npj | climate and atmospheric science

Article

Published in partnership with CECCR at King Abdulaziz University

https://doi.org/10.1038/s41612-024-00769-0

Evaluation of five global AI models for predicting weather in Eastern Asia and Western Pacific

Check for updates

Cheng-Chin Liu[1], Kathryn Hsu[1] ✉, Melinda S. Peng[2], Der-Song Chen[1], Pao-Liang Chang[1], Ling-Feng Hsiao[1], Chin-Tzu Fong[1], Jing-Shan Hong[1], Chia-Ping Cheng[1], Kuo-Chen Lu[1], Chia-Rong Chen[1] & Hung-Chi Kuo[3]

Recent development of artificial intelligence (AI) technology has resulted in the fruition of machine learning-based weather prediction (MLWP) systems. Five prominent global MLWP model, Pangu-Weather, FourCastNet v2 (FCN2), GraphCast, FuXi, and FengWu, emerged. This study conducts a homogeneous comparison of these models utilizing identical initial conditions from ERA5. The performance is evaluated in the Eastern Asia and Western Pacific from June to November 2023. The evaluation comprises Root Mean Square Error and Anomaly Correlation Coefficients within the designated region, typhoon track and intensity predictions, and a case study for Typhoon Haikui. Results indicate that FengWu emerges as the best-performing model, followed by FuXi and GraphCast, with FCN2 and Pangu-Weather ranking lower. A multi-model ensemble, constructed by averaging predictions from the five models, demonstrates superior performance, rivaling that of FengWu. For the 11 typhoons in 2023, FengWu demonstrates the most accurate track prediction; however, it also has the largest intensity errors.

Turing's[1] pioneer work "Computer Machinery and Intelligence" introduced The Turing Test, used by experts to measure computer intelligence even up to date. The term "artificial intelligence" (AI) was coined in 1956 at a Dartmouth summer workshop and machine learning (ML) was referred to by Arthur Samuel[2]. The advancements in AI/ML have experienced intermittent periods of stagnation over the past few decades. Significant progress in AI applications recently stems from vast datasets, rapid computational capabilities, and the availability of improved AI tools. Achievements of AI includes speech and vision recognizers, autonomous vehicles, cognitive computing, and expert systems, etc. These applications are increasingly penetrating various scientific and engineering communities. Many ML algorithms can be thought of as optimizing a nonlinear regression, with deep learning utilizing an extremely high-dimensional model. ML has proven to be an excellent tool for addressing complex, nonlinear, or stochastic challenges encountered in fields such as physics and Earth science.

Numerical weather prediction (NWP) models using nonlinear primitive equations with parameterizations accounting for sub-grid scale physical processes have been developed in the last 70 years. The success of NWP comes from supercomputing capabilities allowing high resolutions, improvements of physical parameterization, data assimilation strategy, and satellite retrievals. Recently, the Earth science community is adept at embracing AI/ML, as many AI concepts align closely with meteorology challenges, particularly data assimilation. Progress in AI meteorology has seen exponential growth since 2016, driven by achievements in ML techniques and sufficient computing resources, such as GPUs. There are many AI applications in Earth science such as PDE solving[3,4] and NWP post processing[5–7].

Recently, there was a growing interest in utilizing AI/ML techniques to build ML-based (data-driven) weather and climate prediction (MLWP) models directly from atmospheric data. A variety proof of concept studies based on a few atmospheric variable comparisons has reveal that MLWP may have a great potential competing with traditional physical-based numerical weather prediction models[8–12]. However, the rather coarse spatial resolution data (~5°) used in these studies limited their forecast performances and applications. A chronicle review of the ML weather and climate modeling is given by de Burgh-Day and Leeuwenburg[13].

Significant breakthrough emerged starting in 2022 for the development of global MLWP models applying to weather and climate predictions. FourCastNet system (FCN)[14] is the first MLWP model producing 0.25° resolution forecasts using ECMWF ERA5 reanalysis[15] as its training data. The system applying Vision Transformer (ViT)[16] with the Fourier Neural Operator (FNO)[17] and Adaptive Fourier Neural Operators (AFNO)[18] for

[1]Central Weather Administration, Taipei, Taiwan. [2]University of Colorado Colorado Springs, Colorado Springs, CO, USA. [3]National Taiwan University, Taipei, Taiwan. ✉e-mail: r94229018@gmail.com

efficient computation in long-range dependencies in spatial-temporal data. A vision transformer is a deep learning model that breaks down an image into patches, processes them using transformers, and aggregates the information for object detection. FourCastNet was subsequently upgraded to version 2, FCN2, that uses Spherical Harmonics Neural Operators (SFNO) to build a more stable autoregressive model for weather prediction[19]. SwinRDM[20] is the first MLWP system to outperform ECMWF's deterministic operational forecasting system, IFS-HRES, in 5-day forecasts at a spatial resolution of 0.25°. Pangu-Weather[21] acquires promising medium-range performance that surpasses IFS-HRES with a multi-timescale model combination strategy based on 3D Earth-Specific Transformers (3DEST) in which the upper-air variables and the surface variables are embedded into a single deep network. The 3D data are propagated through an encoder–decoder architecture derived from the Swin transformer[22], a variant of the vision transformer. GraphCast[23] takes the approach of applying the graph neural networks (GNN)[12] to a six-layer icosahedron grids with increasing resolution globally in which a set of objects and the connections between them are expressed as a graph. A 12-step autoregressive fine tuning is employed as the strategy for increasing the long-lead prediction accuracy. The model is more accurate in predicting 90% of the atmospheric variables compared with the IFS-HRES.

FuXi[24] model backbone is based on Swin transformer v2[25] with 3D cube embedding (called U-transformer) and is designed as a cascade of models (short for 0 ~ 5 days, medium for 5 ~ 10 days, and long for 10 ~ 15 days) optimized for different forecast lead times that is similar to the multiple time steps used in Pangu-Weather[21]. FengWu forecast system is constructed from a multi-modal and multi-task perspective in which each atmospheric state variable is treated as an individual modal and a cross-modal fuser transformer is applied to connect them[26]. To solve the long-lead prediction issue, the replay buffer mechanism is used, which is inspired by the reinforcement learning study[27]. The replay buffer stores the predicted results from previous optimization iterations and uses them as the current model's input, which mimics the intermediate input error during the autoregressive inference stage.

Bouallegue et al.[28] conducted a comprehensive comparison between Pengu-Weather simulation and ECMWF operational IFS using the same initial conditions. The results are very promising for the MLWP model, with comparable accuracy from Pengu-Weather for both global metrics and extreme events, when verified against the IFS analysis and synoptic observations. Meanwhile, overly smooth forecasts, increasing bias with forecast lead time, and poor performance in predicting tropical cyclone intensity are identified as current drawbacks of ML-based forecasts. Charlton-Perez et al.[29] conducted a quantitative evaluation for FCN, Pangu-Weather, GraphCast, and FCN2 on the prediction of Storm Ciaran (2023), which has caused significant casualties and damage in Europe. The simulations of four MLWP models accurately capture the synoptic-scale structure of the winter cyclone including the position of the cloud head, shape of the warm sector and location of the warm conveyor belt jet. Meanwhile, all of the MLWP models underestimate the peak amplitude of winds associated with the storm. This appears to be the commonality of existing machine learning models with resolution limited by ERA5 as their training data.

