Swiss National Science Foundation

UNIVERSITÄT BERN
OESCHGER CENTRE CLIMATE CHANGE RESEARCH

Do AI models predict storm hazards as accurately as physics-based models?

A case study of the February 2020 European storm series

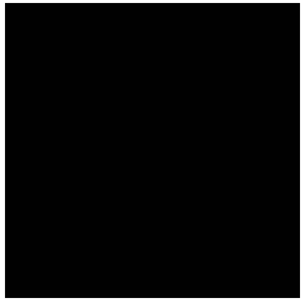
Hilla Afargan-Gerstman

Oeschger Centre for Climate Change Research
University of Bern

With collaborations: Rachel Wu (Columbia University, ETH), Alice Ferrini (ETH), Daniela Domeisen (UNIL, ETH), Fabio Flütsch (ETH).

Motivation: AI weather models are accurate, but can we trust them?

- Data-driven models beat the ECMWF IFS on many “classical” metrics (e.g., RMSE, ACC).
- AI weather models learn from ERA5, do not solve the equations of the atmospheric flow.
- AI weather prediction models systematically underestimate extreme events.
- Specifically, AI weather prediction models **systematically underestimate** the intensity of mid-latitude cyclones (Dacre et al., 2026).



Motivation: AI weather models are accurate, but can we trust them?

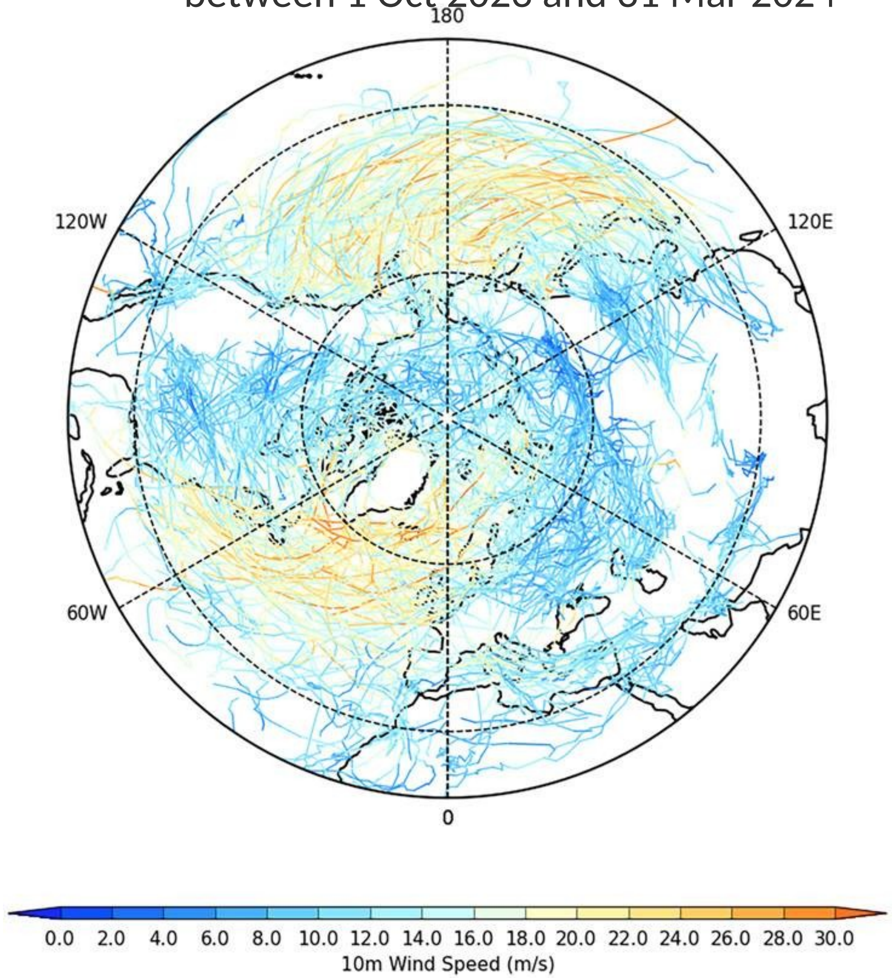
- Data-driven models beat the ECMWF IFS on many “classical” metrics (e.g., RMSE, ACC).
- AI weather models learn from ERA5, do not solve the equations of the atmospheric flow.
- AI weather prediction models systematically underestimate extreme events.
- Specifically, AI weather prediction models **systematically underestimate** the intensity of mid-latitude cyclones (Dacre et al., 2026).

*AI weather predictions are potentially better then traditional Numerical Weather Prediction (NWP) methods, but **do we trust their them**?*

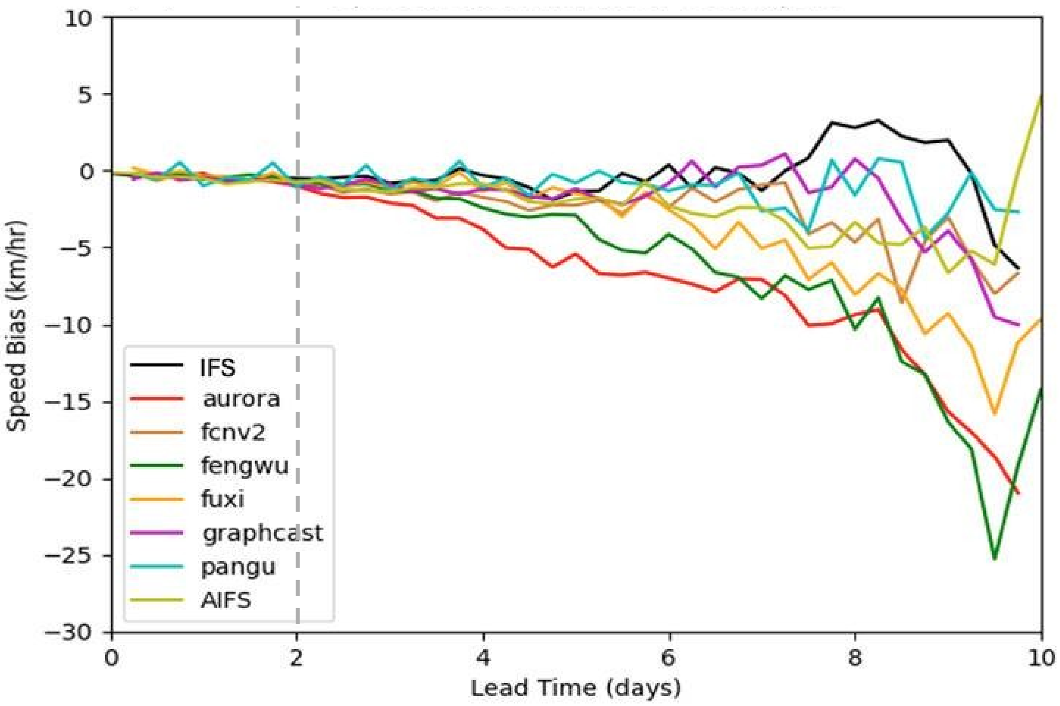
Is it good enough for impact-related decision-makers?

AI weather prediction models **systematically underestimate** midlatitude cyclone intensity

Cyclone tracks identified in the **ECMWF IFS** analysis between 1 Oct 2023 and 31 Mar 2024



10-m wind speed intensity error at cyclone-maximum



…Surface winds underestimation develops after 2-3 days

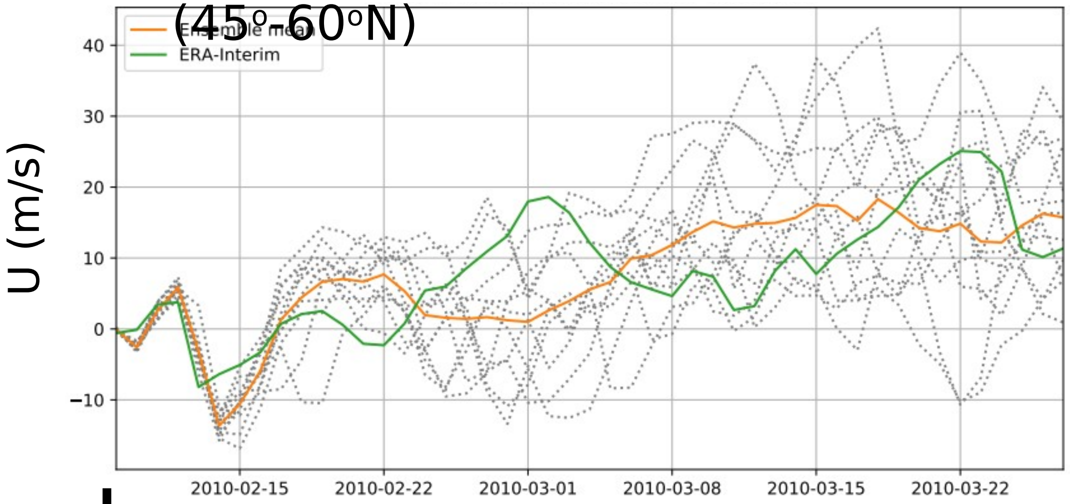
Figures: Dacre et al., 2026, *BAMS*.

