the propagation of signals away from these disturbances **is constrained by dynamics**”.
- “the model encodes **realistic physics** in all experiments”.

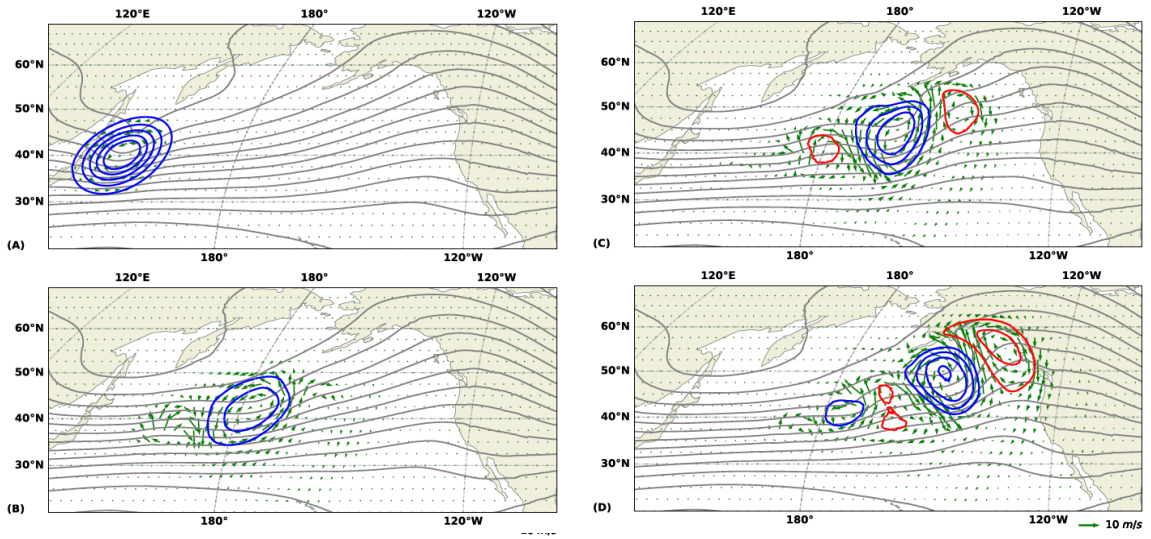
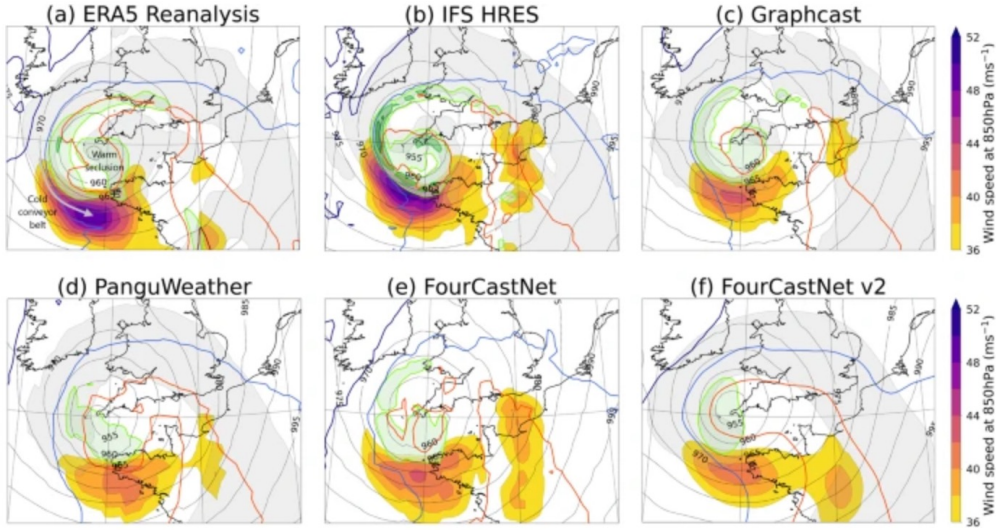


FIG. 3. Solution at 500 hPa for a localized disturbance on the DJF atmosphere. The full geopotential height is shown by gray lines every 60 m, and anomalies from the DJF average are shown by red (positive) and blue (negative) lines every 20 m; the zero contour is suppressed. Green arrows show the anomalous vector wind. Solutions are shown at (a) 0 (the specified initial condition); (b) 2; (c) 3; and (d) 4 days.

Example 2 (Charlton–Perez et al, 2024)

- A case study “to compare the ability of the models to capture the **detailed physical structure of the storm**”.
- “Only some machine learning models resolve the **warm core seclusion** and none of the machine learning models capture the **sharp bent-back warm frontal gradient**.”