The new developments on MLWP systems initiated a new era on weather predictions[30]. While all the aforementioned MLWP models have demonstrated performance compatible with or superior to conventional NWP models, and some comparisons between individual models have been analyzed, a homogeneous, comprehensive comparison among them is lacking. Additionally, the illustrations of tropical cyclone prediction skills, which serve as examples of extreme weather events, are selectively presented in their major publications.

An essential aspect of global weather prediction models is their role in generating initial and boundary conditions for regional weather forecast systems[31]. High resolutions in regional models facilitates representing of complex multi-scale processes crucial for high-impact weathers that directly affect human society directly. Biases and model errors present in the global forecast fields are thus inherited by regional models[32]. The influence of uncertainty in initial conditions usually diminishes over the course of the simulation length, whereas the impact of variations in lateral boundary forcing does not exhibit a clear trend.

As an operational weather prediction agency, the Central Weather Administration (CWA) is responsible for weather forecasts in Taiwan, encompassing a broad spectrum of temporal scales ranging from nowcasting to climate predictions. We utilize global forecast fields to drive our regional models, which feature high resolutions specifically tailored for the East Asia and western Pacific region[33]. The performance of the driving global forecast model significantly influences the accuracy of a limited area significantly[34]. Therefore, performance of the driving global model is critical for the success of our regional models. This study aims to contribute to this understanding.

Result

Synoptic-scale predictions

In this section we present the performance of the five MLWP models and IFS in the area shown in Fig. 1, covering the evaluation period from June to November 2023. Note that IFS is the operational NWP system at ECMWF and our MLWP simulations used ERA5 as the initial conditions, which is different from the IFS. We compute the standard metrics of latitude-

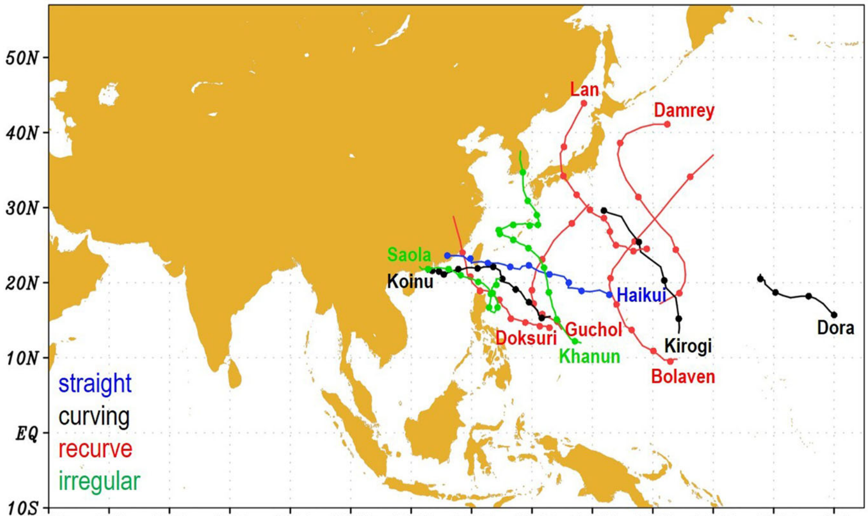

Fig. 1 | The regional domain for evaluating the five MLWP models includes the best tracks of the 11 typhoons that occurred during June to November 2023. The best tracks are categorized as straight (blue), curving (black), recurve (red), or irregular (green), with the names of the typhoons indicated beside their respective tracks. Track data interval is 6-h, and the dots shows 00 UTC of the day.

weighted Root Mean Square Error (RMSE) and Anomaly Correlation Coefficient (ACC), commonly used in the meteorological community. ACC measures the spatial correlation between a forecast anomaly relative to climatology and is widely used for synoptic-scale prediction assessment.

Figure 2 shows the RMSE and ACC for various variables, including the 500 hPa geopotential height, temperature at 850 hPa and at 2 m height, and zonal component of the wind at 10 m height. Based on RMSE, the best model is FengWu, followed by FuXi and GraphCast, with FCN2 trailing behind, and Pangu-Weather ranking last. IFS with a resolution of 9 km, lies in the middle of the group. Near the surface, there is a slight variation between FCN2 and FuXi. It is worth noting that ERA5 contains more observational data post-analysis data, while IFS uses the operational real-time analysis without the post-analysis data. This is reflected by the larger RMSE of IFS compared to the other five models at the initial time. With this in mind, we focus on the error trends when comparing IFS with the other models rather than the absolute magnitudes. In general, the slope of the RMSE from IFS aligns with the others, except for Pangu-Weather, which exhibits a larger upward trend. Regarding ACC (Fig. 2e–h), the performance of the five MLWP models is similar to their RMSEs, with FengWu leading and Pangu-Weather trailing, and IFS falling in the middle of the group.

Weather forecasting inherently carries uncertainties due to the chaotic nature of weather systems and imperfect initial conditions. To address this uncertainty, ensemble forecasting is employed through perturbing the initial conditions and physical parameterizations[35,36]. There is also a multi-model ensemble approach that is widely recognized for its improved performance over individual models in various fields, including weather forecasting, climate modeling, and machine learning[37]. Hagedorn, et al.[38] provided a comprehensive analysis of why multi-model ensembles outperform individual models in seasonal forecasting. A multi-agency effort was established for the North American Multi-Model Ensemble (NMME) prediction system targeted at the seasonal forecasts that showed better performance compared to individual models[39].

Due to their exceptional computational efficiency, MLWP models have been proposed as ideal for ensemble predictions from a single-model ensemble perspective. Here, we investigate the potential benefits of a multi-model ensemble approach by taking a simple multi-model ensemble strategy. We compute the RMSE and ACC of the ensemble by averaging forecast fields from all five MLWP models. Notably, when averaging the forecasts from the first three low-scoring models, the ensemble outperforms the individual models (figure not shown). As more models are added to the ensemble, the additional benefits become less pronounced. Ultimately, the performance of the ensemble comprising all five models approaches that of the best-performing model, which is FengWu in this study. The primary advantage of using a multi-model ensemble is the reduction in the error range. Specifically, for tropical cyclone track predictions, the ensemble consistently stays within the error range, preventing outliers from individual models. This point will be addressed further when evaluating typhoon predictions. Note that in this simple multi-model ensemble strategy, the uncertainty associated with initial conditions from different models is not accounted for. However, the ensemble strategy still demonstrates its conceptual advantage.