Methodology

- **Reanalysis:** ERA-5, for the years 2000-2023
- **NWP Forecasts:** ECMWF IFS reforecasts, S2S Prediction Project database (Vitart et al., 2017)

Model details:
Integrated Forecasting System (IFS)
46-day integrations
10 members
Period: 2000-2020

Zonal wind in the North Atlantic (45º-60ºN)



Initialization date

Ensemble mean (red)
Ensemble members (grey)
Reanalysis (green)

Climatology for ECMWF reforecasts, between 2000-2020; Figure from Afargan-Gerstman et al., 2024, WCD

AI weather forecast models

IFS ENS

ECMWF, physics-based

- 0.25° grid
- Probabilistic
- Operational use

GraphCast

DeepMind, Graph Neural Network

- 0.25° grid
- single
- Trained on ERA5 1979–2019

Pangu-Weather

Huawei, 3D Earth-Specific Transformer

- 0.25° grid
- single
- Trained on ERA5 1979–2017

Forecast lead time: 10 days

Data source: WeatherBench 2 (Rasp et al. 2024). *Storm must lie outside the training period.*

Previous work: Tracking extratropical cyclones in NWP forecasts

- **Forecasts:** ECMWF re-forecasts (hindcasts) from the S2S Prediction Project database (Vitart et al., 2017)
- **Extratropical cyclone algorithm:** Wernli and Schwierz (2006), refined by Sprenger et al (2017)

North Atlantic extratropical cyclone frequency, averaged over all re-forecast initializations in DJFM

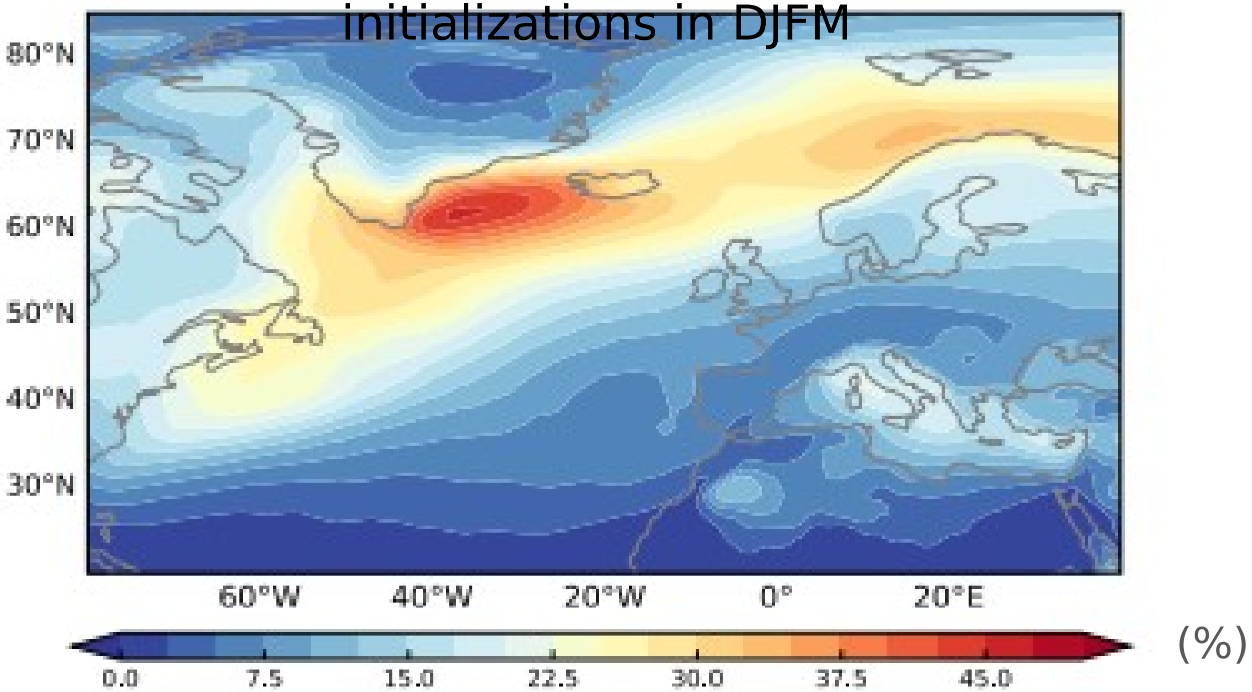
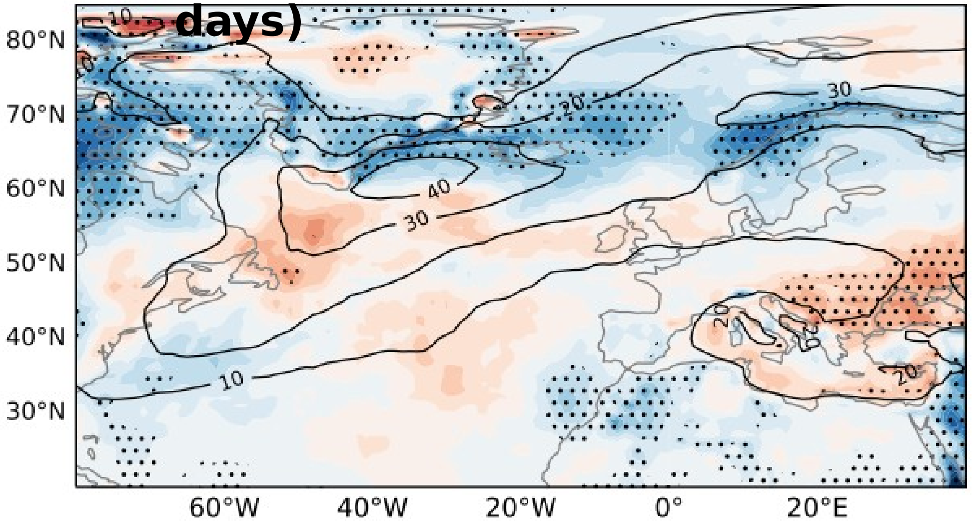


Figure from Afargan-Gerstman et al., 2024, WCD

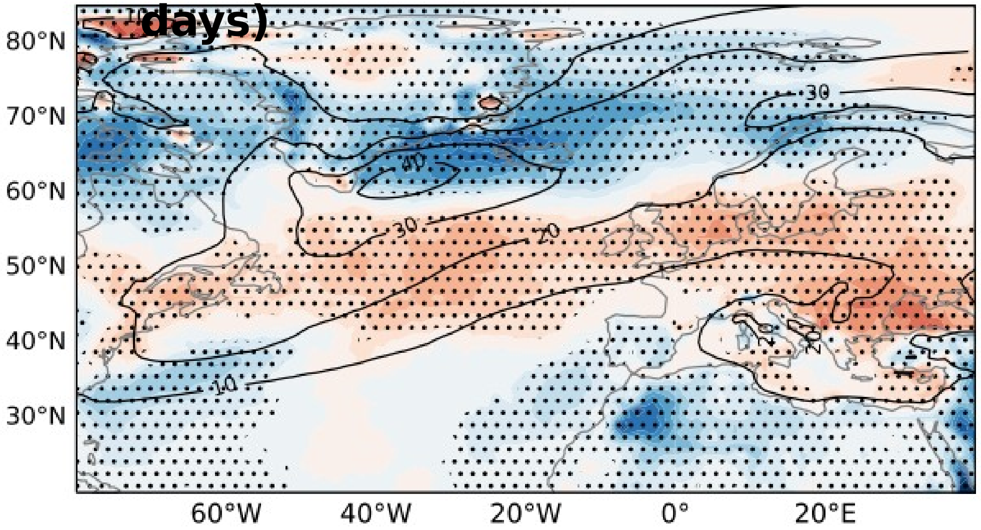
Systematic storm track bias in subseasonal forecasts

Forecast bias (the difference between the forecast and the observation) increases with lead time:

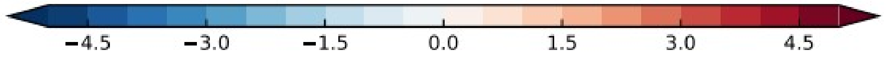
Cyclone frequency bias (after 7 days)



Cyclone frequency bias (after 28 days)



Data: ECMWF forecasts (between 2000 to 2020)
Cyclone frequency anomaly (shading),
Climatology (black contours).



% cyclone frequency

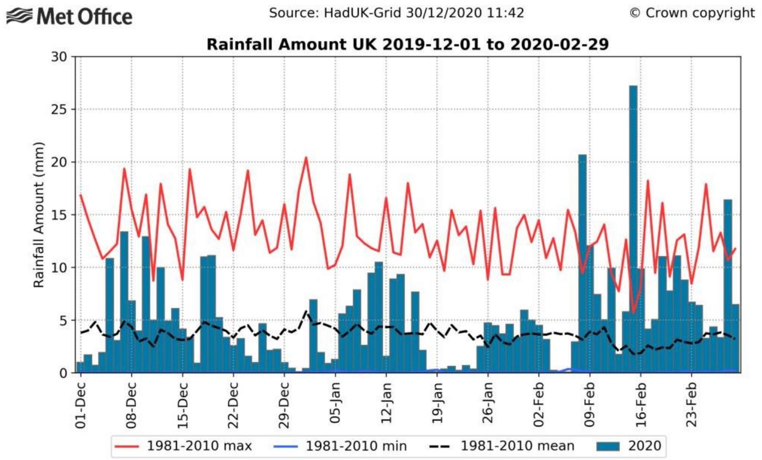
On sub-seasonal timescales, forecasts **overestimate cyclone frequency** over the midlatitude North Atlantic and central Europe, and **underestimate** frequency over Scandinavia.

Figure from Afargan-Gerstman, et al., 2024, *WCD* on time scales beyond one week.

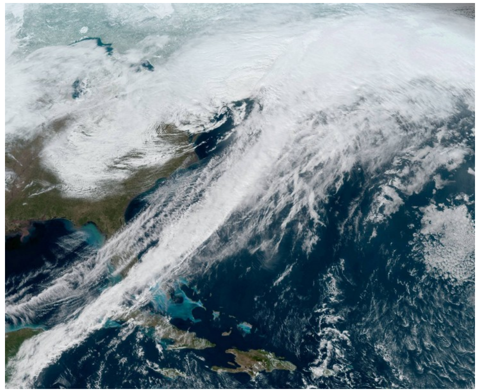
Case study: Forecasting storm series in February 2020

February 2020 was remarkable in many ways:

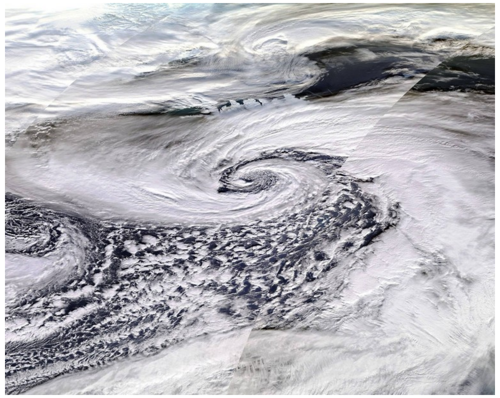
- **The wettest February on record for the UK**
- Strongly positive North Atlantic Oscillation (NAO) and a succession of cyclonic systems:
    - *Ciara* (7-8 February)
    - *Dennis* (15-16 February)
    - *Jorge* (27-28 February)
- Persistent heavy rain that caused severe floods in the UK.



