# *Northern hemisphere midlatitude cyclone intensity biases in machine learning weather prediction models*



Dacre, H. F. ORCID: https://orcid.org/0000-0003-4328-9126, Charlton-Perez, A. J. ORCID: https://orcid.org/0000-0001-8179-6220, Driscoll, S., Gray, S. L. ORCID: https://orcid.org/0000-0001-8658-362X, Harvey, B. ORCID: https://orcid.org/0000-0002-6510-8181, Harvey, N. J. ORCID: https://orcid.org/0000-0003-0973-5794, Hodges, K. I. ORCID: https://orcid.org/0000-0003-0894-229X, Hunt, K. M. R. ORCID: https://orcid.org/0000-0003-1480-3755 and Volonté, A. ORCID: https://orcid.org/0000-0003-0278-952X (2026)
Northern hemisphere midlatitude cyclone intensity biases in machine learning weather prediction models. Bulletin of the American Meteorological Society, 107 (1). E208-E221. ISSN 1520-0477 doi: 10.1175/BAMS-D-25-0129.1 Available at https://centaur.reading.ac.uk/127752/

BAMS Article

# Northern Hemisphere Midlatitude Cyclone Intensity Biases in Machine Learning Weather Prediction Models

H. F. Dacre,[a] A. J. Charlton-Perez,[a] S. Driscoll,[a,b,c] S. L. Gray,[a] B. Harvey,[a,d] N. J. Harvey,[a] K. I. Hodges,[a,d] K. M. R. Hunt,[a,d] and A. Volonté[a,d]

**KEYWORDS:** Forecast verification/skill; Model comparison; Model evaluation/performance; Numerical weather prediction/forecasting; Machine learning

**ABSTRACT:** Forecasting the location and intensity of strong winds associated with midlatitude cyclones is important as they can have significant safety, economic, and environmental impacts. In this study, we use a feature-based evaluation method to assess the performance of both numerical weather prediction and machine learning weather prediction (MLWP) models in forecasting midlatitude cyclone winds. By tracking over 1000 cyclones across the Northern Hemisphere from 1 October 2023 to 31 March 2024 in seven MLWP models, we systematically compare model performance. Our results show that MLWP models predict midlatitude cyclone tracks with accuracy comparable to the ECMWF Integrated Forecasting System (IFS) forecast out to 10 days. However, MLWP models exhibit a persistent intensity bias, underestimating cyclone minimum sea level pressure by more than 5 hPa at 10-day forecast lead times, whereas the ECMWF IFS forecast has no bias. Additionally, all MLWP models produce weaker than observed peak 10-m winds, even at short lead times. In contrast, the ECMWF IFS forecast exhibits no bias in 10-m wind speed. These differences highlight the limitations of current MLWP models in capturing important high-impact weather features like peak wind speeds.

**SIGNIFICANCE STATEMENT:** Machine learning weather prediction models predict midlatitude cyclone tracks comparably to numerical weather prediction models but underestimate cyclone intensity, particularly minimum surface pressure and peak wind speeds, highlighting areas where machine learning models need improvement.

DOI: 10.1175/BAMS-D-25-0129.1
*Corresponding author*: H. F. Dacre, h.f.dacre@reading.ac.uk
Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/BAMS-D-25-0129.s1.

Manuscript received 2 May 2025, in final form 17 December 2025, accepted 20 December 2025

**AFFILIATIONS:** [a] Department of Meteorology, University of Reading, Reading, United Kingdom; [b] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, United Kingdom; [c] National Centre for Earth Observations, University of Reading, Reading, United Kingdom; [d] National Centre for Atmospheric Science, University of Reading, Reading, United Kingdom

## 1. Introduction

Weather forecasts are used by a wide variety of decision-makers across numerous sectors and industries. The decisions they make can have significant economic, social, and even life-saving consequences. Trust in weather forecasts ensures that decisions are made confidently, efficiently, and in a timely manner, which is critical in both everyday operations and during extreme weather events. Building trust in weather forecasts is a complex process that is developed over time and depends on several factors, including the accuracy of the forecasts, the transparency of the methods, and effective communication.

Quantifying the accuracy of weather forecasts typically involves comparing the model's predictions to observations, operational analysis, or reanalysis data. There are many methods for assessing the performance of weather forecasts. Broadly speaking, these methods fall into two categories: unconditional evaluation and conditional evaluation. Unconditional evaluation uses statistical measures to evaluate the overall performance of a model across a broad region or time period. Two commonly used global metrics are root-mean-square error (RMSE), which measures the average magnitude of the difference between predicted and observed values, and anomaly correlation coefficient (ACC), which is used to assess the correlation between predicted anomalies (deviations from the climatology) and observed anomalies. Conditional evaluation methods evaluate how well a model simulates certain weather systems or features, evaluating the model's accuracy in reproducing these phenomena when they occur.

The performance of numerical weather predictions (NWPs) has been extensively evaluated over the last 50 years resulting in a high level of trust (Bauer et al. 2015). In contrast, machine learning weather predictions (MLWPs) are gaining attention for their potential to rival traditional NWP methods, but the level of trust they have from decision-makers and meteorologists is still evolving. The training process of most MLWP models aims to minimize the mean-squared error of the outputs; therefore, to date, RMSE has often been used as the main forecast verification metric. Over the last few years, many MLWP models, including FengWu (Chen et al. 2023a), Pangu-Weather (Bi et al. 2023), GraphCast (Lam et al. 2023), FourCastNet (Pathak et al. 2022), FuXi (Chen et al. 2023b), Aurora (Bodnar et al. 2025), and ECMWF Artificial Intelligence/Integrated Forecasting System (AIFS) (Lang et al. 2024), have been shown to outperform traditional NWP forecasts when predicting a selection of variables over a variety of lead times using unconditional global metrics. However, since global metrics average errors across time and/or space, they can hide localized issues or poor performance in predicting extreme events that are critical for accurate forecasting and decision-making (Gilleland et al. 2009).

Recent studies have started to apply conditional evaluation methods to assess MLWP forecasts. These methods include dynamical analysis of atmospheric motions via spectral analysis (Bonavita 2024; Selz and Craig 2023) and the synoptic analysis of physical attributes of features of interest via case studies or climatologies. Weather features (such as cyclones,

fronts, thunderstorms, and precipitation bands) are coherent spatial structures which can shift in location or timing. Conditional feature-based evaluation investigates the accuracy of these spatial patterns rather than focusing on whether the model predicts exact values at individual points. Feature-based approaches can translate more directly into decision-making tools for forecasters, emergency managers, and industries that rely on weather predictions. For instance, knowing whether a model consistently overpredicts the size of thunderstorms, underpredicts the speed of hurricane movement, or the intensity of a midlatitude cyclone is more useful than looking at errors in temperature or wind speed at specific locations averaged over long time periods. Thus, feature-based evaluation is essential for building trust in weather predictions.