Another important metric for assessing model performance is the model bias, which represents the difference between the mean forecast state and the verification. In this study, we examine the position and strength of the western Pacific subtropical high (WPSH) system, a key feature of Hadley circulation's substance. The variation of the WPSH is primarily influenced by central Pacific cooling/warming and positive atmosphere-ocean feedbacks between the WPSH and the Indo-Pacific warm pool oceans[40]. The position and strength of this system significantly influences various regional weather features such as monsoon circulations and TC movements. Interactions between a TC and the WPSH can lead to changes in cyclone trajectories, making them more prone to recurvature or maintaining a straight path based on the strength and position of the WPSH[41].

We use the 500 hPa geopotential height contour as a measurement of the system, with the 5880 m line serving as a common reference in local operational communities and predicting of it from different models can be easily assessed. Figure 3 illustrates the average of the 168 h forecasted 5880 m contours in summer (JJA) and fall (SON) from each MLWP model, alongside the three-month mean from ERA5 serving as the verification. There is a seasonal migration of the WPSH westward from summer to fall.

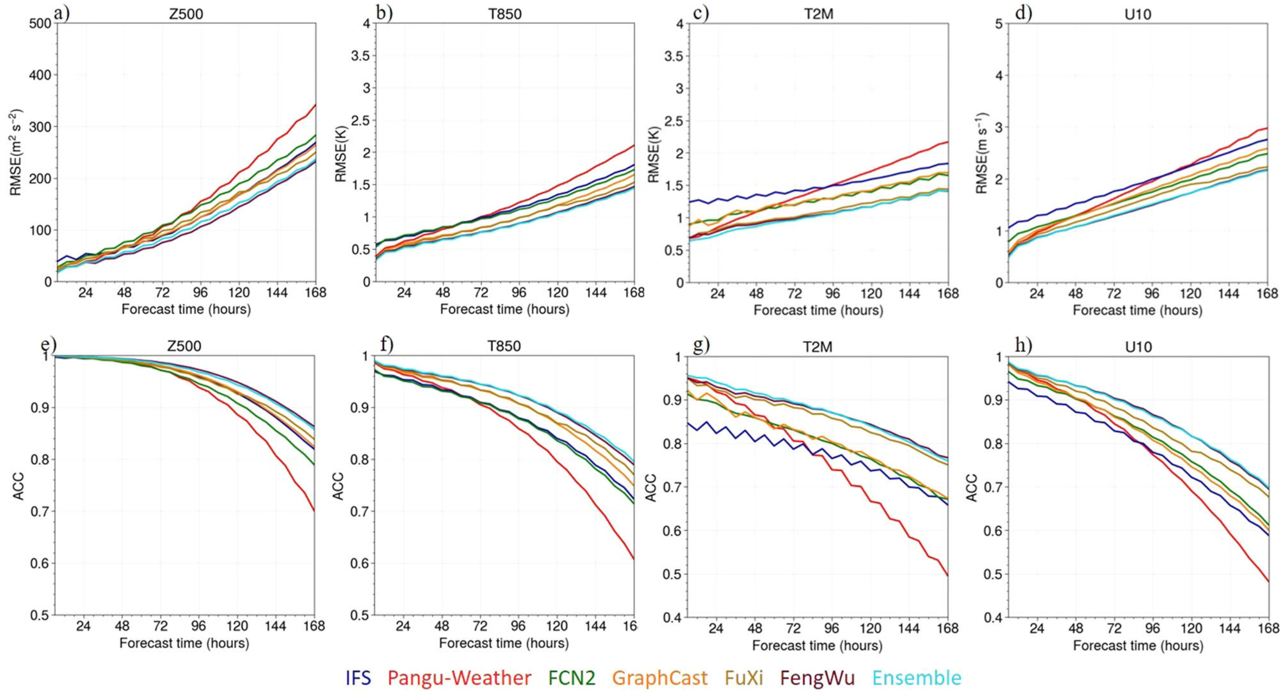

Fig. 2 | The evaluation includes RMSE and ACC of five MLWP models, their ensemble, and IFS. RMSE for **a** 500 hPa geopotential, **b** 850 hPa temperature, **c** temperature at 2 m height, and **d** zonal component wind at 10 m height of Pangu-Weather (red), FCN2 (green), GraphCast (orange), FuXi (light brown), FengWu (brown), the ensemble (light blue) of the five MLWP models, and IFS (blue), **e–h** ACC correspond to the same variables.

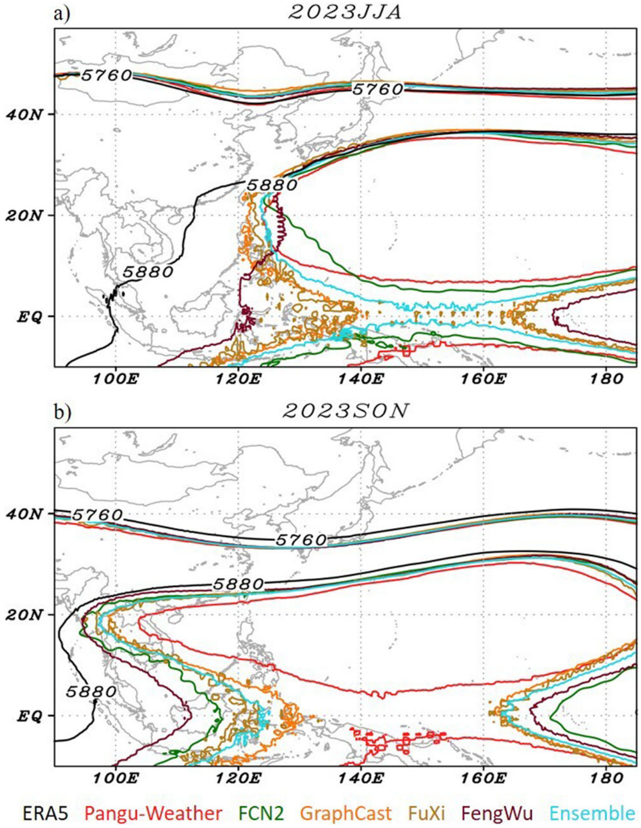

Fig. 3 | The analysis (ERA5) and 168 h predicted 5880 m and 5760 m geopotential height at 500 hPa for the five MLWP models and their ensemble. Contours show the period of (**a**) June, July, and August and (**b**) September, October, and November 2023 of Pangu-Weather (red), FCN2 (green), GraphCast (orange), FuXi (light brown), FengWu (brown), the ensemble (light blue) of those MLWP models mentioned above, and ERA5 (black).