Davies et al., 2021



Storm Ciara on February 7, 2020



Storm Dennis on February 15, 2020

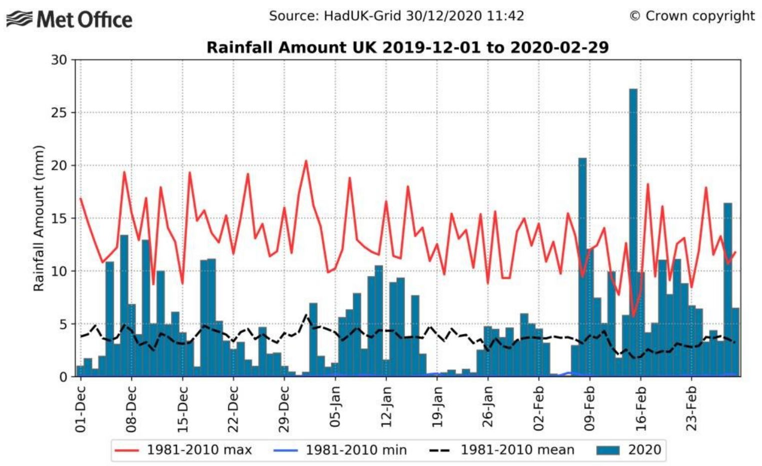
Storm Jorge on February 27, 2020

on time scales beyond one week.

Case study: Forecasting storm series in February 2020

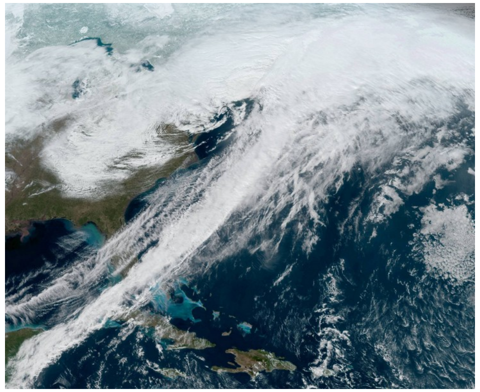
February 2020 was remarkable in many ways:

- **The wettest February on record for the UK**
- Strongly positive North Atlantic Oscillation (NAO) and a succession of cyclonic systems:
    - *Ciara* (7-8 February)
    - *Dennis* (15-16 February)
    - *Jorge* (27-28 February)
- Persistent heavy rain that caused severe floods in the UK.

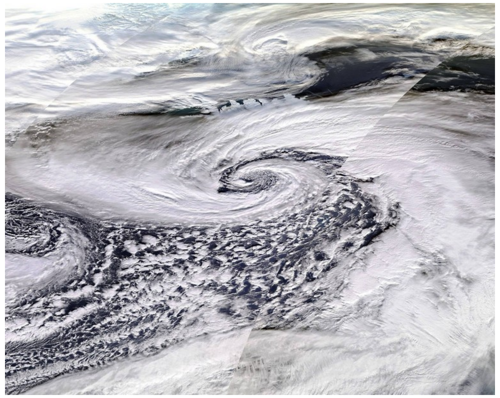


Davies et al., 2021

- **How well can we predict the 2020 storm clustering event?**
- **Does predicting the first storm in a series improve the forecast of successive storms?**



Storm Ciara on February 7, 2020

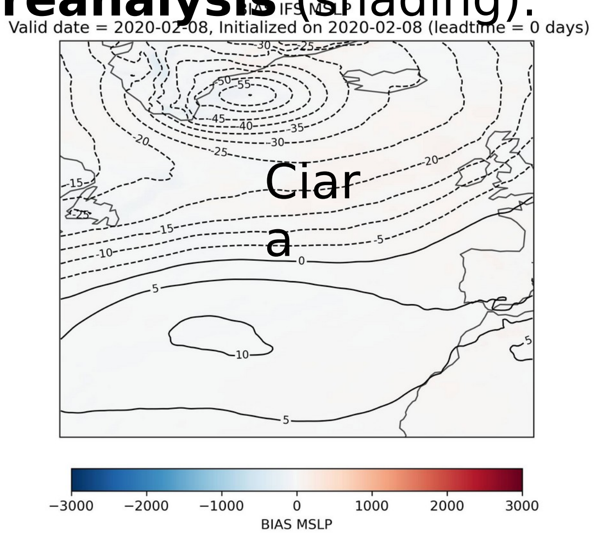


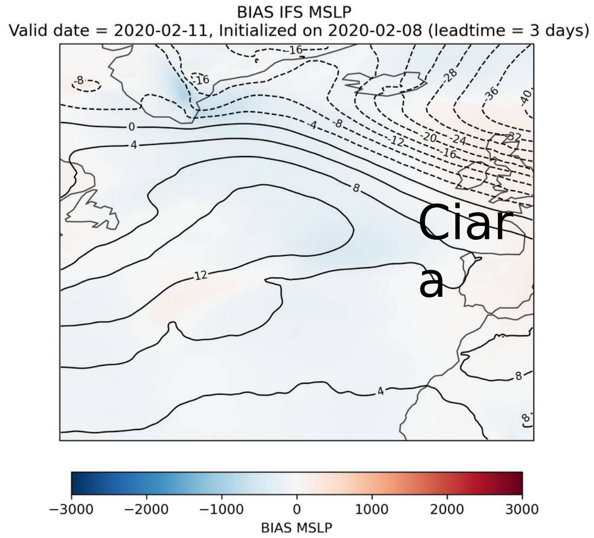
Storm Dennis on February 15, 2020

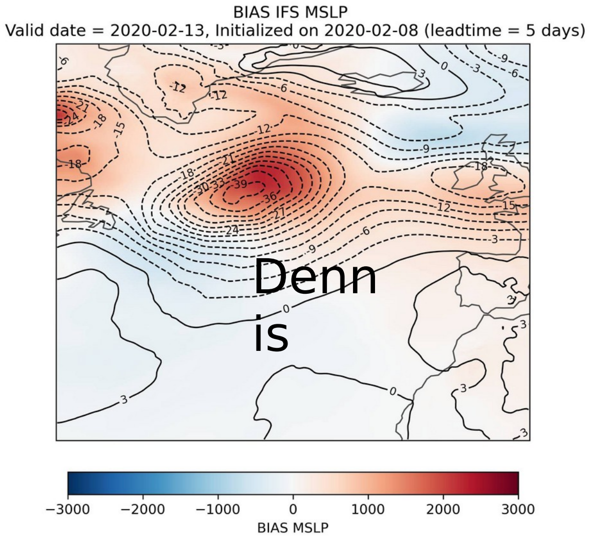
Storm Jorge on February 27, 2020

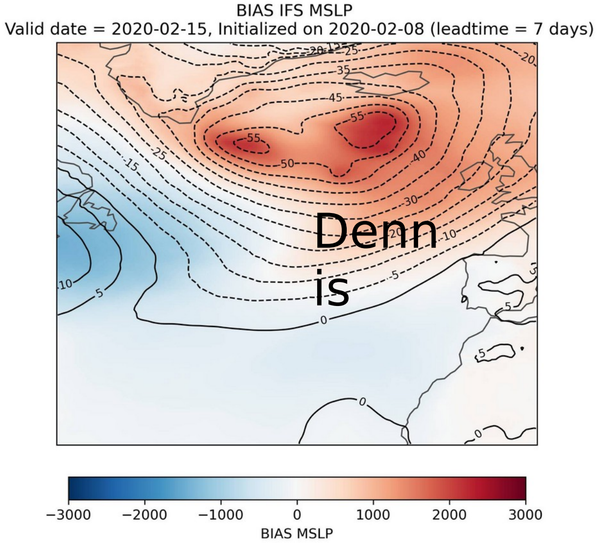
Forecasting storms Ciara and Dennis

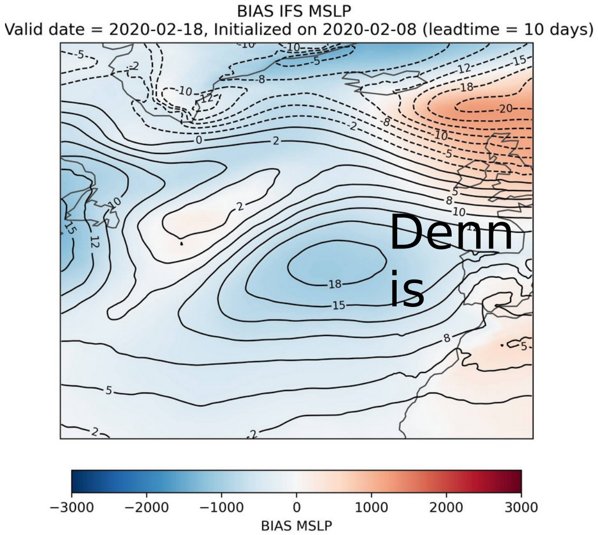
Bias in mean sea level pressure (MSLP) anomalies between **ECMWF IFS forecast** and **ERA-5 reanalysis** (shading):