In the MLWP model evaluation literature to date, there has been a focus on evaluating a model's ability to forecast tropical features, including hurricanes and typhoons. In general, it has been demonstrated that MLWP models can produce more accurate tracks than NWP models (Bi et al. 2023; Lam et al. 2023), but that the rapid pressure drop in the eye of the storms is not so well captured (Pathak et al. 2022). Most recently, DeMaria et al. (2024) and Liu et al. (2024) have assessed multiple MLWP models and found that their tropical cyclone forecasts exhibit a bias with weaker-intensity tropical cyclones compared with NWP forecasts from 12- to 168-h lead times.

There have been fewer studies evaluating the performance of MLWP models in forecasting midlatitude weather features. Midlatitude cyclones (also known as windstorms, depressions, extratropical cyclones, and low pressure centers) are synoptic-scale surface weather features that affect midlatitude regions globally, with their highest frequency and intensity occurring in boreal autumn and winter. In both the North Atlantic and North Pacific, these cyclones typically follow similar paths, impacting regions such as North America, Europe, and East Asia. The winds associated with midlatitude cyclones can cause significant damage, leading to substantial insurance losses (Priestley et al. 2018). Charlton-Perez et al. (2024) evaluated the performance of various MLWP models in the prediction of the structure and evolution of a case study, Storm Ciarán. They find that, although the MLWP models are able to accurately forecast the minimum mean sea level pressure (MSLP), a common shortcoming is that none of them is able to capture the peak 10-m wind speed. Low-level wind extremes can cause significant impacts, so understanding MLWP model's ability to predict this metric is crucial for public safety and infrastructure planning. The aim of this study is to perform a systematic evaluation of MLWP midlatitude cyclone prediction over a full extended winter season to test how far the conclusions from the Storm Ciarán case study are valid for a broad set of midlatitude cyclones with varying intensities and structures.

## 2. Data and models

### *a. Verification data.*

**1) IFS ANALYSIS.** In this study, the ECMWF operational analysis [Integrated Forecasting System (IFS) analysis] is used as a benchmark for verifying and evaluating the accuracy of the NWP and MLWP forecasts (Vitart et al. 2022). The IFS is run on a spectral grid (resolution O1280) with the forecast output converted to a Gaussian grid with a grid spacing of approximately 9 km (0.1°) globally, and the atmosphere is represented using 137 vertical levels, extending from the surface to 0.01 hPa (approximately 80 km above sea level). The state-of-the-art ECMWF operational analysis uses a 4D-Var data assimilation system to integrate real-time observations into the ECMWF IFS (IFS Cycle 48r1), ensuring the best possible estimate of the atmospheric state over a 12-h time window.

**2) ERA5.** For comparison with the performance of the NWP and MLWP forecasts, we also compare the MLWP forecasts with the fifth generation ECMWF atmospheric reanalysis

(ERA5) (Hersbach et al. 2020). ERA5 has a horizontal grid spacing of about 31 km (0.25°) globally, so it has a similar resolution to the MLWP models, and it also provides the data that the MLWP models are trained on. Comparison to a dataset with a similar resolution ensures a fair comparison for smaller-scale features such as peak 10-m wind speed. It also uses 137 vertical levels, extending from the surface up to 80 km. ERA5 also uses a 4D-Var data assimilation scheme to combine historical observations with an older version of the ECMWF IFS Cycle 41r2.

Both the IFS analysis and ERA5 rely on NWP models that assimilate diverse data sources, including surface observations from weather stations, buoys, ships, upper-air measurements from radiosondes and aircraft, as well as satellite-based remote sensing data. Despite this, their spatial resolution may be too coarse to accurately capture localized wind speed maxima within cyclones, such as sting jets or within convective thunderstorms that form within the conducive environment produced by midlatitude cyclones.

***b. Forecast models.*** We perform 10-day forecasts initialized at 0000 UTC daily for the period 1 October 2023–31 March 2024 (183 forecasts). Some significant European midlatitude cyclones in this season include storm Babet (18–21 October 2023) (Thompson et al. 2024) and storm Ciarán (29 October–4 November 2023) (Gray and Volonté 2024), which both affected large parts of western Europe due to heavy rain and strong winds. All MLWP model forecasts are initialized from ERA5 to avoid potential incoherence in variable relationships (since the MLWP models are trained on ERA5). The IFS forecast is initialized with the IFS analysis. The impact of this experimental design is that the IFS forecasts may be initialized using a higher resolution, more accurate atmospheric state, and thus the short-term (24 h) forecasts may perform better than the MLWP model forecasts.

#### 1) NWP model.

***IFS forecast*** In this study, we compare the performance of the MLWP models to the ECMWF IFS deterministic forecast (IFS forecast). The IFS deterministic forecast is produced at spectral truncation wavenumber 1279 (TCo1279) corresponding to a 9-km grid spacing in physical space (IFS Cycle 48r1), the same configuration as the IFS analysis. It uses 137 vertical levels, with a top boundary extending up to 0.01 hPa.

#### 2) MLWP models.

At their core, MLWP models usually have an encoder–decoder architecture, typically powered by one or more transformer layers. In simple terms, an encoder–decoder network first compresses (i.e., encodes) the input data (e.g., meteorological variables on spatial grids) into a lower-dimensional abstract representation, known as a latent space, which is a compressed representation that captures essential spatiotemporal patterns. It then reconstructs (i.e., decoding) the target output. Transformers, originally introduced in natural language processing, are specific neural network architectures that use attention mechanisms. Attention allows the model to automatically learn which parts of the input are most relevant for forecasting; so, when applied to weather, the transformer layers decide which regions or variables to pay more attention to based on their importance for the final forecast.

In MLWP models, the first step of encoding is typically a patch-embedding procedure. The input spatial fields are divided into small, nonoverlapping patches, each of which is then flattened to a one-dimensional array. This array is then projected into a fixed-dimensional embedding space, i.e., a more compact representation of the patch. This projection is repeated several times during the encoding stage.

Typically, MLWPs are trained using selected near-surface and three-dimensional fields (on isobaric surfaces) from ERA5 reanalysis from 1979 through to around 2020 (depending

on when the model was published and how much data was reserved for testing). They produce forecasts autoregressively, meaning that the output from one 6-h forecast step effectively becomes the input to the next. This enables multistep forecasts to be produced without needing to retrain the model for different lead times, although such retraining is done for some MLWP models. We describe the nuances that separate the seven MLWPs listed in Table 1 in the appendix.