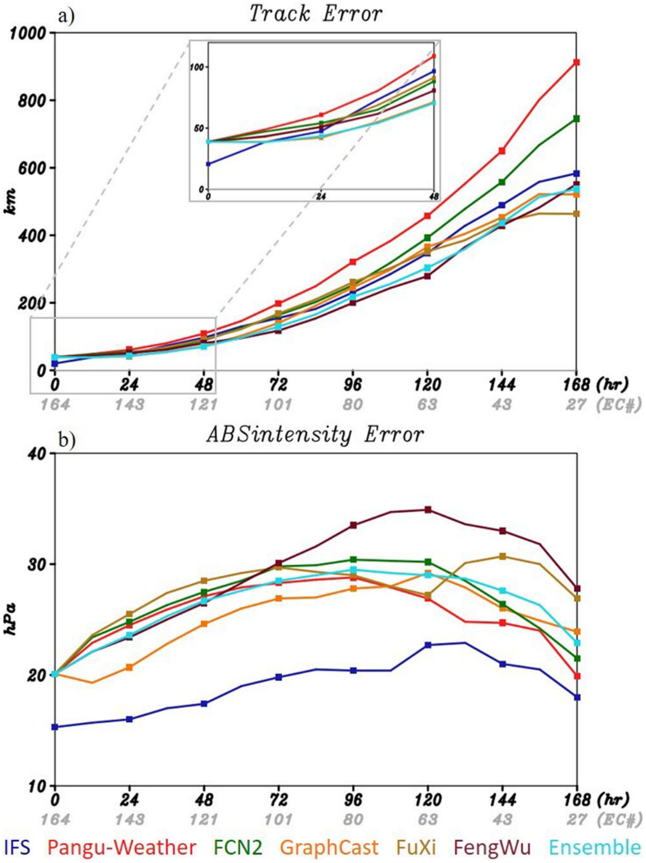

Fig. 4 | Averaged track and absolute intensity errors for the five MLWP models, their ensemble, and IFS. a track errors (km), **b** absolute intensity errors (hPa). The total number of cases are listed at the bottom. Colors according to the models are the same as Fig. 2. The filled squares show averaged track and intensity error every 24 h starting from the initial time.

In both periods, the 5880 m contour line from ERA5 encompasses all lines from the MLWP models, indicating that all these models exhibit a weak bias for the WPSH. Among them, Pangu-Weather shows the largest weak bias, while FengWu exhibits the smallest bias, consistent with their RMSE and ACC scores. This observation holds significant importance in the maritime continent, where the Madden-Julian Oscillation (MJO)[42,43] is active, leading to large variabilities and far-reaching influences across the globe.

Additionally, the 5760 m contour lines at higher latitudes from the five MLWP models are more closely clustered in both seasons. In general, tropical regions present greater challenges in weather prediction compared to mid-latitudes[44], primarily due to more vigorous convective activities[45]. The simple analysis presented here suggests that MLWP models may also face greater challenges at lower latitudes, likely due to limitations inherited from their training data, which is the reanalysis relying on NWP model as its backbone.

Typhoon predictions

We assess the performance of the five MLWP models alongside IFS in predicting the 11 typhoons occurring between June and November 2023, excluding three short-lived ones (see best tracks in Fig. 1). The positions of the typhoons are identified by the storm center, determined as the minimum sea-level pressure in the forecast fields. Typhoon best track data provided by CWA was utilized for this analysis, and they are similar to those from the International Best Track Archive for Climate Stewardship (IBTrACS) database[46,47]. Track and intensity errors from the 7 models, including the five MLWP models, their ensemble, and IFS are depicted in Fig. 4. Note that the averages are taken by the forecast lead time, which likely represent different stages of the life cycle for individual storms. In a case study presented later, we will illustrate how predictions for a storm can vary significantly in different stage.

Among the compared MLWP models, Pangu-Weather exhibited the largest track error, followed by FCN2, GraphCast, FuXi, and FengWu, progressively with smaller error up to 144 h. Subsequently, FengWu's track error is higher than FuXi and slightly higher than GraphCast at 168 h. Overall, errors of IFS lie roughly in the middle of the group even though it has the smallest position error at the initial time (in the zoomed inlet). The ensemble performance was very close to FengWu. The benefit of using a multi-model ensemble is to reduce the range of error in tropical cyclone track prediction in which the ensemble is always within the range so that one would not get outliers (to be discussed more using Table 1). Comparing Fig. 4 with Fig. 2 indicates that, in general, track prediction performance aligns with the ranking by RMSE and ACC. Because TCs are steered, to first order, by the large-scale flow at mid-tropospheric levels[48]. Therefore, a weather prediction model with better ACC and RMSE scores usually also has better TC track prediction.

In terms of intensity prediction, the absolute errors are displayed to avoid cancellation of positive and negative relative errors, and all models exhibit weak biases. Although FengWu exhibited the smallest overall track errors, it had the largest intensity errors. Meanwhile, GraphCast and Pangu-Weather tied for the lowest intensity errors. It is noteworthy that IFS exhibited a much lower initial intensity bias compared to the five MLWP models, which used ERA5 as their initial conditions. Nevertheless, the trend of intensity errors was similar across all models. When comparing Pangu-Weather's performance with IFS on TC intensity prediction, Bouallèguea et al.[28] also demonstrated that Pangu-Weather performed poorly than IFS.

Table 1 | Comparison of TC track and intensity forecasts at 96 h by IFS, Pangu-Weather (PW), FCN2, GraphCast (GC), FuXi, FengWu (FW), and IFS for 11 typhoons in 2023 in the western North Pacific

Storm name	No of cases	Track types	Storm maximum intensity (hPa)	Best track performer @ 96 h (km)	Worst track performer @ 96 h (km)	Best intensity performer @ 96 h (hPa)	Worst intensity performer @ 96 h (hPa)	Average track error from all models @ 96 h (km/No of cases)	Average intensity error from all models @ 96 h (hPa/No of cases)
Haikui	15	straight	945	FW (41.3)	IFS (1012.5)	FuXi (5.2)	P-W (46.6)	442.4/48	26.0/42
Dora @ 72 h	7	curving	975	FW (63.7)	P-W (195.7)	P-W (16.0)	FW (16.8)	111.0/3	14.9/3
Kirogi @ 72 h	8	curving	985	GC (19.3)	P-W (458.9)	P-W (6.5)	FuXi (9.7)	287.7/8	7.2/8
Koinu	19	curving	930	P-W (49.9)	FW (441.8)	FuXi (27.7)	GC (63.3)	179.1/61	44.6/55
Guchol	13	recurve	960	FW (22.3)	IFS (533.8)	P-W (0.4)	FW (23.0)	162.4/30	8.7/30
Doksuri	14	recurve	935	FuXi (0.0)	FW (642.4)	GC (0.2)	FuXi (50.8)	158.9/42	32.2/42
Lan	19	recurve	940	IFS (24.3)	FuXi (599.3)	P-W (0.1)	FCN2 (41.6)	268.4/66	14.9/66
Damrey	8	recurve	985	IFS (102.5)	FuXi (447.5)	GC (10.9)	FW (20.0)	303.6/6	14.1/6
Bolaven	14	recurve	900	P-W (38.0)	IFS (575.8)	GC (1.1)	FuXi (93.7)	222.8/36	53.6/36
Khanun	28	irregular	930	FW (19.4)	P-W (760.1)	FuXi (0.2)	FW (60.5)	256.2/120	15.5/120
Saola	19	irregular	915	P-W (43.2)	GC (618.8)	FCN2 (0.4)	FW (79.3)	263.6/66	46.8/66
Averages	----- -	--------- -	945.5	(38.5)	(571.5)	(2.9)	(45.9)	251.1/486	27.9/474

IFS is excluded from the intensity comparison (columns 7 and 8) but included in the average result (column 10). For Typhoon Dora and Kirogi, due to their short lifespans, the comparison is based on the 72 h forecasts. The colors used for the track types are the same as those shown in Fig. 1, while the colors used for the models are the same as those shown in Fig. 2.