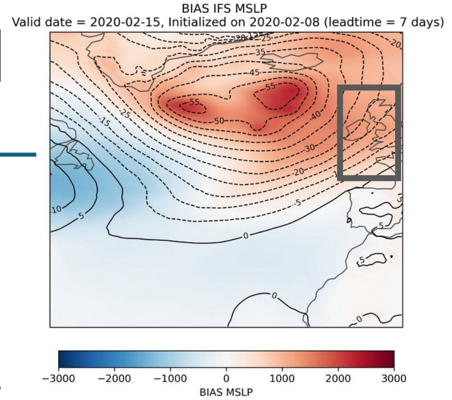


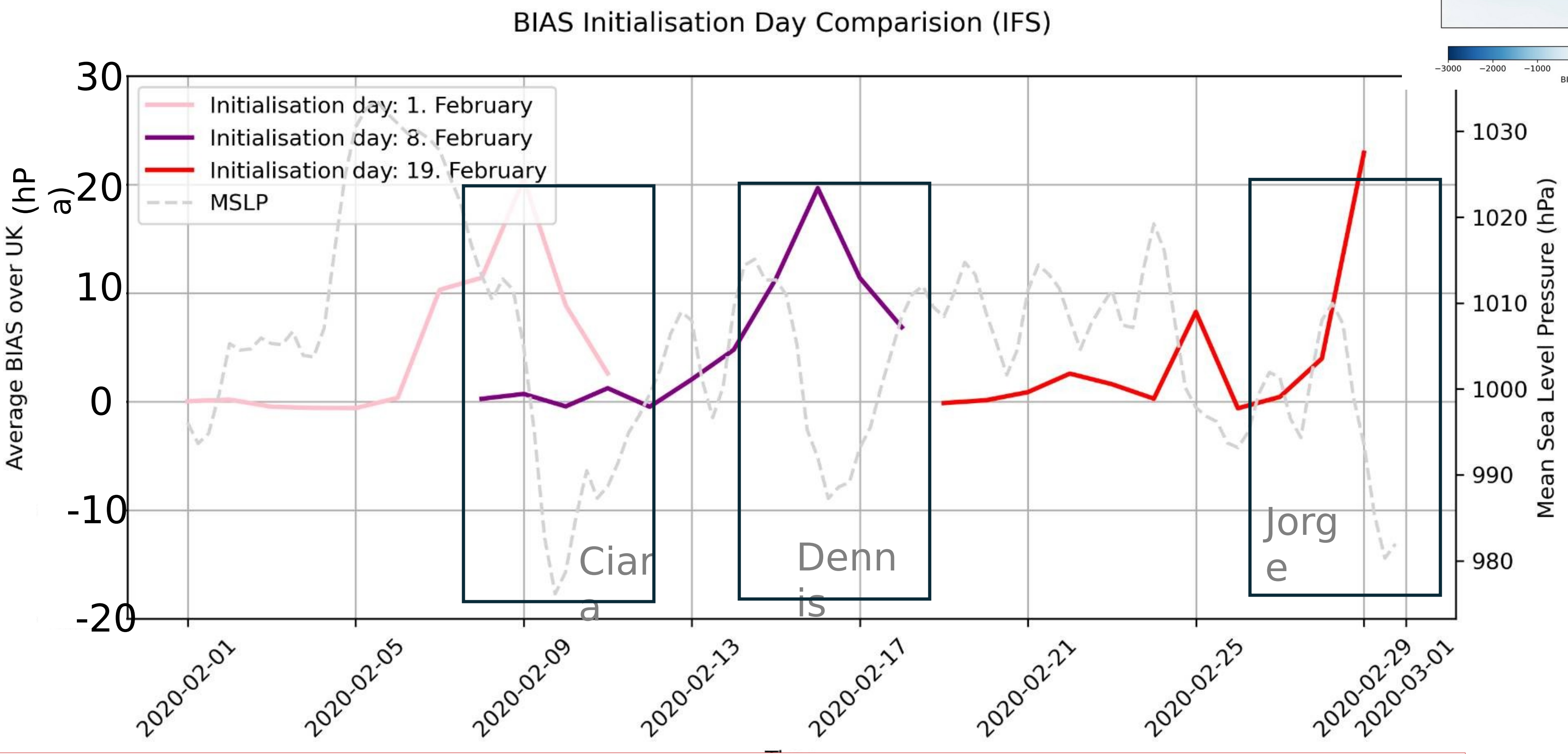


Black contours: MSLP in ERA-5 reanalysis

When initialized on the 8 February (storm Ciara), IFS (physics-based model) tends to underestimates cyclone intensity of the successive storm (storm Dennis) as it propagates through the North Atlantic

Average storm intensity bias over the UK





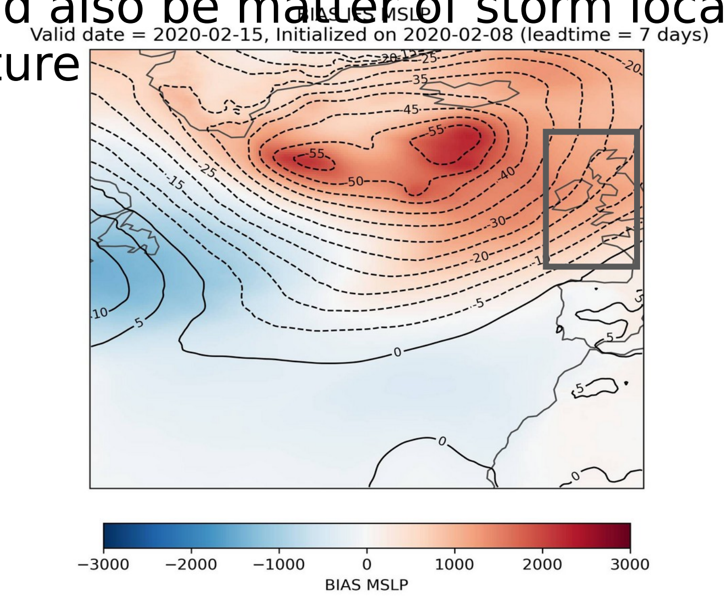
→ All storms have a comparable bias in IFS
→ Initializing the forecast with storm Ciara (8 Feb) does not reduce the bias for storm Dennis

Average storm intensity bias over the UK

Do AI models exhibit the same rate of error growth?

- AI Models are optimized to minimize their error

- Could also be matter of storm location / structure



MSLP at 8 days after forecast initialization

MSLP at 8 days after forecast initialization

Surface wind at 8 days after forecast initialization

Surface wind at 8 days after forecast initialization

Storm Severity Index (SSI), lead time 8 days

$$SSI = \left(\frac{V}{V_{ref}} - 1\right)^3$$

V_ref threshold: 20 m/s
(Klawa and Ulbrich, 2003)

All models underestimate severity index, especially for **Storm Jorge**

AI models: possibly smoother fields (?)

Forecast timeseries for storms Ciara, Dennis and Jorge;
UK box (12°W-5°E, 48°N-60°N)

Are some members as good as AI forecasts?

We can define the “best member” according to absolute error.

Forecast timeseries for storms Ciara, Dennis and Jorge; UK box (12°W-5°E, 48°N-60°N)

Are some members as good as AI forecasts?

We can define the "best member" according to absolute error.

We quantify the error in forecasting cyclone intensity in terms of MSLP and wind speed, using the Mean Absolute Error (MAE)

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|F_i - O_i|$$

Forecast timeseries for storms Ciara, Dennis and Jorge;
UK box (12°W-5°E, 48°N-60°N)

Are some members as good as AI forecasts?

We can define the “best member” according to absolute error.

We quantify the error in forecasting cyclone intensity in terms of MSLP and wind speed, using the Mean Absolute Error (MAE).

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|F_i - O_i|$$

The best-performing member is selected as the member with smallest MAE in 10-m wind speed, averaged over all lead times.

The “best member” of the physics-based model is **as good as** the AI models (for wind and MSLP)

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|F_i - O_i|$$

Low wind error

Low MSLP error

Errors of physically-linked variables are not strongly correlated in AI models

Cyclone-centered box, 6-deg wide

Correlation between variables

Errors of physically-linked variables are not strongly correlated in AI models

Cyclone-centered box, 6-deg wide

Correlation between variables

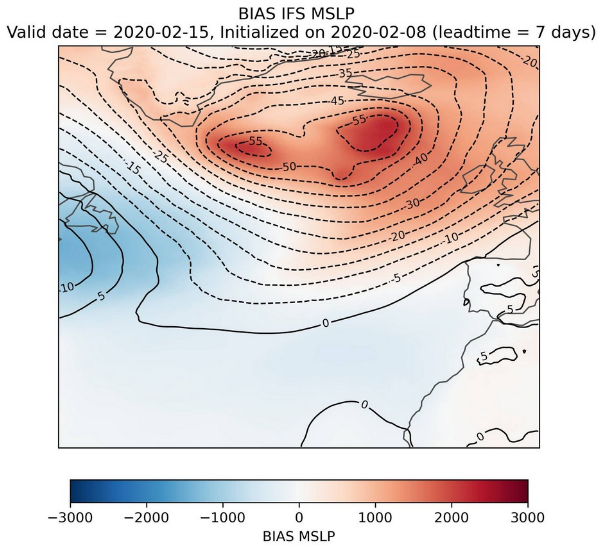
Correlation between errors

→ **Weaker correlation between errors of minimum MSLP and maximum 10-m wind**

Afargan-Gerstman et al., 2026, *npj Natural Hazards*

Summary: Can we trust AI models?

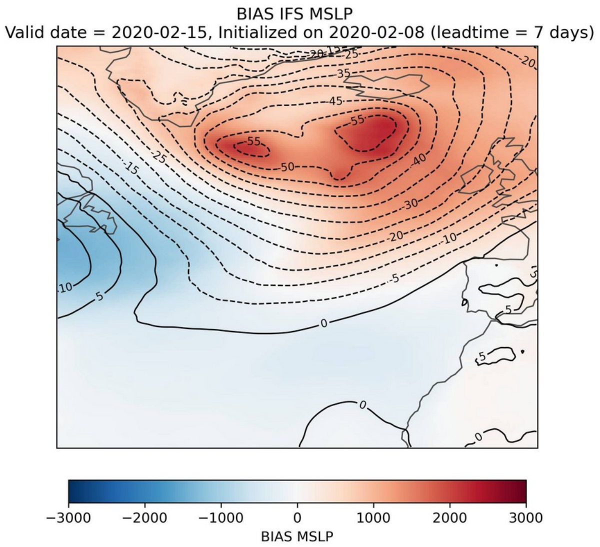
- The ability of weather model to accurately predict **extratropical cyclones series or clustering events** is critical to support effective disaster preparedness – especially challenging due to their aggregated impacts.



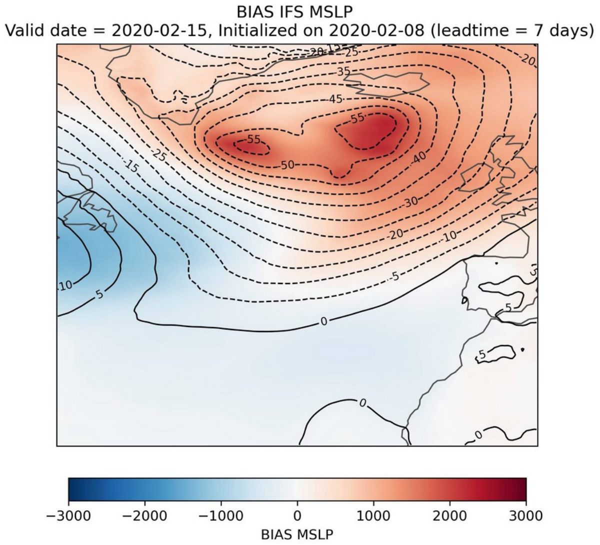
Summary: Can we trust AI models?

- The ability of weather model to accurately predict **extratropical cyclones series or clustering events** is critical to support effective disaster preparedness – especially challenging due to their aggregated impacts.