**Table 1. Datasets from NWP, MLWP, analysis, and reanalysis evaluated in this study.**

| Model name | Model type | Horizontal grid spacing |
|---|---|---|
| IFS analysis | Analysis | 0.1° |
| ERA5 | Reanalysis | 0.25° |
| IFS forecast | NWP | 0.1° |
| Pangu-Weather | MLWP | 0.25° |
| GraphCast | MLWP | 0.25° |
| FengWu | MLWP | 0.25° |
| ECMWF-AIFS | MLWP | 0.25° |
| Aurora | MLWP | 0.25° |
| FourCastNetv2 | MLWP | 0.25° |
| FuXi | MLWP | 0.25° |

***c. Midlatitude cyclone identification and tracking.*** We use an objective cyclone identification and tracking algorithm (Hodges 1994, 1995) to analyze data from the forecasts and analyses (Table 1) for the period 1 October 2023–31 March 2024. Cyclone tracks are identified every 6 h based on 850-hPa relative vorticity, truncated to a T42 resolution (approximately 2.8°), and have the large-scale background removed, to highlight synoptic-scale phenomena. The start of a cyclone is defined when the filtered vorticity maximum first exceeds $1.0 \times 10^{-5}$ $s^{-1}$ and ends when it falls below this threshold. Cyclone tracks that are confined to the tropics (<30°N) and are short-lived (lifetimes shorter than 48 h) are filtered out following completion of the tracking. Figure 1 shows the tracks of the 1008 midlatitude cyclones identified in the IFS analysis between 1 October 2023 and 31 March 2024. The tracks are colored according to the maximum 10-m wind speed within a 6° geodesic radius of the vorticity maxima location. Cyclones occur preferentially over the North Atlantic and North Pacific Ocean basins with the strongest 10-m wind speeds between 30° and 60°N.

***d. Evaluation metrics.*** We evaluate diagnostics only for cyclone tracks that exist in both a forecast dataset and the analysis/reanalysis dataset. Following Froude et al. (2007), forecast cyclone tracks are said to be matched with an analysis/reanalysis cyclone track if the two tracks meet certain spatial and temporal constraints. The separation distance between each of the first four points (24 h) of the forecast track and their corresponding points in the analysis track must be less than 4°. Matched tracks are limited to forecast cyclone tracks

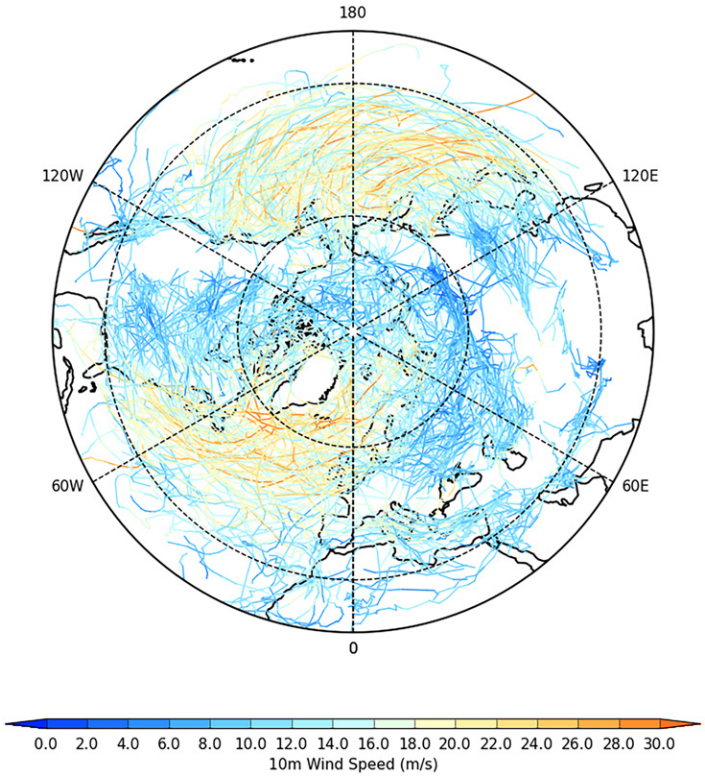


**Fig. 1. NH midlatitude cyclone tracks identified in the ECMWF IFS analysis between 1 Oct 2023 and 31 Mar 2024. Tracks are colored according to the maximum 10-m wind speed within a 6° radius of the cyclone center along the track. The dashed latitude circles are at 30° and 60°N.**

which exist at forecast day 0 or whose genesis occurs within the first 3 days of the forecast. Tracks that have their genesis beyond 3 days are not included in the analysis to avoid randomly generated cyclones in the model forecasts being matched by chance. There are similar numbers of matched tracks in each of the forecast-verification datasets. The average number of matched tracks in a 10-day period for the IFS forecast is 26.3, with a hit rate of 0.70 (see Table 2). By comparison, the average number of matched tracks in the seven MLWP models ranges from 23.8 to 26.7 with hit rates between 0.63 and 0.71. The high hit rates demonstrate that the results are representative of overall model forecast performance. On average, the number of matched tracks used to calculate the diagnostics increases from approximately 2250 on day 0 to over 3000 on day 2 of the forecast (see Fig. S1a in the online supplemental material). The number of matched tracks then decreases rapidly to less than 500 by day 7, as cyclones present in the first 3 days of the forecast dissipate (known as lysis). Note that the same cyclone will contribute multiple times to the sample size, as it is present in forecasts initialized at successive times (see Fig. S1b).

A cyclone center is defined as the location of the maximum 850-hPa relative vorticity. The location error represents the geodesic separation distance (great circle) between a cyclone center A, on a forecast track, and a cyclone center B, on an analysis/reanalysis track. We use the angular distance (°) between A and B to quantify the location error, which is equivalent to the geodesic distance since 1° ≈ 111 km. Since the location error is measured between analysis and forecast cyclone centers that occur at the same time, it will include components of both along-track error (propagation speed error), cross-track error (error occurring because the forecast cyclone takes a different path than the analyzed cyclone), and error in the genesis location of the cyclone. Note, however, that the mean error in the cross-track direction is substantially smaller than in the along-track direction. This difference is primarily due to the cancellation of errors on either side of the track, indicating that there is no significant tendency for cyclones to move to the left or right of the analysis track. The bias in cyclone propagation speed represents the difference between the cyclone movement in the forecast and analysis/reanalysis datasets. The cyclone propagation speed is calculated using the distance between consecutive 6-hourly points on the forecasts and analysis tracks independently. Positive values of propagation speed bias indicate that the forecast cyclone movement is too fast, and negative values indicate the cyclone movement is too slow.