Their Fig. 8b suggests that the superiority of IFS may stem from better initial (higher) intensity due to its higher resolution. However, the slope of the error does not necessarily indicate that IFS predicts intensity changes better than Pangu-Weather. An interesting observation is that all models exhibited a decreasing intensity error from 120 to 168 h. This is because TCs near the end of their life cycle are usually weak and the ranges of predicted intensities are smaller.

We conducted two sets of statistical significance tests for typhoon track and intensity predictions: one comparing Pangu-Weather (PW) with the other models, and one comparing FengWu (FW) with the other models. Since the typhoon track and intensity errors do not follow a normal distribution, we adopted the Mann-Whitney U test[49,50] to assess statistical significance. The results indicate a 95% confidence level in the comparison of track errors between the selected model (PW or FW) and other MLWP models. Additionally, there is a 90% confidence level in the difference in TC intensity errors between FW and the other MLWP models.

It is widely acknowledged that TC track predictions can exhibit significant diversification for individual typhoons. Instead of displaying predicted tracks for all typhoons from all models, Table 1 offers further statistical insights into our evaluations. We subjectively classify track characteristics into 4 categories: straight, curving, recurve, and irregular, as illustrated in Fig. 1. Evaluation is conducted at 96 h due to its substantially higher number of verifications available, as the number of cases decreases significantly with longer lead times. The best-performing model for track prediction is FengWu, with one typhoon prediction leading in all 4 categories, consistent with the averaged track errors shown in Fig. 4. Interestingly, despite Pangu-Weather exhibiting the largest average track error, it performs the best for three individual typhoons. Additionally, IFS achieves the highest scores for two typhoons in the recurving category, while GraphCast and FuXi each lead in one storm.

While TC intensity prediction may not be the primary focus of a global model, it is still pertinent to discuss the performance of the five models. FengWu, despite excelling in track prediction, exhibits a reversal in the performance for intensity prediction, failing to rank as first for any individual typhoon—a result consistent with its largest absolute average errors depicted in Fig. 4b. Conversely, Pangu-Weather leads in intensity prediction for 4 typhoons, while both FuXi and GraphCast lead for 3 typhoons each. It's worth noting that the numbers shown in column 7 of Table 1 for the best-performing model are significantly smaller than the averages displayed in Fig. 4b, which range from 25 to 35 hPa. These more detailed performance assessments on track and intensity predictions from different models further underscore the diversification of TC predictions. The IFS is not included in the intensity evaluation (columns 7 and 8) due to its advantage of low intensity bias at the initial time, attributed to its high resolution.

Case Study-Typhoon Haikui (2023)

Typhoon Haikui (blue track in Fig. 1) was the first major storm to hit Taiwan since Bailu in 2019. Behind Typhoon Saola (left green track in Fig. 1), Haikui began its life as a broad low-pressure system on August 27 near the Northern Mariana Islands. The system intensified to a tropical storm the next day and was named Haikui by the Japan Meteorological Agency (JMA). In the subsequent days, Haikui reached a tropical storm strength and eventually became a typhoon, before making landfall near Taitung City, Taiwan, on September 3. Haikui also enhanced the southwest monsoon in the Philippines, causing extensive rainfall in Luzon. As it stalled over Pearl River Delta in China, the remnants of Haikui induced torrential rain in Hong Kong resulting in the issuance of a Black Rainstorm Signal for 16 h, the longest duration ever since the rainstorm warning system was implemented in 1992. Overall, Haikui caused US$2.31 billion worth of damage during its onslaught.

As listed in Table 1, the prediction of the track for Typhoon Haikui was the most inaccurate (average and individual errors) among all typhoons in the western North Pacific in 2023. We present the track predictions from 12 UTC on August 28 to 12 UTC on September 4 for Typhoon Haikui from individual models, including IFS and the five MLWP models (Fig. 5). Among them, IFS exhibited the largest average track error, exceeding 1000 km at 96 h with a significant poleward bias, mainly from the early stage (Fig. 5f). Both GraphCast and FCN2 (Fig. 5b, c) also showed substantial track errors for the first two watches, with their predicted tracks resembling that of IFS. Meanwhile, FengWu's predicted tracks were closely aligned with the best track (Fig. 5e), with the averaged track error of only 41 km at 96 h, making it the best-performing model for Haikui's track prediction. The second best is Fuxi (Fig. 5d), followed by Pengu-Weather (Fig. 5a).

Next, we examine the potential link between the predicted tracks and the position of the WPSH. For consistency and ease of comparison among model predictions, the WPSH system is represented by the 5880 m height at