- Our results show that both data-driven and physics-based models tend to **underestimate storm intensity (MSLP)** and **surface wind**, particularly at lead times beyond 5–7 days.



Summary: Can we trust AI models?

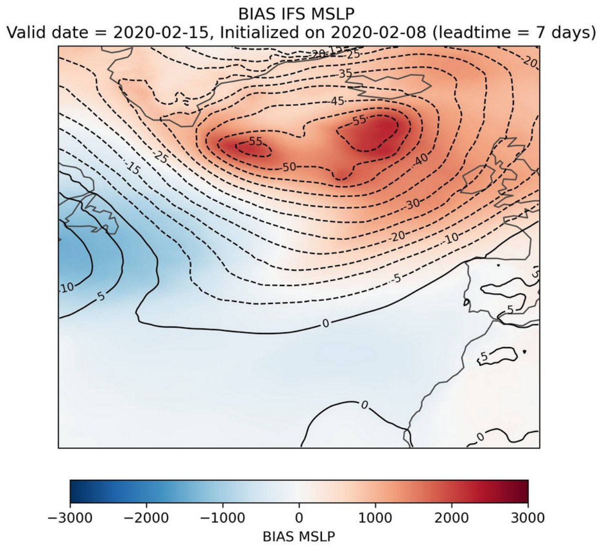


- The ability of weather model to accurately predict **extratropical cyclones series or clustering events** is critical to support effective disaster preparedness – especially challenging due to their aggregated impacts.

- Our results show that both data-driven and physics-based models tend to **underestimate storm intensity (MSLP)** and **surface wind**, particularly at lead times beyond 5–7 days.

- Unlike the physics-based model, the data-driven models show weaker error correlations between two physically linked variables (storm intensity and surface wind, averaged at 6-deg around cyclone center).

Summary: Can we trust AI models?



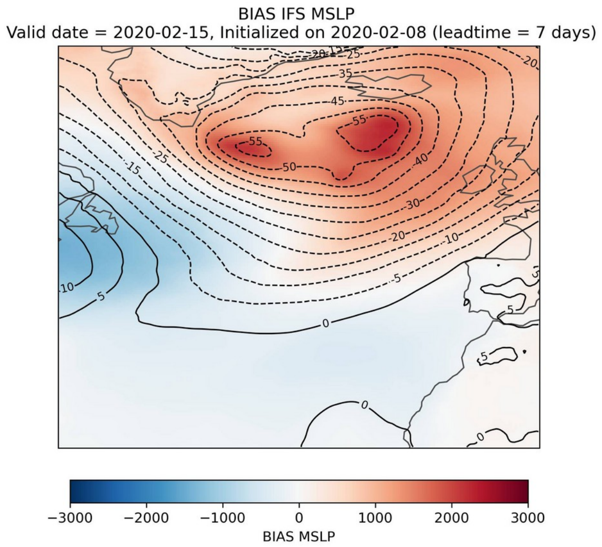
- The ability of weather model to accurately predict **extratropical cyclones series or clustering events** is critical to support effective disaster preparedness – especially challenging due to their aggregated impacts.

- Our results show that both data-driven and physics-based models tend to **underestimate storm intensity (MSLP)** and **surface wind**, particularly at lead times beyond 5–7 days.

- Unlike the physics-based model, the data-driven models show weaker error correlations between two physically linked variables (storm intensity and surface wind, averaged at 6-deg around cyclone center).

→ suggesting that despite improved ability of AI weather models in predicting the surface wind field (and consequently, windstorm-related impacts), these models potentially misrepresenting physical constraints.

Summary: Can we trust AI models?



- The ability of weather model to accurately predict **extratropical cyclones series or clustering events** is critical to support effective disaster preparedness – especially challenging due to their aggregated impacts.

- Our results show that both data-driven and physics-based models tend to **underestimate storm intensity (MSLP)** and **surface wind**, particularly at lead times beyond 5–7 days.

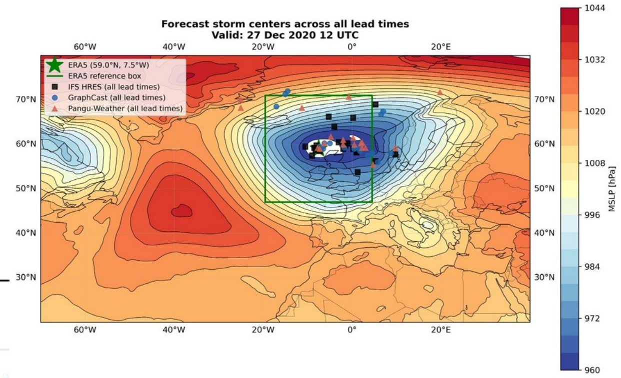
- Unlike the physics-based model, the data-driven models show weaker error correlations between two physically linked variables (storm intensity and surface wind, averaged at 6-deg around cyclone center).

→ suggesting that despite improved ability of AI weather models in predicting the surface wind field (and consequently, windstorm-related impacts), these models potentially misrepresenting physical constraints.

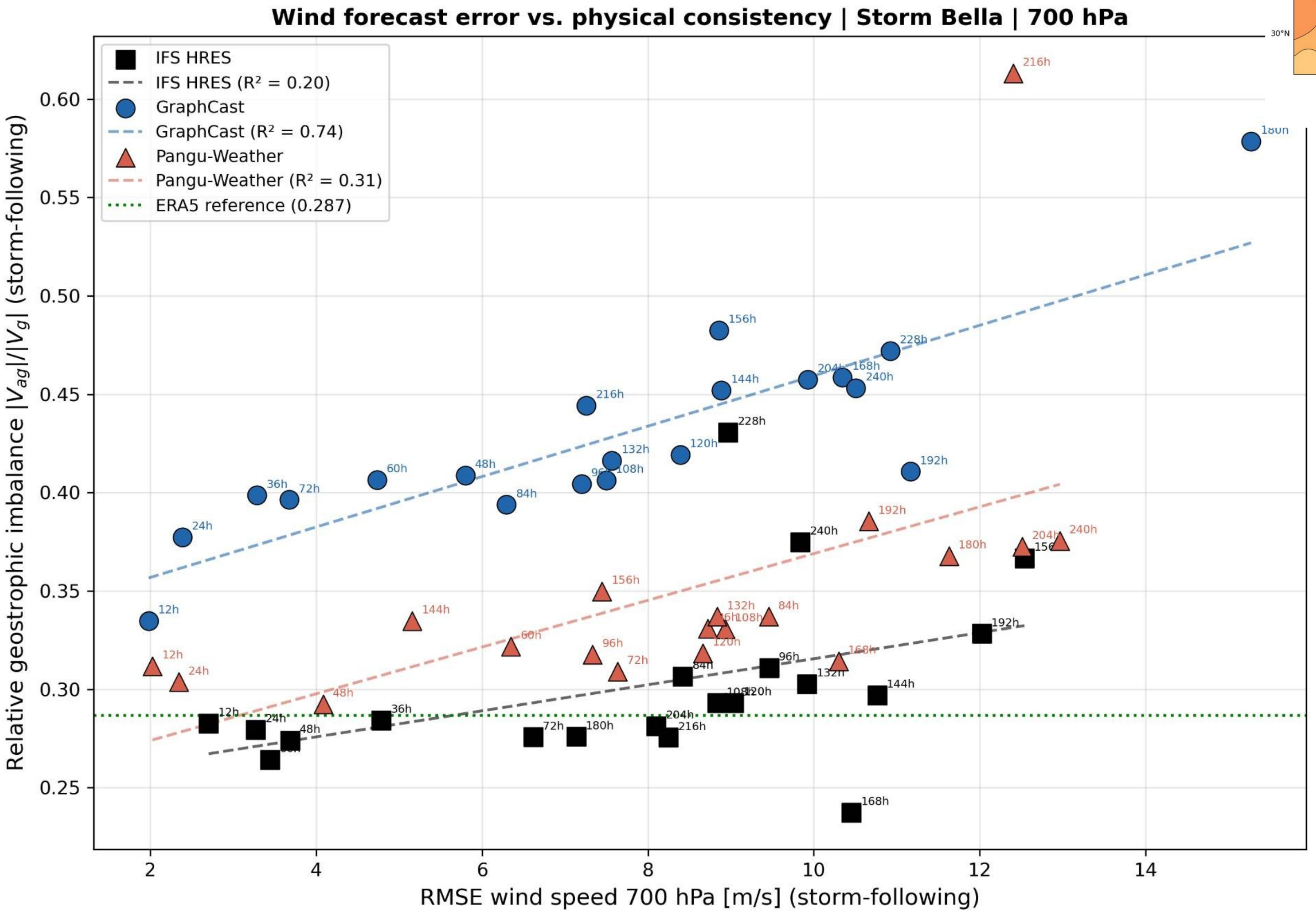
Next Steps:

- Compare to AI ensemble predictions | Extend prediction horizon
- Find useful metric to compare sources of error: misrepresentation of physical processes?

Bridge: physical consistency vs. forecast errors



How much "geostrophically balanced" is the cyclone wind?



The correlation between geostrophic imbalance and wind speed error is postive for all three models.