The cyclone-maximum 10-m wind speed intensity error is the absolute difference between the forecast and analysis/reanalysis maximum wind speed within 6° of each

**Table 2. Average number of NH cyclones with genesis north of 30°N in a 10-day period from the IFS analysis, the IFS forecast, and each of the MLWP model forecasts (column 2). The average number of forecast cyclones matched with an analysis cyclone (column 3), hit rate (column 4), and false alarm rate (column 5). Averages are taken over 183 forecasts between 1 Oct 2023 and 31 Mar 2024.**

| | No. of cyclones in 10-day analysis/forecast period | No. of cyclones matched with IFS analysis | Hit rate | False-alarm rate |
|---|---|---|---|---|
| IFS analysis | 38.3 | 38.3 | 1.00 | 0.00 |
| IFS forecast | 47.4 | 26.3 | 0.70 | 0.44 |
| Pangu-Weather | 45.4 | 24.8 | 0.65 | 0.46 |
| GraphCast | 45.5 | 25.3 | 0.67 | 0.44 |
| FengWu | 37.7 | 24.7 | 0.65 | 0.34 |
| ECMWF-AIFS | 47.2 | 26.7 | 0.71 | 0.43 |
| Aurora | 35.7 | 25.3 | 0.67 | 0.29 |
| FourCastNetv2 | 43.5 | 23.8 | 0.63 | 0.45 |
| FuXi | 39.9 | 25.3 | 0.67 | 0.36 |

cyclone vorticity center. The intensity bias is the signed difference, with positive values indicating overestimated winds and negative values indicating underestimated winds. Similarly, the MSLP intensity error is the absolute difference between the forecast and analysis/reanalysis minimum MSLP within 5° of each cyclone vorticity center. The MSLP bias is the signed difference, where positive values mean forecast MSLP is too shallow, and negative values mean it is too deep. Confidence intervals are shown as shaded regions in Figs. S2–S4. They generally widen after forecast day 7 due to reduced sample size. The 95% confidence intervals are computed as ±1.96× standard error of the mean. If the confidence intervals do not overlap, then, in general, differences in model performance are considered statistically significant.

## 3. Results

***a. Cyclone track location performance.*** Figure 2a shows the average distance of the matched Northern Hemisphere (NH) cyclone tracks from the IFS analysis tracks as a function of forecast lead time for the NWP IFS forecast and seven MLWP model forecasts. As expected, the accuracy of the cyclone tracks decreases as the forecast lead time increases, and for all models, the location error is greater than 10° by day 10. For context, 10° is the distance between San Francisco, California, and Seattle, Washington. For most MLWP models, the performance is indistinguishable from the NWP IFS forecasts, i.e., the 95% confidence intervals overlap (see Fig. S2), with the exception of ECMWF-AIFS and FuXi, which exhibit an improved track position compared to the NWP IFS forecast for 4–8-day lead times.

The propagation speed of the cyclones is also important, as knowing when as well as where a cyclone may have damaging impacts is useful for decision-makers. The NWP IFS forecast cyclone propagation speed bias is close to zero for the first 7 days of the forecast (black curve in Fig. 2b). Pangu-Weather, GraphCast, ECMWF-AIFS, and FourCastNetv2 also exhibit negligible cyclone propagation speed bias over this time window. However, FengWu, Aurora, and FuXi have an increasing cyclone propagation speed bias from day 4 onward, with cyclones propagating too slowly in these forecasts. There is also some evidence of oscillatory behavior in Pangu-Weather and FuXi. Similar propagation speed biases were observed by Froude et al. (2007) in their evaluation of IFS cycle 31r1 forecasts when compared to the verifying analysis. Since the IFS cycle 31r1 grid resolution is the same as the MLWP models (25 km), one potential reason for the underestimation of propagation

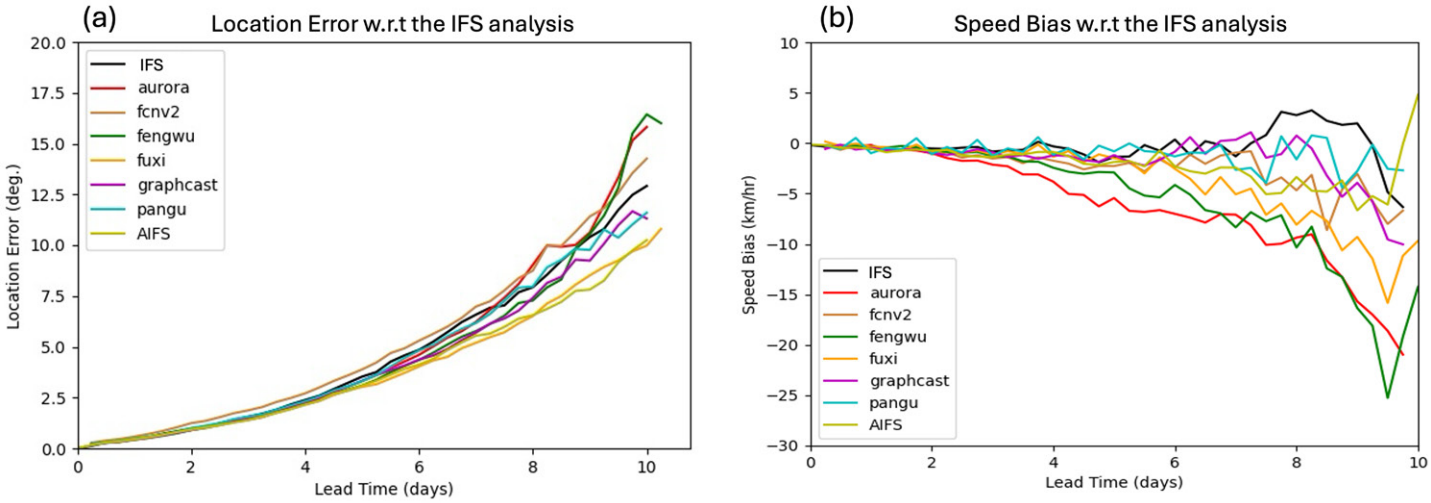


**Fig. 2. NH cyclone track location error with respect to the IFS analysis as a function of forecast lead time (days). (a) Track location error (°) and (b) cyclone propagation speed bias (km h$^{-1}$). Note that the geodesic distance can be converted to kilometers using 1° ≈ 111 km.**

speed may be representation of low-level frictional drag in the training dataset, although the magnitude of the cyclone propagation speed error varies between the MLWP models. See Fig. S2 for an estimation of uncertainty.