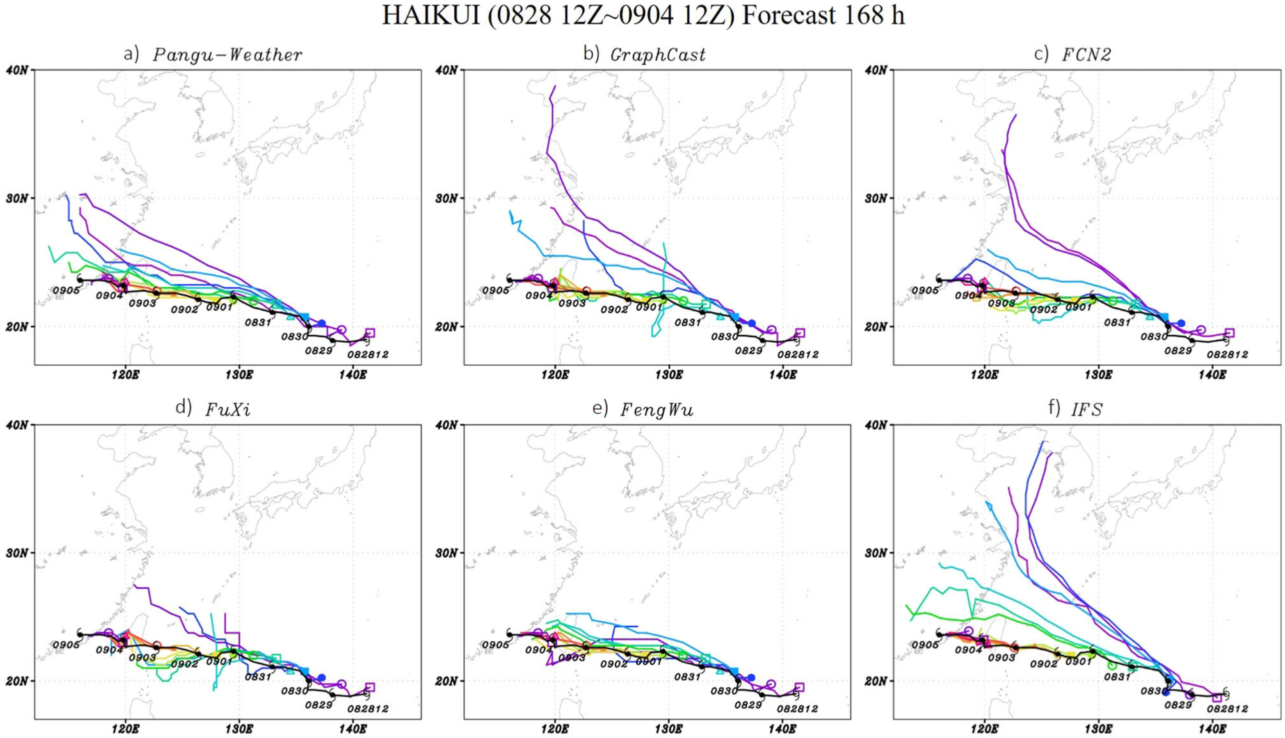

Fig. 5 | The best (CWA) and the forecast tracks for Typhoon Haikui to 168 h from five MLWP models and IFS during period from 12 UTC 28 August to 12 UTC 4 September, 2023. The black line with typhoon marks represents the best track, while lines with other colors and marks indicate forecasts at different initial times of (**a**) Pengu-Weather, (**b**) GraphCast, (**c**) FCN2, (**d**) FuXi, (**e**) FengWu, and (**f**) IFS. The open typhoon mark shows the best track location at 12 UTC 28 August while closed typhoon marks show the best track locations at 00 UTC each day with date beside. Marks in other colors indicate the forecast initial time.

500 hPa, as referenced for its seasonal verification (Fig. 3). Figure 6 illustrates two 96 h forecasts, starting at 00 UTC 29 August and then 6 h later at 06 UTC on 29 August, from IFS and the five MLWP models. For the first forecast, all models exhibit a poleward bias in their predictions, with the IFS showing the largest and FengWu and Fuxi the smallest (Fig. 6a). While only the 96-h geopotential pattern are displayed (Fig. 6b), the distribution of the western edges of the WPSH generally aligns with their individual tracks. In the second example, starting 6 h later, the WPSH predicated by the five MLWP models have all migrated westward, while in IFS, the system lagged and remained close to its position predicted 6 h earlier (blue lines in Fig. 6b, d). Following the movement of the WPSH, all five MLWP models adjusted their predicted tracks to be mainly westward, in line with the best track, while the IFS maintained its northwestward track (Fig. 5f). The predicated tracks by IFS are very similar in these two forecasts, showing a significant poleward bias (blue lines in Fig. 6a, c), which contributed to its overall large track error for Haikui (Table 1).

Tropical cyclone movement can be influenced by many surrounding synoptic-scale and mesoscale features[51,52]. The analysis presented here only provide the first-order influence by the WPSH, as commonly assessed by operational weather prediction centers in eastern Asia and western Pacific region. As shown in Fig. 5f, the IFS eventually adjusted its predicted track to be mainly westward after the first four bad forecasts. Further in-depth analysis is required to fully understand the dynamics behind it.

High-resolution data and model configuration are critical for representing and simulating complex mesoscale phenomena. Thus, we rely on regional models for local area weather predictions, and for extreme weather systems such as typhoons. These models can offer significant benefits where local weather is influenced by factors like islands, coastlines, topography, and land/sea contrast. While MLWP models have demonstrated impressive performance in synoptic-scale systems, their effectiveness can vary for systems involving multiple scales (such as typhoons) and remain to be thoroughly examined. The TWRF (Typhoon WRF) system is a regional NWP system developed by CWA based on the ARW WRF model[53,54], dedicated for typhoon prediction in East Asia with a focus on Taiwan. TWRF is two-way nested regional model with 15 km and 3 km resolutions in the outer and inner domain. Previous studies have verified its superior performance[33,34], and its predictions are currently displayed on the NOAA hurricane analysis and forecast system website. In the following analysis of Haikui during its passage over Taiwan, we also include the prediction from TWRF (15 km version) for comparison.

Rainfall induced by a TC on Taiwan is intricately linked to its track, which determines how the typhoon interacts with the island's complex terrain[55]. The case study starts on 12 UTC 2 September and the two-day accumulated rainfalls ending on 12 UTC 4 September will be evaluated. Here, we focus on the rainfall prediction from two MLWP models, GraphCast and FuXi, both of which include precipitation forecasts in their outputs. Also included in the comparison are two NWP models, IFS and TWRF (15 km resolution), along with the precipitation data from ERA5. First, we discuss the predicted tracks and intensities for the two-day period (Fig. 7). The predicted tracks from all models moved westward, hitting Taiwan on the second day with a small diversion after passing over the island. Notably, the IFS (blue line) has the best track prediction for this period, having corrected its significant poleward bias observed in the early stage (Fig. 5f). All five models place their cyclone centers near the west coast of the island, to the west of the Central Mountain, at the 24 h forecast. By 48 h, the centers are positioned in the middle of the Taiwan Strait.