Extra Slides

Storm Bella
(25–28 December 2020)

Key facts

- Period: 25–28 December 2020
- Hit England & Wales overnight 26–27 December
- Min MSLP ~953 hPa at peak (27 Dec 12 UTC)

Impact (UK Met Office)

- Fallen trees and structural damage
- Large waves along the coast
- Widespread flooding

Flood and wind warnings across UK as Storm Bella sweeps in

Severe flood warnings are in place and winds of 60mph are expected into Sunday, Met Office says

The Great Ouse burst its banks at Bedford on Boxing Day. Photograph: Dan Kitwood/Getty Images

The Guardian

Physics-based and AI Models

IFS HRES

ECMWF, physics-based

- 0.10° Grid --> 0.25° Grid
- Operational baseline

GraphCast

DeepMind, Graph Neural Network

- 0.25° grid
- Trained on ERA5 1979–2019

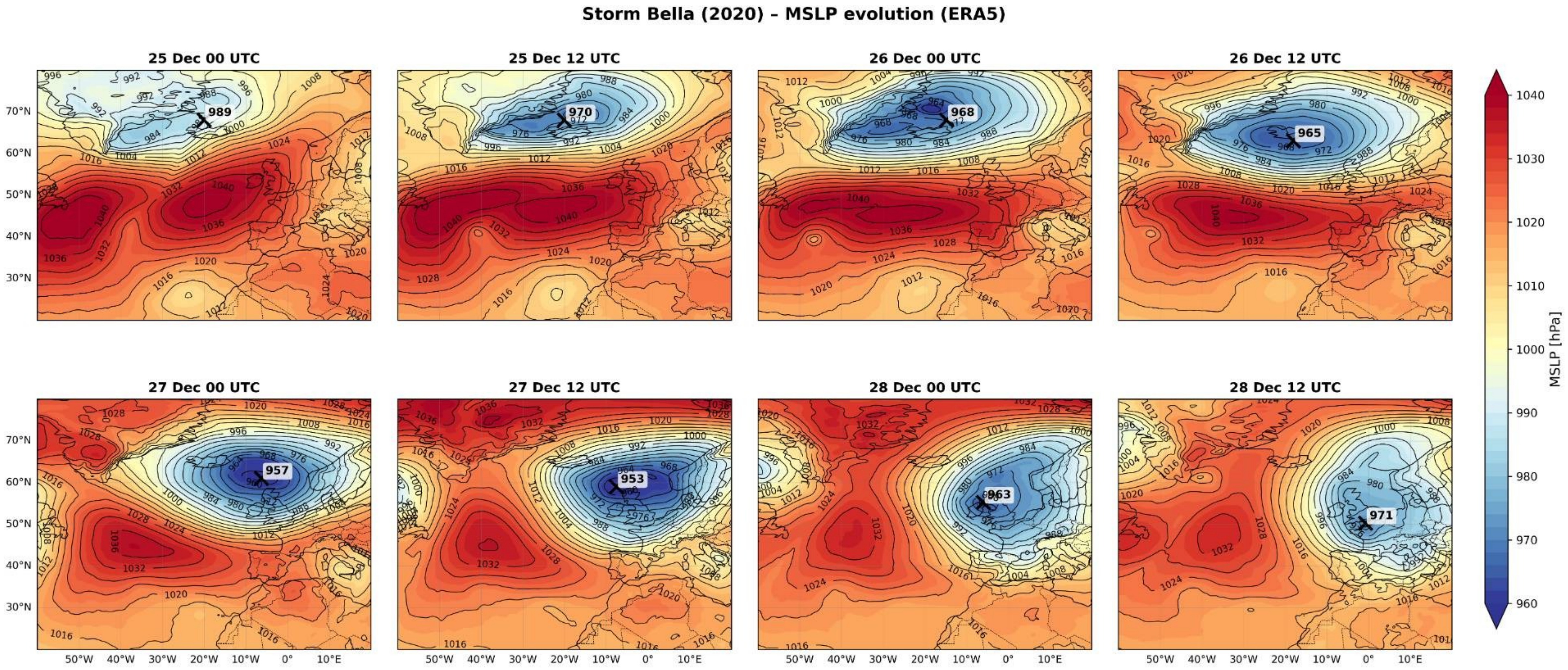
Pangu-Weather

Huawei, 3D Earth-Specific Transformer

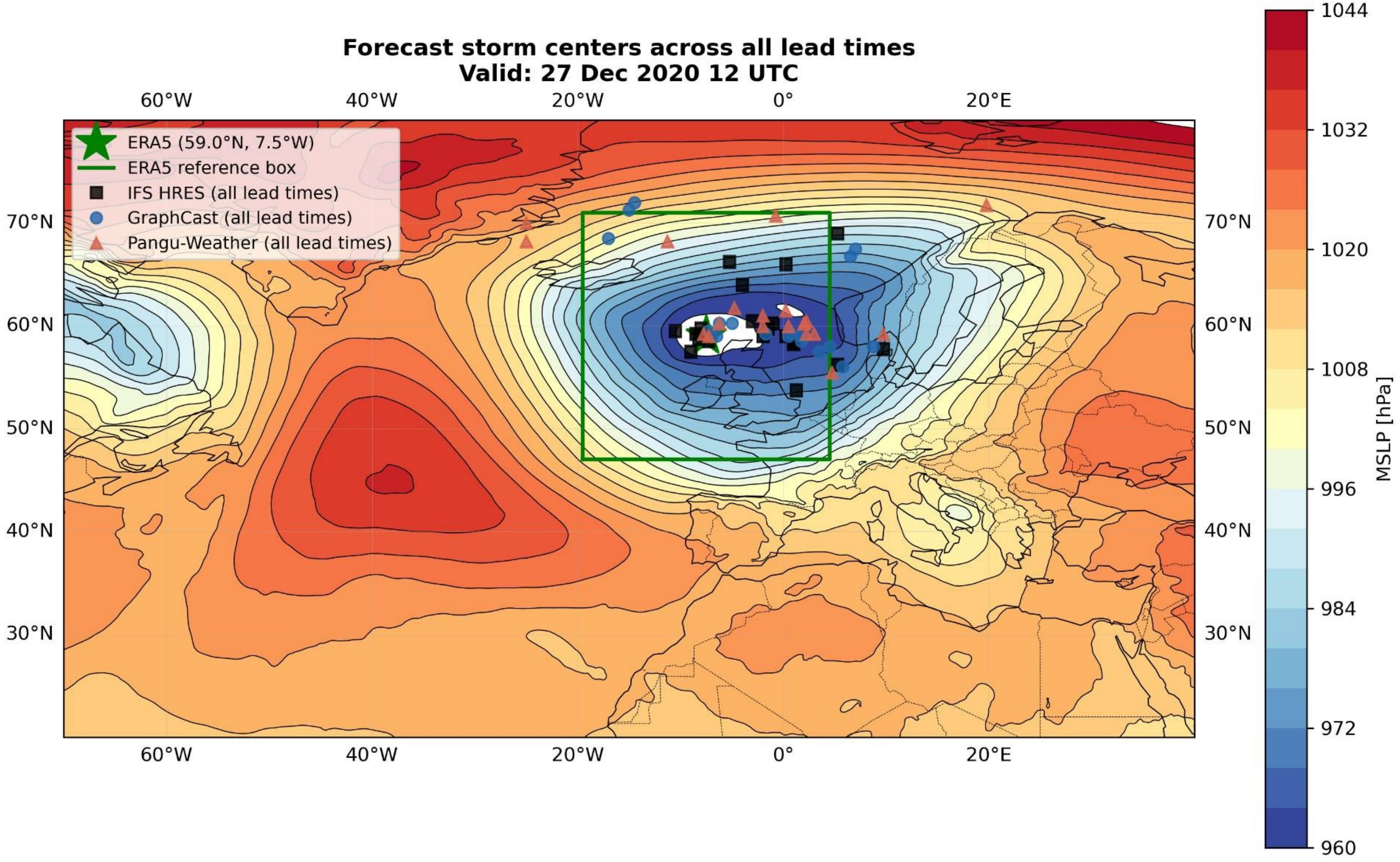
- 0.25° grid
- Trained on ERA5 1979–2017

Data source: WeatherBench 2 (Rasp et al. 2024).

Storm Bella — MSLP evolution

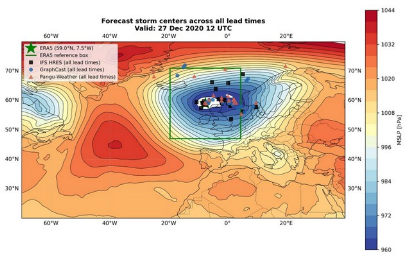


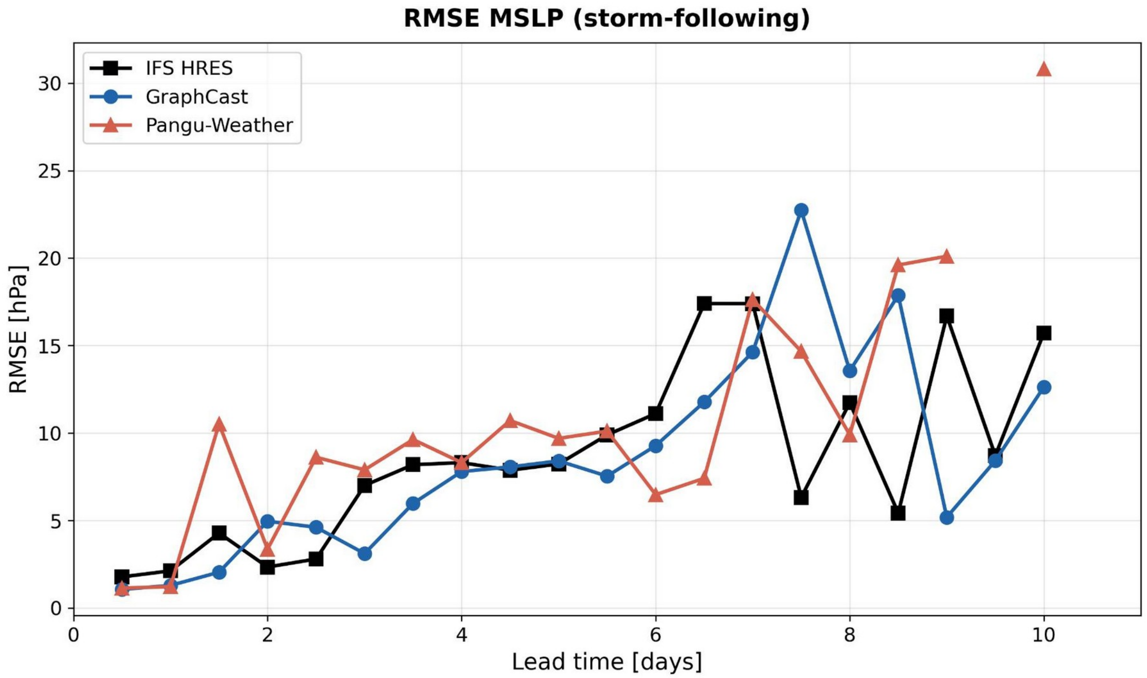
Forecast storm centres across all lead times