***b. Wind intensity performance.*** Figure 3a shows absolute maximum cyclone wind speed error as a function of forecast lead time, and Fig. 3b shows cyclone-maximum 10-m wind

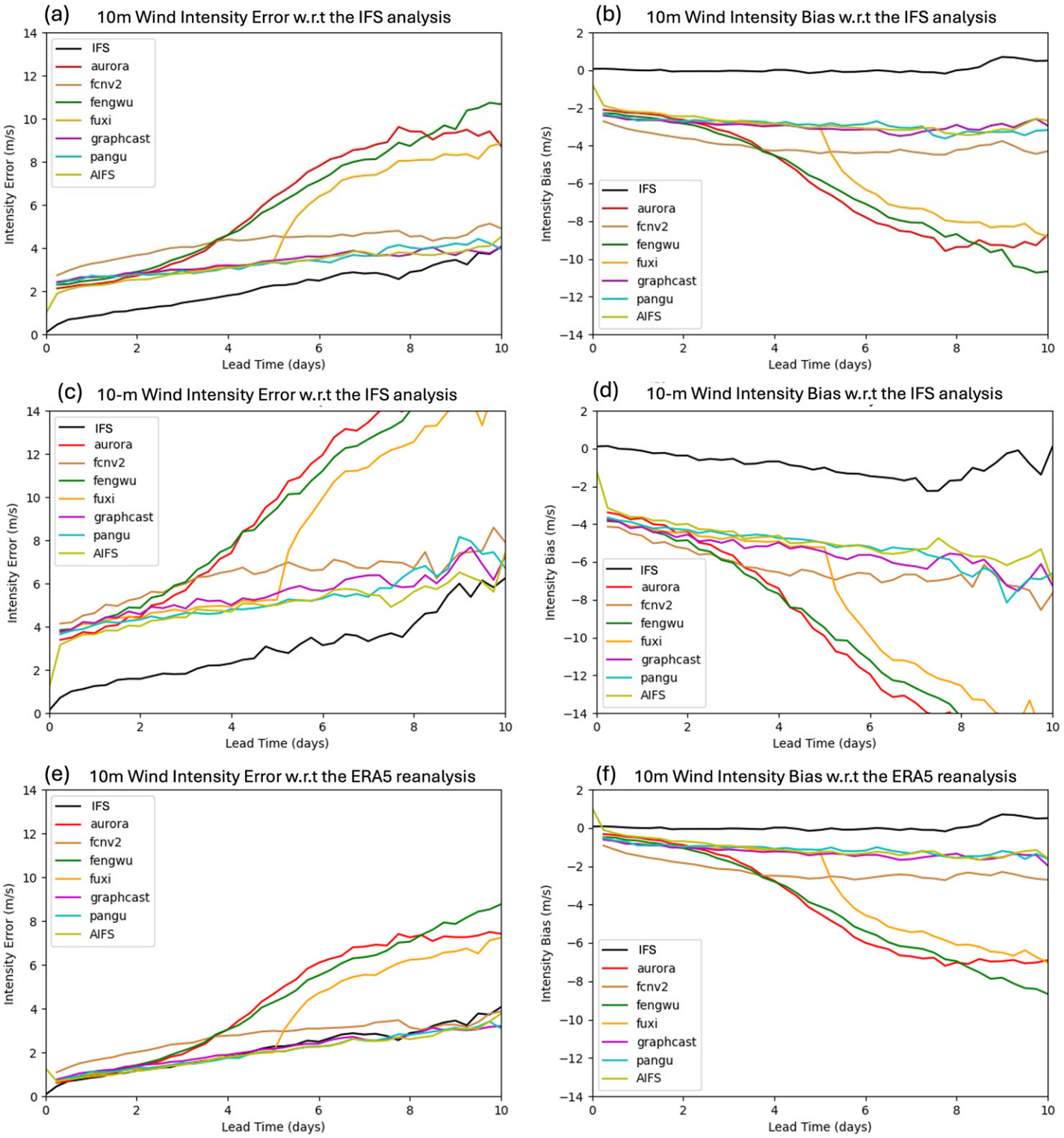


**FIG. 3. NH cyclone-maximum 10-m wind speed error as a function of forecast lead time (days). (a),(e) Wind speed error (m $s^{-1}$) for all cyclones and (c) wind speed error for strong gale periods only. (b),(f) Wind speed bias (m $s^{-1}$) for all cyclones and (d) wind speed bias for strong gale periods only. (a)–(d) Error with respect to the IFS analysis, and (e),(f) error with respect to the ERA5 reanalysis.**

speed bias as a function of forecast lead time with respect to the IFS analysis. All models show an increase in cyclone-maximum 10-m wind speed error as the forecast lead time increases. The NWP IFS forecast error increases over the first 12 h of the forecast due to decorrelation between the IFS analysis and the IFS forecast error, but after this, the error growth rate is fairly constant, reaching 3–4 m $s^{-1}$ by day 10 (black curve in Fig. 3a). The cyclone-maximum 10-m wind speed error observed after 12 h in the NWP IFS forecast does not lead to a systematic over- or underprediction of the cyclone-maximum 10-m wind speed (black curve in Fig. 3b).

All of the MLWP models have larger cyclone-maximum 10-m wind speed errors than the NWP IFS forecast model. For all MLWP models, a cyclone-maximum 10-m wind speed error of 2–3 m $s^{-1}$ occurs from 6 h into the forecast, which manifests as a systematic underprediction of the cyclone-maximum 10-m wind speed (Fig. 3b). There is, however, significant variability in the growth rate of the MLWP model cyclone-maximum 10-m wind speed error. In the Pangu-Weather, GraphCast, ECMWF-AIFS, and FourCastNetv2 MLWP models, the error growth rate is fairly constant with lead time, resulting in cyclone-maximum 10-m wind speed errors of 4–5 m $s^{-1}$ by day 10. In FengWu and Aurora, the error growth rate increases with lead time, resulting in cyclone-maximum 10-m wind speed errors of around 10 m $s^{-1}$ by day 10. Finally, in FuXi, there is a sharp increase in cyclone-maximum 10-m wind speed error at day 5 of 3 m $s^{-1}$, doubling the error. The FuXi weather forecasting system employs a cascade approach, where multiple fine-tuned models are used for different forecast time ranges (see appendix for details). This appears to result in discontinuities when forecasting features that exist across these time windows.

In Figs. 3a and 3b, the 10-m cyclone wind speed errors are averaged across cyclones of different intensities and at different stages of their life cycle. Analysis of strong wind periods only (maximum 10-m wind speed exceeding 21 m $s^{-1}$, strong gale on Beaufort scale) (Figs. 3c,d) suggests that errors are even larger for more intense wind periods. This has important consequences if MLWP models are used to issue hazard warnings. Note, however, that the cyclone sample size is small from day 3 or 4 onward resulting in noisy statistics.

Since the IFS analysis resolution is higher than that of the MLWP models, the larger cyclone-maximum 10-m wind speed error and underestimation of cyclone-maximum 10-m wind speed observed in the MLWP models might be expected. We therefore also compare the MLWP forecasts to ERA5 (Figs. 3e,f). For all MLWP models, the cyclone-maximum 10-m wind speed errors compared to ERA5 are typically between 1.5 and 3 m $s^{-1}$ smaller than when compared to IFS analysis throughout the forecast (cf. Figs. 3a,e), with a consistent reduction in the wind bias (cf. Figs. 3b,f). For Pangu-Weather, GraphCast, ECMWF-AIFS, and FuXi (up to 5-day lead time), the cyclone-maximum 10-m wind speed error when compared to ERA5 is similar to the NWP IFS forecast compared to the IFS analysis; however, there remains a systematic underestimation in the cyclone-maximum 10-m wind speed of 1–1.5 m $s^{-1}$. Note that since wind damage can be estimated as proportional to the wind speed cubed (over a threshold) (Klawa and Ulbrich 2003), a forecast error of 1–1.5 m $s^{-1}$ on a 25 m $s^{-1}$ 10-m wind speed would result in a 12%–19% underestimation of the predicted wind damage. In FengWu, Aurora, and FourCastNetv2, the cyclone-maximum 10-m wind speed error when compared to ERA5 (Fig. 3e) remains significantly larger than the NWP IFS forecast compared to IFS analysis (Fig. 3a) at all lead times greater than 2 days, and significant cyclone-maximum 10-m wind speed biases occur from 6 h onward, i.e., the 95% confidence intervals do not overlap (see Fig. S3).