Fig. 6: As Fig. 5 but for the dynamical structure of Storm Ciarán at 00 UTC on 2 November 2023.



Contours of wind speed at 250 hPa and some of the contours of wet-bulb potential temperature at 850 hPa are not present as the associated values are not reached in the maps shown.

the concept of falsification*

- Falsification or refutation of theories
- Drives scientific progress

  "Scientific progress consist in moving forward theories that tell us more and more [...] not the accumulation of observations but the overthrown of less good theories".

- Here, falsification of ML models with respect to a trusted source of information: are ML predictions and physics theories compatible?

* See, for example, "Unended Quest", Karl Popper, 1974.

Relationship between type 1 and type 2

Sharpness principle*

Maximizing the sharpness of the predictive distributions subject to calibration = aiming for an ideal forecasts.

Activity-Accuracy trade-off

Improve structural realism or functional realism? Might be contradictory goals in a non-probabilistic setting.

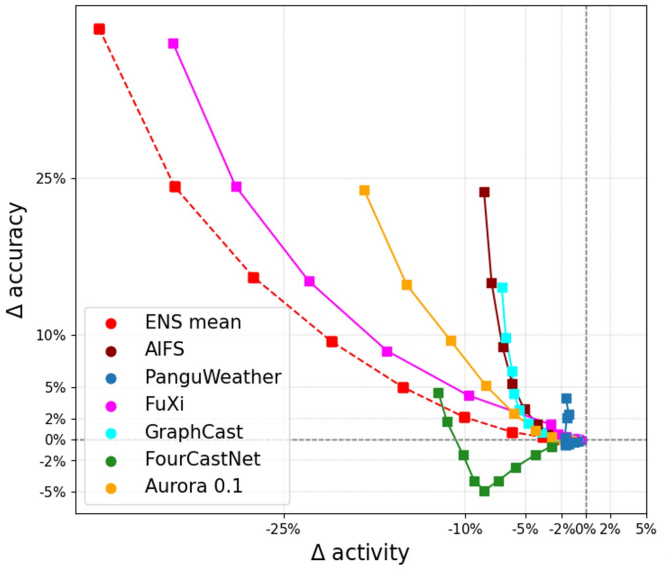


Figure 2: Forecast accuracy-activity trade-off for geopotential hight at 500 hPa over the Northern Extratropics. The plot shows the forecast relative accuracy with respect to the IFS control forecast versus the forecast activity relative with respect to the IFS. Results for forecasts at lead time day 1 (dot) up to day 10 (squares) are plotted. Accuracy is here measured with the anomaly correlation coefficient.

* see “Probabilistic forecasts, calibration and sharpness”, Geniting et al, 2007.

Relationship with type 3 ?

Type 1 and Type 3

- does improved physical realism imply improved functional realism? (Better physics -> better scores?)
- vice versa?

Type 2 and Type 3

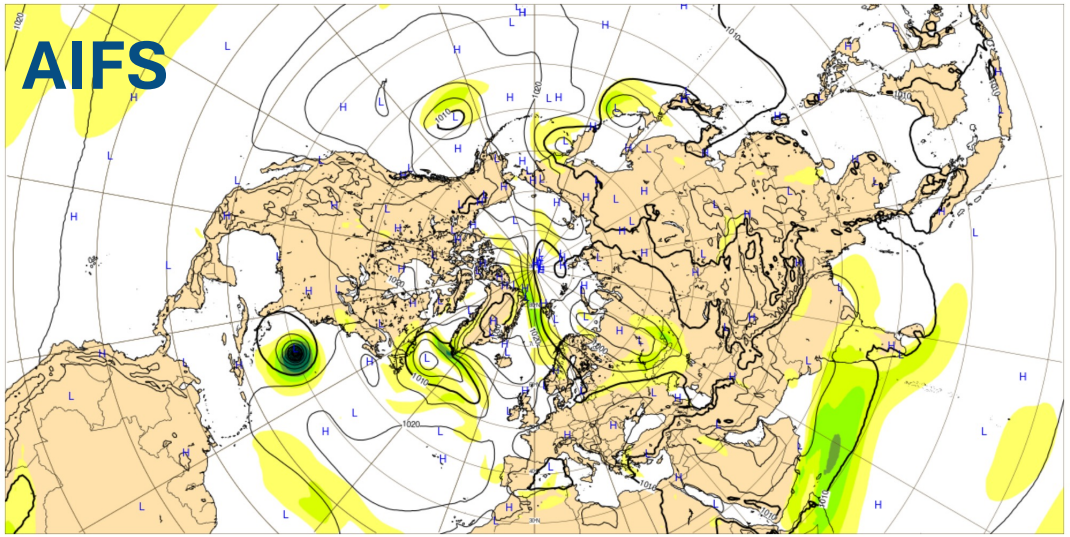
- does improved structural realism leads to improved physical realism? (Better structures -> better physics?)
- vice versa?