Figure 7b, c illustrate the predicted maximum wind speed and minimum sea-level pressure associated with Haikui over the 48 h period starting at 12 UTC 2 September. Haikui weakened significantly after passing over Taiwan, with the maximum wind decreasing from 40 to 25 m s^{-1} and the central pressure increasing from 950 hPa to 987 hPa, according to the best track data. Among the model intensity forecasts, IFS and TWRF are similar, both showing higher intensity and closely matching the best track due to their higher resolutions. The intensity evolution predicted by GraphCast nearly aligns with that from ERA5. Meanwhile, Fuxi predicted the weakest typhoon for the first 24 h. There is a strong relationship between the wind

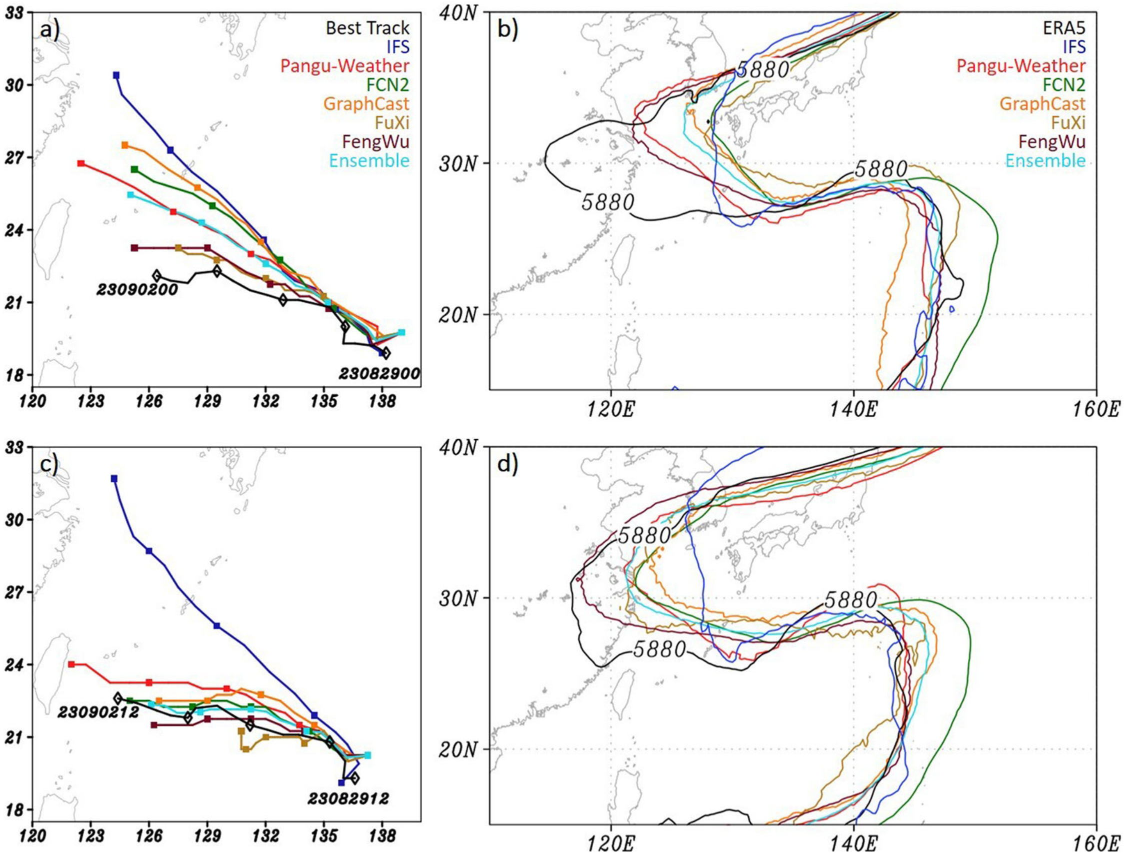

Fig. 6 | The 96 h tracks for Haikui and the corresponding 5880 m geopotential height at 500 hPa at 96 h of the five MLWP models, their ensemble, and IFS. The 96 h best track and predicted tracks are starting from 00 UTC 29 September (**a**) and from 12 UTC 29 September (**c**). **b** and **d** show the corresponding geopotential height contour line at 96 h forecast of five MLWP models, their ensemble, IFS, and the analysis of ERA5 at the same time. Colors according to the models are the same as Fig. 2 and the black line in (**a**) and (**c**) show the best track, while it in (**b**) and (**d**) show the geopotential height of ERA5. The open diamonds and the filled squares on (**a**) and (**c**) show Haikui's location every 24 h starting from the initial time.

and pressure profile for this case, which is also observed in other typhoon cases we examined and at longer lead time (figures not shown). In investigating the prediction of a mid-latitude winter storm, Charlton-Perez et al.[29] did not find as strong a relationship between wind and pressure variations from the MLWP models they evaluated.

The capability of MLWP models to predict TC formation is also examined using Haikui as an example. The tropical disturbance that can be traced as precursor of Haikui developed into a tropical storm at 00 UTC on 28 August 2023. This time is referred to as the TC formation time in our analysis, Table 2 lists the dates in the predictions of each model in which a disturbance as the precursor of Haikui can be identified, following the criteria established by Tsai et al.[56]. Among them, FCN2 was the earliest to predict Haikui's formation with the lead time of six days, while the others demonstrated a predicative ability to forecast formation within four to five days. Predicting tropical cyclone genesis usually involves more longer time scales and is best handled by ensemble system[57]. The single case presented here only provides a glimpse of the potential capability of MLWP models. We devote more in-depth research to this topic in another study.

Figure 8 illustrates the two-day accumulated rainfall from FuXi, GraphCast, ERA5, IFS, and TWRF (15 km). The last panel (Fig. 8h) displays the accumulated rainfall retrieved by the Quantitative Precipitation Estimation and Segregation Using Multiple Sensor (QPESUMS) algorithm from CWA[58], used for verification. Taiwan island is predominantly covered by the Central Mountain range with e peak about 4000 meters (Fig. 7a). As Haikui approached Taiwan from the east and passed its southern part, the rainfalls are accumulated mostly on the upwind side of the Central Mountain ridge, which in this case on the eastern side.

The maximum accumulated rainfall observed in the QPESUMS verification exceeds 700 mm over two days (Fig. 8h). Among all five models examined here, only TWRF reached this extreme value (Fig. 8e). The operational model IFS produced very good result in terms of distribution (Fig. 8d), with its maximum reaching 600 mm. The ERA5 (Fig. 8c) shows a rainfall pattern similar to IFS but with less details structure and a weaker peak intensity, in the range of 300–400 mm. For comparison, the IFS and ERA5 have resolutions of 0.25° and TWRF presented here has a 15 km resolution.

The only two precipitation outputs available from the five MLWP models are Fuxi and GraphCast (Fig. 8a, b). While the general patterns of the precipitations align with the verification, both AI model-generated rainfall amounts are smaller than those of the verification and are also less than those of the two dynamic NWP models (IFS and TWRF). Between them, GraphCast produced larger amount, reaching 200 mm, while Fuxi recorded rainfall in the range of 130–150 mm. Despite a small track difference, GraphCast predicted higher intensity than Fuxi during the first 24 h (Fig. 7b, c). This intensity bias in Fuxi contributes to its smaller accumulated rainfall. Additionally, the circular rainfall pattern over Taiwan Strait between Taiwan Island and south eastern China, is also much weaker in Fuxi compared to other models. These intensity and precipitation bias may stem from Fuxi's longer time step, which prioritizes longer lead time forecasts.