***c. MSLP intensity performance.*** Figure 4a shows MSLP error as a function of forecast lead time, and Fig. 4b shows cyclone MSLP bias as a function of forecast lead time. As expected, for all models, there is an increase in cyclone MSLP error with lead time. The NWP IFS

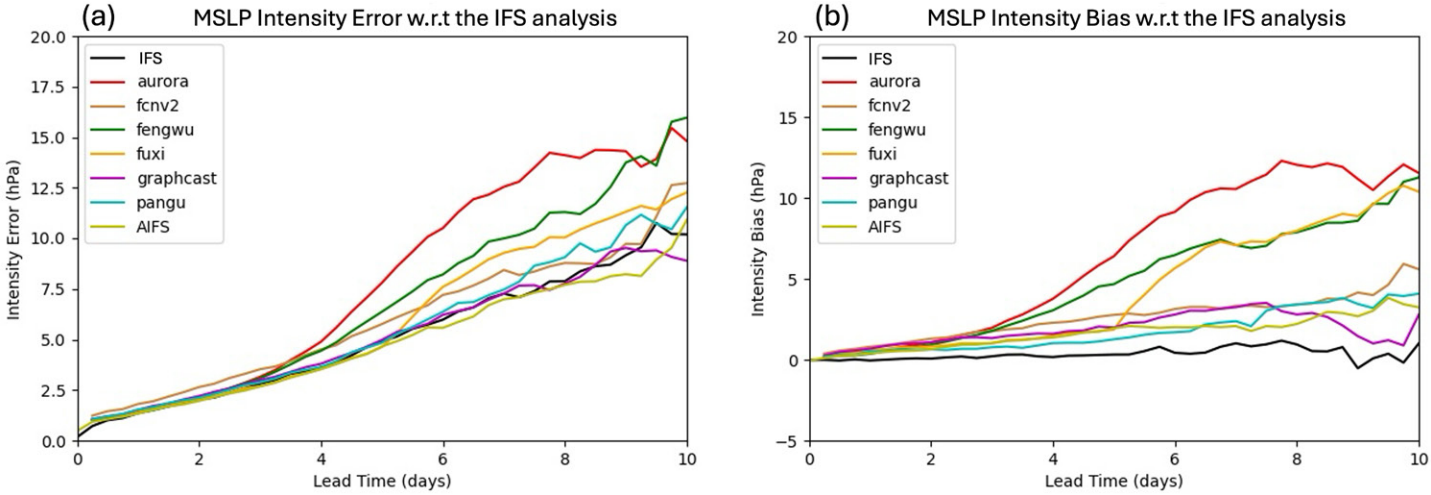


**Fig. 4. NH cyclone MSLP error (hPa) with respect to the IFS analysis as a function of forecast lead time (days). (a) Cyclone MSLP error (hPa) and (b) cyclone MSLP bias (hPa).**

forecast cyclone MSLP error increases at a constant rate, reaching approximately 10 hPa by day 10 (black curve in Fig. 4a). For the first 8 days of the NWP IFS forecast, the cyclone MSLP bias is close to 0 hPa (black curve in Fig. 4b).

For the first 3–4 days, all of the MLWP model forecasts have a similar cyclone MSLP error growth to the NWP IFS forecast (cf. black and colored curves in Fig. 4a). At longer lead times, Pangu-Weather, GraphCast, ECMWF-AIFS, and FourCastNetv2 remain similar to the NWP IFS forecast, but the FengWu, Aurora, and FuXi cyclone MSLP error growth rates become significantly larger than the NWP IFS forecast (at days 3, 4, and 5, respectively), i.e., the 95% confidence intervals do not overlap (see Fig. S4). FuXi shows an increase in cyclone MSLP error at 5 days (as observed in the wind speed error forecasts). For FengWu, Aurora, and FuXi, the cyclone MSLP errors when compared to ERA5 (not shown) are the same as when compared to IFS analysis (Figs. 4a,b). This suggests that the resolution of the training data is not the primary cause of the forecast error.

Despite comparable performance of MLWP models and the NWP IFS forecast model in forecasting cyclone MSLP in terms of absolute error, the MLWP models underestimate cyclone depth at all forecast lead times (Fig. 4b). This underestimation in the depth of cyclone MSLP is, on average, less than 5 hPa for the Pangu-Weather, GraphCast, ECMWF-AIFS, and FourCastNetv2 MLWP models by day 10, but for FengWu, Aurora, and FuXi, the underestimation is approximately twice as large, reaching 10 hPa by day 10. As for cyclone MSLP error, the bias does not change when compared to the lower-resolution ERA5 dataset (not shown), suggesting that resolution is not the dominant factor contributing to the underestimation of cyclone depth.

## 4. Discussion

In this paper, we have applied a feature-based approach to evaluate the position and intensity of midlatitude cyclones in both NWP and MLWP model forecasts. By identifying and tracking the position of over 1000 midlatitude cyclones occurring between 1 October 2023 and 31 March 2024 in the Northern Hemisphere, we systematically evaluate MLWP model midlatitude cyclone forecasts using a feature-based evaluation metric for the first time.

We have shown that MLWP models are able to capture the position of midlatitude cyclone tracks with comparable accuracy to the NWP IFS forecast for lead times out to 10 days. This is consistent with similar studies evaluating the accuracy of tropical cyclone tracks in MLWP models (Bi et al. 2023; Lam et al. 2023). However, in some MLWP models (FengWu, Aurora, and FuXi), the cyclones propagate too slowly. In these models, a cyclone propagation speed

error of 15 km $h^{-1}$ over 24 h would result in a 12-h timing error in the passage of a typical cyclone traveling over a given location.

We have also examined the intensity metrics of the forecast midlatitude cyclones by evaluating their minimum MSLP and peak 10-m wind speeds. It is shown that all of the MLWP models analyzed produced cyclone-maximum 10-m wind speeds that are too weak when compared to IFS analysis and ERA5, whereas the NWP IFS forecast has no bias. This is consistent with the wind speed underestimation that Charlton-Perez et al. (2024) found in their case study analysis of storm Ciarán. This result also supports the conclusion of Bonavita (2024) that MLWP models produce overly smooth predictions and that weather phenomena at spatial scales shorter than 100 km are not well represented. Of particular note is that there is an underestimation of cyclone-maximum 10-m wind speed (1 m $s^{-1}$) present, even at 6-h lead times. We also show that the MLWP cyclone MSLP minima are generally too shallow (by >5 hPa at 10-day lead time) when compared to IFS analysis and ERA5. In contrast, the NWP IFS forecast has similar absolute errors but no bias. This is consistent with studies identifying a weak intensity bias in MLWP forecasts of tropical cyclones (DeMaria et al. 2024; Liu et al. 2024).