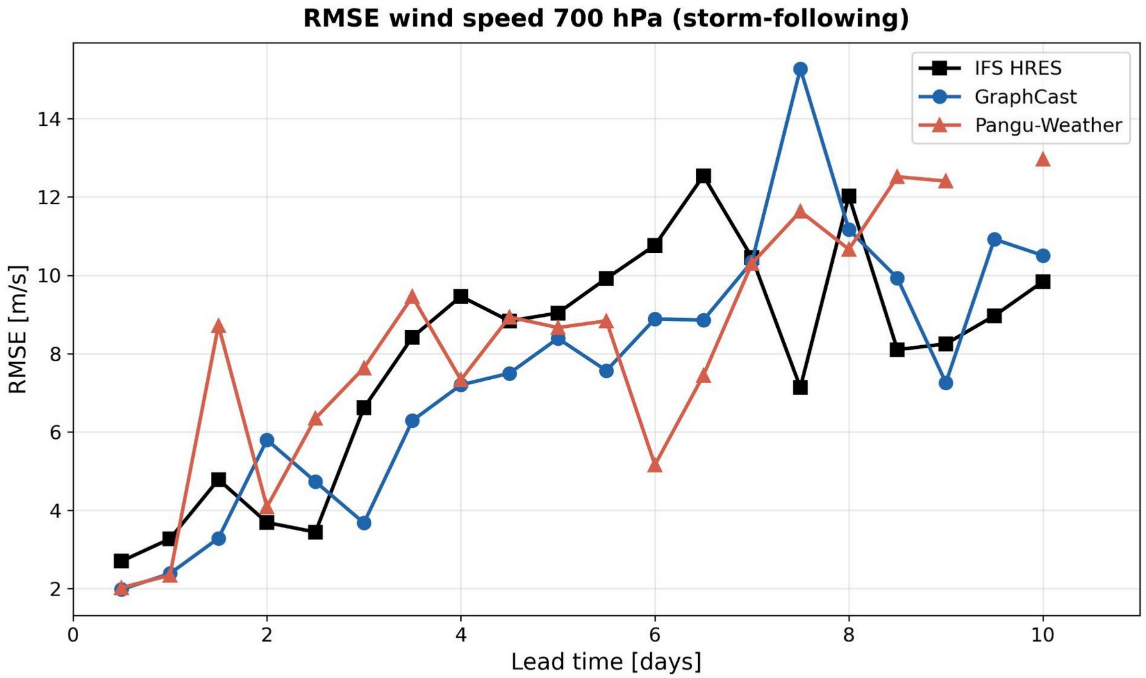


Verification metrics: RMSE

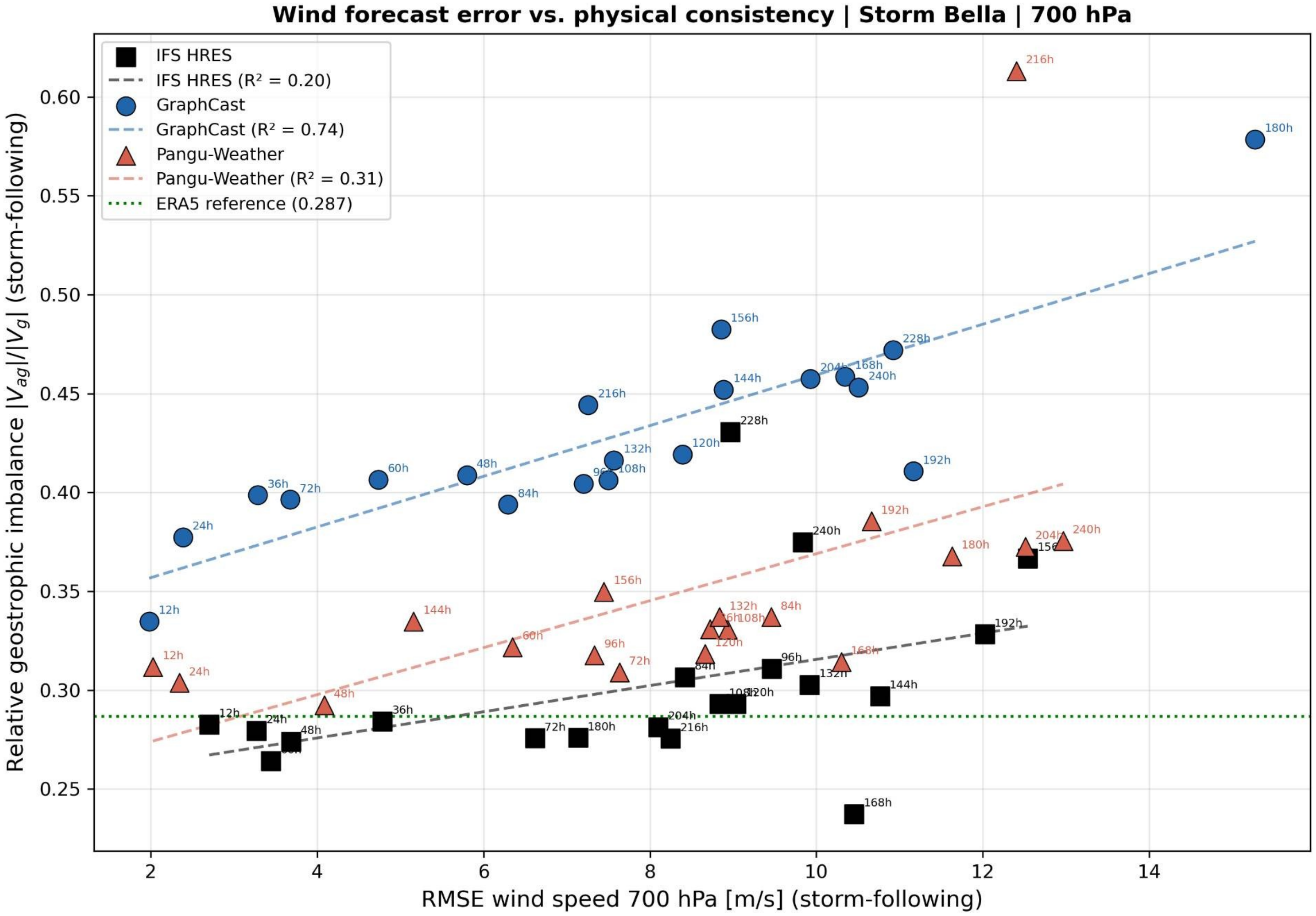
Lagrangian
perspective







Bridge: physical consistency vs. forecast errors



The correlation between geostrophic imbalance and wind speed error is postive for all three models.

Summary and outlook

Current findings

- All Models can accurately capture the storm structure up to 8-10 d in advance

- Geostrophic imbalance: Graphcast overestimates the imbalance
- Wind bias: All models underestimate the surface wind with lead times, consistent with recent studies in the field

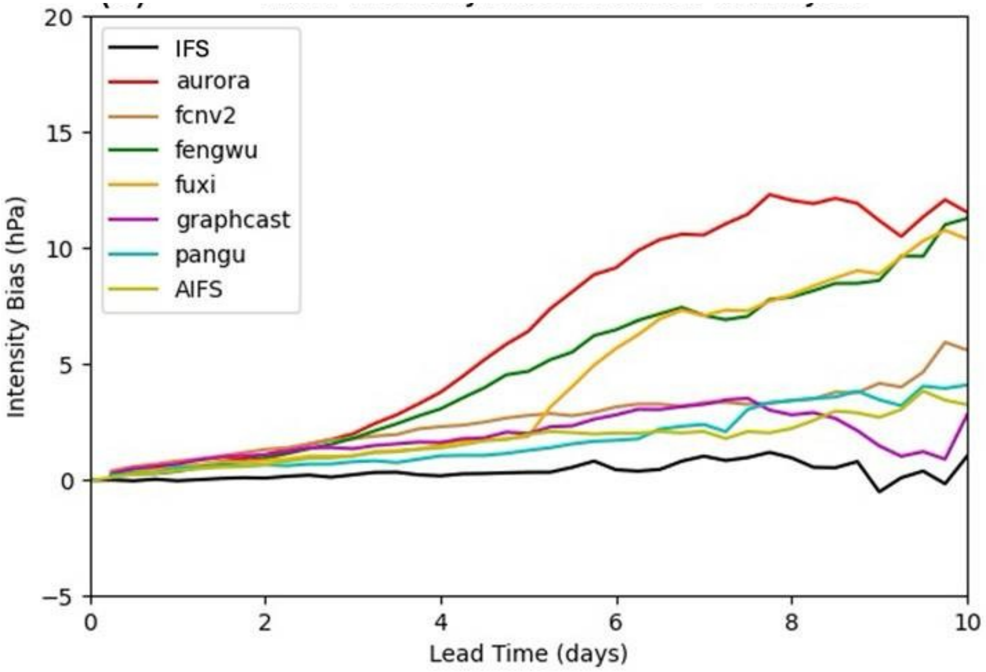
  ➢ These findings highlight that extreme weather events intensity is underestimated in AI forecast models and highlight the importance of physical consistencies.

Next steps

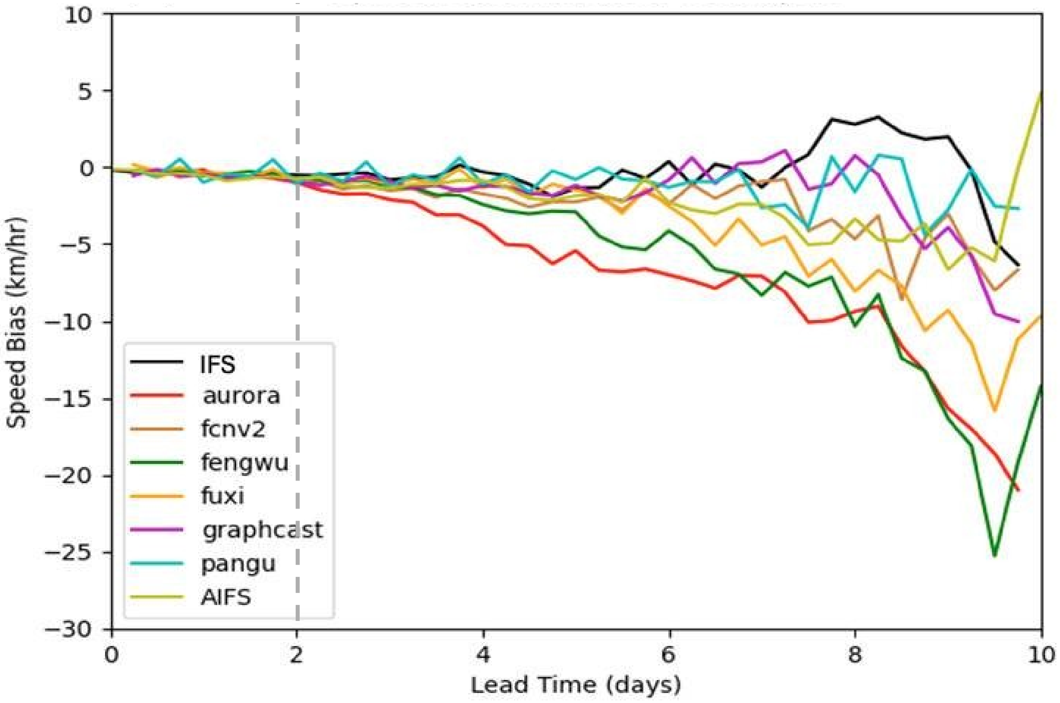
- Ensembles of AI–weather forecasts
- Extend prediction horizon:

AI weather prediction models **systematically underestimate** midlatitude cyclone intensity

Cyclone MSLP bias



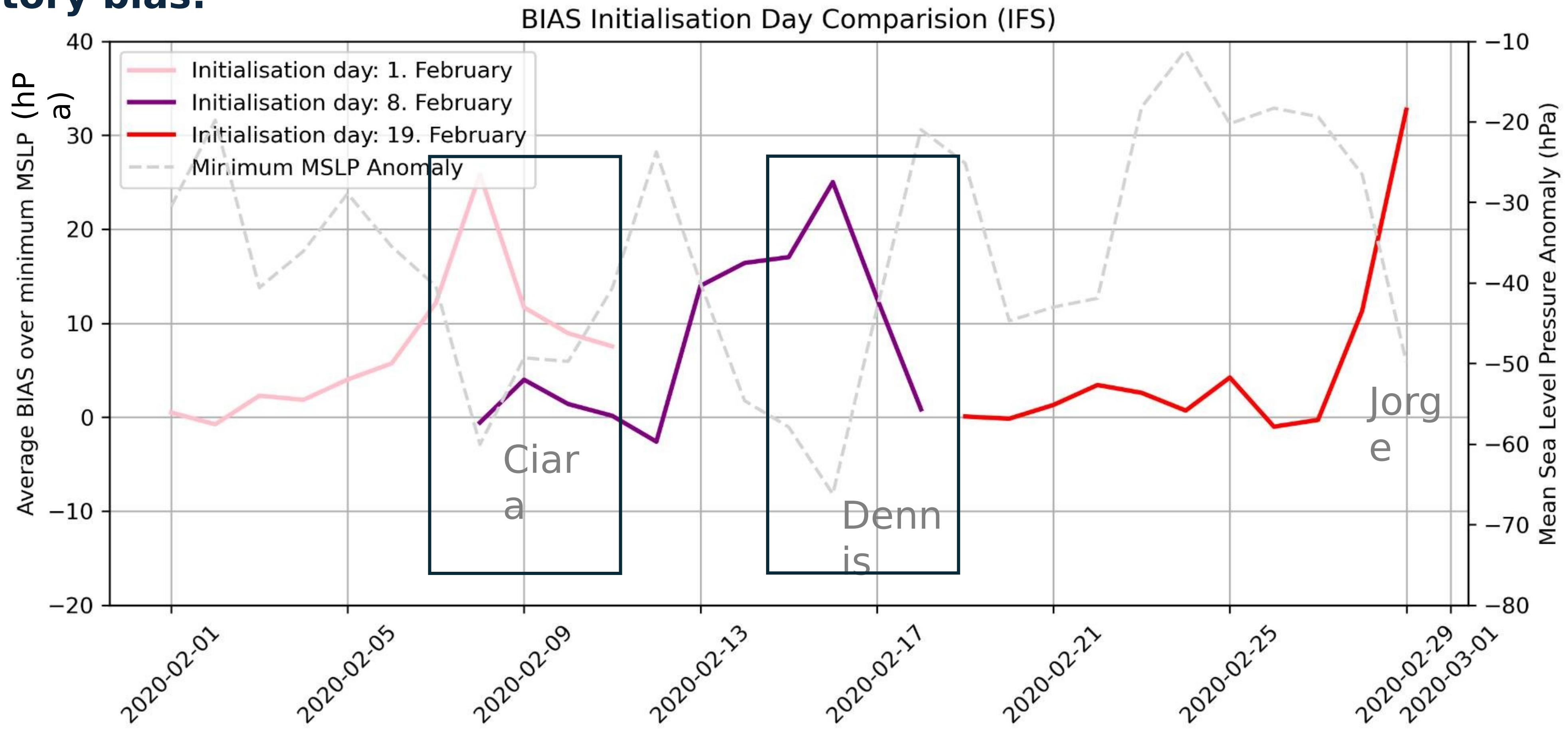
10-m wind speed intensity bias at cyclone-maximum



Figures: Dacre et al., 2026, *BAMS*.

Average bias from a cyclone-center perspective

Using this approach, we can analyze the **intensity bias** regardless of the **trajectory bias:**

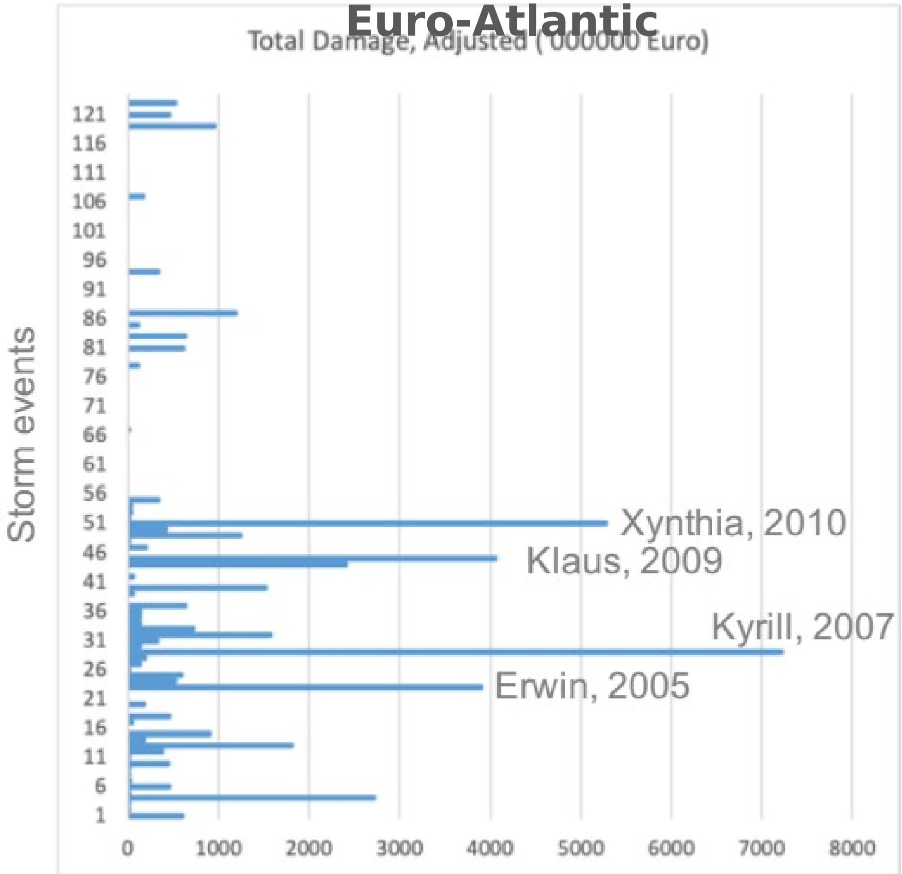


→ Forecast biases increase ~5 days after the initialization of storm Ciara, with a sharper increase.
→ Storms may increase forecast uncertainty (Rupp et al., 2024).

High-impact midlatitude storms

- In Europe alone, windstorms are one of the costliest hazards (Manning et al., 2022).
- Damages caused by extratropical cyclones amounted to 77 billion Euros between 1998 and 2022.
- Between 2000 to 2023, storms have led to more than 340 casualties in Europe.

Total damage related to extratropical cyclones over the Euro-Atlantic

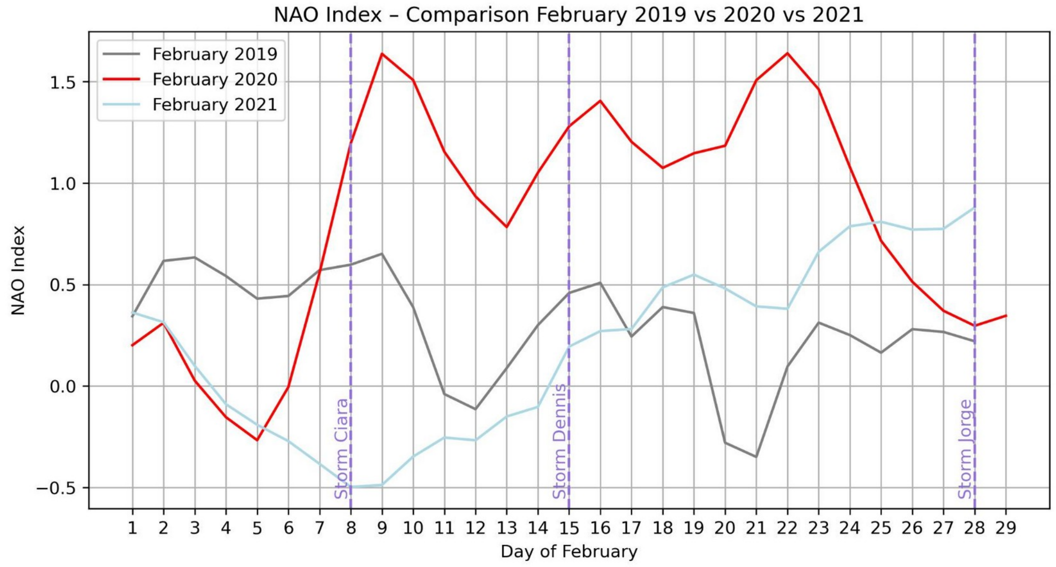


Damage (in million Euro)

Total casualties related to extratropical cyclones over the Euro-Atlantic

Number of casualties

Data: **Emergency Events Database (EM-DAT),** global database on natural disasters Guha-Sapir, 2021.



NAO Index – Comparison February 2019 vs 2020 vs 2021
February 2019
February 2020
February 2021
NAO Index
Day of February
Storm Ciara
Storm Dennis
Storm Jorge