Day +6 forecasts

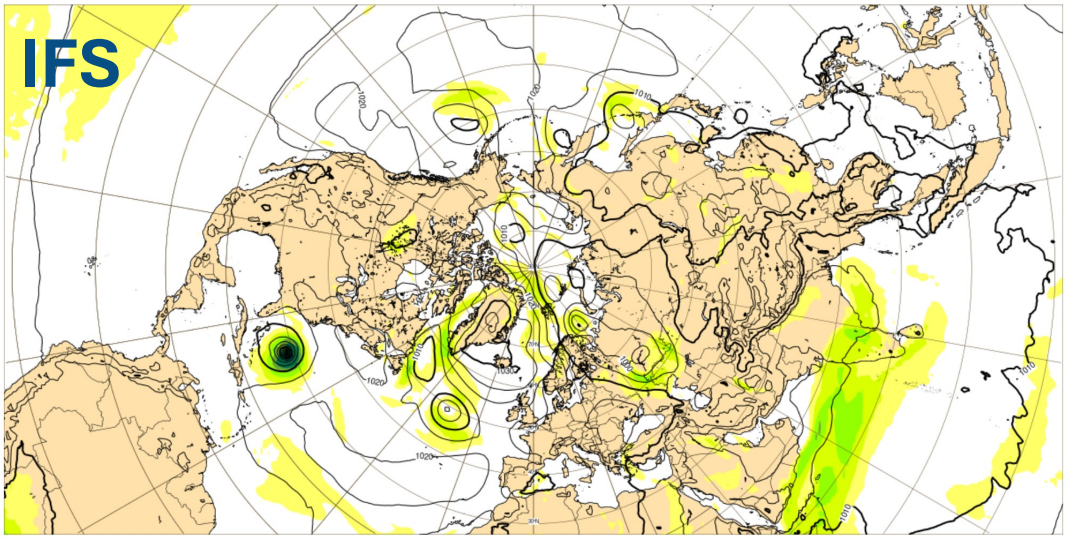
Mean sea level pressure and 850 hPa wind speed

.

Base time: Thu 14 Aug 2025 06 UTC Valid time: Wed 20 Aug 2025 06 UTC (+144h) Area : North Pole



Base time: Thu 14 Aug 2025 00 UTC Valid time: Wed 20 Aug 2025 00 UTC (+144h) Area : North Pole



850 hPa wind speed (m/s)

10 15 20 25 30 40 50 60 80 100 (m/s)

Mean sea level pressure (hPa)

ECMWF

summary

is the forecast fit-for-purpose?

Evaluation of ML forecasts in 3 steps:

1. Verification of the forecasts (assessing the **functional realism**):
Q: is the forecast skilful?
T: measure forecast performance with scores

2. Diagnostic of models (checking the **structural realism)**:
Q: is the forecast consistent with the observations?
T: assess forecast characteristics

3. Falsification of ML models (testing the **physical realism)**:
Q: is the model trustworthy?
T: test forecast adequacy to our knowledge

References

- Bonavita 2024, On Some Limitations of Current Machine Learning Weather Prediction Models,
  - https://doi.org/10.1029/2023GL107377
- Charlton-Perez *et al 2024* Do AI models produce better weather forecasts than physics-based models?
  - https://www.nature.com/articles/s41612-024-00638-w
- Hakim, G. J., and S. Masanam, 2024: Dynamical Tests of a Deep Learning Weather Prediction Model.,
  - https://doi.org/10.1175/AIES-D-23-0090.1
- Ji et al 2024, WeatherReal: A Benchmark Based on In-Situ Observations for Evaluating Weather Models,
  - https://arxiv.org/abs/2409.09371
- Pasche et al 2024, Validating Deep-Learning Weather Forecast Models on Recent High-Impact Extreme Events,
  - https://arxiv.org/abs/2404.17652
- Price ate al 2024, GenCast: Diffusion-based ensemble forecasting for medium-range weather,
  - https://arxiv.org/abs/2312.15796

Resources

- Forecast Verification methods (JWGFVR):
  - https://cawcr.gov.au/projects/verification/
- AIFS blog posts (ECMWF):
  - https://www.ecmwf.int/en/about/media-centre/aifs-blog