Various explanations for the blurring of features by MLWP models have been hypothesized. One explanation is that because the MLWP models are trained to minimize forecast errors, this leads them to produce forecasts that smooth out unpredictable details. All of the MLWP models evaluated in this paper are trained using deterministic data, which provides a single estimate of the atmospheric state. Consequently, these models learn to produce point forecasts, limiting their capacity to capture extreme events. In contrast, real-world weather is inherently nondeterministic, i.e., identical initial conditions can lead to a range of possible outcomes. For example, an MLWP model trained to predict 2-m temperature 3 days ahead using ERA5 as the target may struggle to represent the spread of plausible temperatures under similar atmospheric conditions. Thus, as lead time increases, the models effectively shift their predictions closer to the average of the forecast distribution (Bonavita 2024). Furthermore, autoregressive models, which generate future predictions by feeding previous outputs into the system, can propagate errors over time. If these models smooth out small-scale features early, it can lead to loss of fine-scale features for progressive forecasts, especially for extreme weather events.

The midlatitude cyclone MSLP biases in Pangu-Weather, GraphCast, ECMWF-AIFS, and FourCastNetv2 are smaller than those in FengWu, Aurora, and FuXi. It is difficult to connect model architectural differences (see the appendix) to performance variations. Both GraphCast and ECMWF-AIFS use graph neural networks suggesting that this may be a good approach to take. Aurora is the only model to make use of reanalysis, climate model, and NWP forecast data in its training dataset, which may have resulted in poorer model performance. Finally, the FuXi model displays a jump in cyclone MSLP and wind speed error at day 5, which is likely due to the cascade approach it employs. This is visible, but not so prominent, in the global RMSE evaluation performed by Chen et al. (2023b) (their Fig. 1), illustrating the value of feature-based evaluation for identifying potential issues with model performance. These speculative observations require further work to verify the link between model architecture and forecast accuracy.

ERA5 data are used as the sole training dataset for all but one of the MLWP models evaluated in this study. This lack of diversity could result in models that are overfitted to the characteristics of ERA5. Consequently, the models may perform well during verification against ERA5 data, but this might not reflect true predictive skill in real-world conditions. ERA5 has some known biases, such as an underestimate of near-surface wind speeds and extreme precipitation (Chen et al. 2024). The results presented in this study should thus be considered as an upper estimate of the MLWP model's ability to capture cyclone position and intensity.

Systematically underestimating cyclone-maximum 10-m wind speed will result in poor hazard warnings. If wind warnings do not reflect the true severity, authorities and the public may fail to take necessary precautions, such as securing structures, increasing the risk of damage and injuries. Emergency services may be underprepared, leading to slower responses and delayed recovery efforts. MLWP models are currently not able to capture winds that lead to the largest impacts and thus should not be used to issue wind warnings.

Feature-based evaluation helps to identify specific areas where MLWP models are not yet competitive with traditional NWP. These deficiencies cannot be identified using metrics to assess globally averaged variables such as 500-hPa geopotential height, 850-hPa temperature, 2-m temperatures, or 10-m wind speed. Forecast skill may vary between early and late winter, by cyclone intensity, or by stage in the cyclone life cycle. Due to the need to consider cyclones that are not in the data that the MLWP models have been trained on, in this study, we were limited to analyzing only the most recent winter season (at the time of writing), limiting our sample size and preventing subseasonal, extreme cyclone intensity, or life cycle stage evaluation. This study is one of the first to systematically evaluate MLWP models' ability to forecast extratropical cyclones and will hopefully serve as a foundation for future more detailed assessments. Future work will investigate the life cycle–relative 3D structure of the forecast midlatitude cyclones to further evaluate their physical realism, which is important for building trust and testing the hypothesis that synoptic-scale features become increasingly fuzzy with increasing forecast lead time (Keisler 2022).

**Acknowledgments.** We thank Alan Thorpe for useful discussions about this work. We also thank the modeling groups who made the code for the MLWP models publicly available. H. F. Dacre would like to acknowledge the support of the NERC Climate Change Impact on Midlatitude Cyclone Intensity, Tracks and Impacts (CLIM-CITI) Grant (NE/Y001273/1).

**Data availability statement.** We thank the ECMWF labs team for building the publicly available AI-models library, which enabled us to produce and compare forecasts from the MLWP models. This library can be accessed at https://github.com/ecmwf-lab/ai-models. We also thank the modeling groups who made the code for the ML models publicly available through the following repositories: FourCastNet: https://github.com/NVlabs/FourCastNet; PanguWeather: https://github.com/198808xc/Pangu-Weather; GraphCast: https://github.com/google-deepmind/graphcast. We are also grateful to ECMWF for providing access to operational analysis products to members of the research team through national research accounts held through the Met Office.

## APPENDIX

### MLWP Model Descriptions

Pangu-Weather was developed by Huawei (Bi et al. 2023) and uses an "Earth-specific transformer." In other words, the input variables are embedded into cubic patches containing additional information on their geographical location. Pangu-Weather was trained on 1-h forecasts, and to overcome large errors accumulating during autoregressive rollout, the authors trained additional 3-, 6-, and 24-h forecast models. In this study, we use the 6-h forecast model (applied autoregressively) only.

GraphCast was developed by Google DeepMind (Lam et al. 2023) and differs considerably in architecture from the other six models. Rather than encoding into patches and applying a transformer, GraphCast encodes the input variables into seven polyhedral meshes of varying resolution, using these as input to a graphical neural network. Information can be exchanged between meshes of neighboring resolution. Output is then decoded by reversing the process.

FourCastNetv2, developed mainly by Nvidia (Pathak et al. 2022), also does not use transformers. While they do use an embedding-based autoencoder, the core computation is handled by spherical Fourier neural operators. Neural operators learn the solution operators for partial differential equations and are resolution invariant so they can handle inputs of varying resolutions (rather than spatially invariant, as, e.g., convolutional neural networks). The trained model is fine-tuned on longer lead times to reduce the error of medium-range forecasts.

FengWu was developed by a collaboration of Chinese universities (Chen et al. 2023a) with medium-range forecast skill in mind. They used a "cross modal" approach, where a transformer is applied to each input variable separately rather than together, after embedding, as in many other models. The subsequent representations are then passed through another set of transformers before the encoding is reversed to produce output. They improve medium-range forecasts by fine-tuning the model on its shorter-range forecasts through what they term a "replay buffer."

FuXi, developed at Fudan University (Chen et al. 2023b), uses a standard approach (patch embedding plus transformer), although their transformer layer is a U-Net, allowing some higher-resolution information to bypass the latent space. Once trained, the FuXi model is split into three versions, fine-tuned, respectively, to maximize skill at lead times of 0–5, 5–10, and 10–15 days. This strategy helps minimize accumulation errors and improve forecast accuracy across various time horizons (Chen et al. 2023b; Zhong et al. 2024).

ECMWF-AIFS, developed by ECMWF (Lang et al. 2024), is based on an attention-based graph neural network encoder and decoder and a sliding window transformer processor. This approach allows for flexibility in handling atmospheric data, which is processed on grids that can change resolution across the globe. The processor works on an octahedral reduced Gaussian grid, the same kind of grid that is used in the IFS forecast.

Aurora, developed by Microsoft (Bodnar et al. 2025), also uses a shifted-windows (SWIN) transformer approach. However, unlike all the other models that just train on ERA5, this model is trained on a wide range of data. This data include the MERRA-2 reanalysis, operational NWP analyses, operational NWP forecasts, and climate model output. This training creates a more flexible type of model, called a "foundational model," which can be fine-tuned for various purposes. For example, pretuned versions exist for producing weather forecasts at both 0.25° and 0.1° resolution. In this study, we performed simulations at 0.25° for comparison with the other MLWP forecasts.

## References


Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2023: Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, **619**, 533–538, https://doi.org/10.1038/s41586-023-06185-3.

Bodnar, C., and Coauthors, 2025: A foundation model for the Earth system. *Nature*, **641**, 1180–1187, https://doi.org/10.1038/s41586-025-09005-y.

Bonavita, M., 2024: On some limitations of current machine learning weather prediction models. *Geophys. Res. Lett.*, **51**, e2023GL107377, https://doi.org/10.1029/2023gl107377.

Charlton-Perez, A. J., and Coauthors, 2024: Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciarán. *npj Climate Atmos. Sci.*, **7**, 93, https://doi.org/10.1038/s41612-024-00638-w.

Chen, K., and Coauthors, 2023a: Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. arXiv, 2304.02948v1, https://doi.org/10.48550/arXiv.2304.02948.

Chen, L., X. Zhong, F. Zhang, Y. Cheng, Y. Xu, Y. Qi, and H. Li, 2023b: FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate Atmos. Sci.*, **6**, 190, https://doi.org/10.1038/s41612-023-00512-1.

Chen, T.-C., F. Collet, and A. Di Luca, 2024: Evaluation of ERA5 precipitation and 10-m wind speed associated with extratropical cyclones using station data over North America. *Int. J. Climatol.*, **44**, 729–747, https://doi.org/10.1002/joc.8339.

DeMaria, M., J. L. Franklin, G. Chirokova, J. Radford, R. DeMaria, K. D. Musgrave, and I. Ebert-Uphoff, 2024: Evaluation of tropical cyclone track and intensity forecasts from Artificial Intelligence Weather Prediction (AIWP) models. arXiv, 2409.06735v1, https://doi.org/10.48550/arXiv.2409.06735.

Froude, L. S., L. Bengtsson, and K. I. Hodges, 2007: The prediction of extratropical storm tracks by the ECMWF and NCEP ensemble prediction systems. *Mon. Wea. Rev.*, **135,** 2545–2567, https://doi.org/10.1175/mwr3422.1.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, https://doi.org/10.1175/2009waf2222269.1.

Gray, S. L., and A. Volonté, 2024: Extreme low-level wind jets in Storm *Ciarán*. *Weather*, **79**, 384–389, https://doi.org/10.1002/wea.7620.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hodges, K. I., 1994: A general method for tracking analysis and its application to meteorological data. *Mon. Wea. Rev.*, **122**, 2573–2586, https://doi.org/10.1175/1520-0493(1994)122<2573:agmfta>2.0.co;2.

——, 1995: Feature tracking on the unit sphere. *Mon. Wea. Rev.*, **123**, 3458–3465, https://doi.org/10.1175/1520-0493(1995)123<3458:ftotus>2.0.co;2.

Keisler, R., 2022: Forecasting global weather with graph neural networks. arXiv, 2202.07575v1, https://doi.org/10.48550/arXiv.2202.07575.

Klawa, M., and U. Ulbrich, 2003: A model for the estimation of storm losses and the identification of severe winter storms in Germany. *Nat. Hazards Earth Syst. Sci.*, **3**, 725–732, https://doi.org/10.5194/nhess-3-725-2003.

Lam, R., and Coauthors, 2023: Learning skillful medium-range global weather forecasting. *Science*, **382**, 1416–1421, https://doi.org/10.1126/science.adi2336.

Lang, S., and Coauthors, 2024: AIFS-ECMWF's data-driven forecasting system. arXiv, 2406.01465v2, https://doi.org/10.48550/arXiv.2406.01465.

Liu, C.-C., and Coauthors, 2024: Evaluation of five global AI models for predicting weather in eastern Asia and western Pacific. *npj Climate Atmos. Sci.*, **7**, 221, https://doi.org/10.1038/s41612-024-00769-0.

Pathak, J., and Coauthors, 2022: Fourcastnet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. arXiv, 2202.11214v1, https://doi.org/10.48550/arXiv.2202.11214.

Priestley, M. D., H. F. Dacre, L. C. Shaffrey, K. I. Hodges, and J. G. Pinto, 2018: The role of serial European windstorm clustering for extreme seasonal losses as determined from multi-centennial simulations of high-resolution global climate model data. *Nat. Hazards Earth Syst. Sci.*, **18**, 2991–3006, https://doi.org/10.5194/nhess-18-2991-2018.

Selz, T., and G. C. Craig, 2023: Can artificial intelligence-based weather prediction models simulate the butterfly effect? *Geophys. Res. Lett.*, **50**, e2023GL105747, https://doi.org/10.1029/2023gl105747.

Thompson, V., S. Y. Philip, I. Pinto, and S. F. Kew, 2024: The influence of the Atlantic Multidecadal Variability on storm Babet-like events. EGUsphere, https://doi.org/10.5194/egusphere-2024-1136.

Vitart, F., M. A. Balmaseda, L. Ferranti, and M. Fuentes, 2022: The next extended-range configuration for IFS cycle 48r1. *ECMWF Newsletter*, No. 173, ECMWF, Reading, United Kingdom, 21–26, https://www.ecmwf.int/sites/default/files/elibrary/102022/20502-newsletter-no-173-autumn-2022_0.pdf.

Zhong, X., L. Chen, J. Liu, C. Lin, Y. Qi, and H. Li, 2024: FuXi-extreme: Improving extreme rainfall and wind forecasts with diffusion model. *Sci. China Earth Sci.*, **67**, 3696–3708, https://doi.org/10.1007/s11430-023-1427-x.