In investigating the capability of AI models in predicting an extreme weather event of Storm Ciaran (2023), Charlton-Perez et al.[29] noted that the four AI models they examined (FCN, Pangu-Weather, GraphCast, and FCN2) failed to accurately capture the structure and magnitude of the winds as revealed in ERA5. Their study suggested

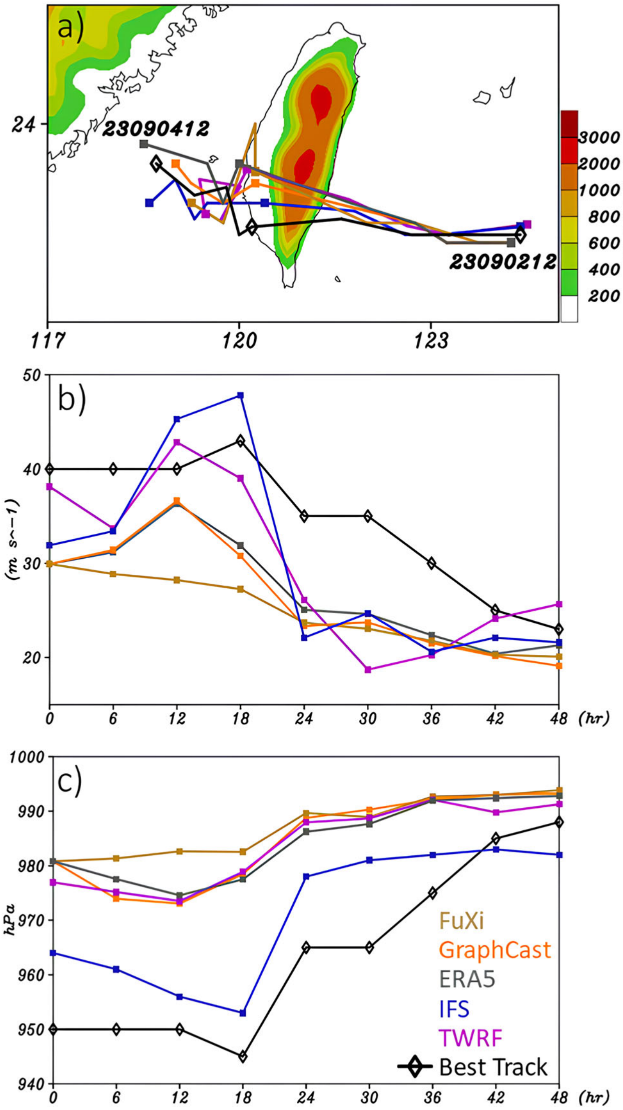

Fig. 7 | Best (CWA), analysis (ERA5), and predicted tracks, 10-m maximum wind speed, and minimum sea-level pressure of Typhoon Haikui of MLWP and NWP models. **a** best, analysis, and predicated tracks for two days starting on 12 UTC 2 September 2023 on map with terrain height (m). The open diamonds and the filled squares on (**a**) show Haikui's location every 24 h starting from the initial time. **b** shows the 10-m maximum wind speed (m s^{-1}) and **c** shows the minimum sea-level pressure (hPa) corresponding to (**a**). The open diamonds and the filled squares on (**b**) and (**c**) show Haikui's intensity every 6 h starting from the initial time. Colors according to the models are the same as Fig. 2 with ERA5 in color gray.

that the weaker winds predicted by these ML models, compared to NWP model forecasts, are not merely a result of being trained on a coarse resolution dataset. AI models often produce smoother results due to several factors: regularization techniques[59], data averaging[60], optimization objectives[37], and noise reduction[61]. The analysis presented in this study and by Charlton-Perez et al.[29] may offer AI modelers insights for future improvements of their models.

Discussion

In the rapidly advancing field of MLWP models, five standout global weather prediction systems have emerged: Pangu-Weather, FCN2, GraphCast, FuXi, and FengWu. Despite employing distinct AI/ML technologies, all five systems use ERA5 for training, spanning approximately 39 years. Publications on these models have demonstrated comparable, and in some cases, superior performance compared to the ECMWF's traditional NWP model, IFS, while exhibiting several orders of magnitude greater computational efficiencies.

Table 2 | The capability of MLWP models and IFS to predict formation of Haikui, occurred at 00 UTC on 28 August

Model	The earliest Initial time of the forecast predicting TC formation (UTC)	Predicted days before TC formation (8/28 00 UTC) (days)
IFS	8/23 12	4.5
Pangu-Weather	8/23 00	5
FCN2	8/22 00	6
GraphCast	8/23 12	4.5
FuXi	8/24 00	4
FengWu	8/24 00	4

The earliest time a model predicted formation of Haikui, and the number of days predicted ahead of the formation time.

Given the intricacies of local geographic profile in regional areas, high resolution models are essential for accurately simulating multi-scale processes responsible for high-impact weather phenomena. The efficacy of regional models, however, depends critically on the initial and boundary conditions provided by global models[34,62,63]. The recent emergency of MLWP opened a new avenue for the adoption of these models in operational forecasting. Therefore, evaluating the performance of global MLWP models within a limited area is warranted.

This study independently evaluates the performance of the five aforementioned MLWP models in East Asia and Western Pacific over a six-month period from June to November in 2023. We conducted our simulations using the codes provided by the respective model developers, which are available on their websites. The initial conditions for our simulations are sources from sourced from ERA5, remaining identical for all models, with a forecast integration period set at 168 h. Additionally, we include comparisons with the predictions generated by IFS, noting that IFS has a higher resolution of 9 km compared to 0.25° resolution of the MLWP models inherited from ERA5. Using ERA5 for initial conditions is not feasible for operational purposes, thereby potentially altering the performance comparisons when transitioning to an operational environment. Nevertheless, our primary focus remains on a consistent comparison of MLWP models, aiming to mitigate uncertainties stemming from varied initial conditions. In future studies, we plan to utilize other reanalysis, such as those from NCEP, to further enhance our insights in this endeavor.

We computed the latitude-weighted RMSE and ACC against valid ERA5 reanalysis for the evaluation in the East Asia and western Pacific region, which is of significant interest for our operational weather prediction. The region is also of great interest to meteorological community due its vigorous phenomena across a broad spectrum, including tropical cyclone activities, monsoon circulations, and the Madden-Julian Oscillation (MJO). The ACC and RMSE scores indicate that the best-performing model is FengWu, followed by FuXi and GraphCast, then FCN2, with Pangu-Weather ranking last. Notably, the IFS, with a resolution of 9 km, places it in the middle of the group. The trends and ranks of RMSE and ACC are consistent. Additionally, we conducted a simple ensemble average of the five MLWP models. The performance of the ensemble is comparable to FengWu.

We evaluated the prediction skill for TCs as an extreme event example. For the 11 typhoons (excluding three short-lived ones) that occurred in the western North Pacific in 2023, FengWu demonstrated the best track prediction among all models and led in four individual typhoons. The performance of the track prediction generally aligned with the ACC and RMSE scores. While Pangu-Weather has the largest averaged track error, it still performed best for three individual typhoons. The IFS excelled in the

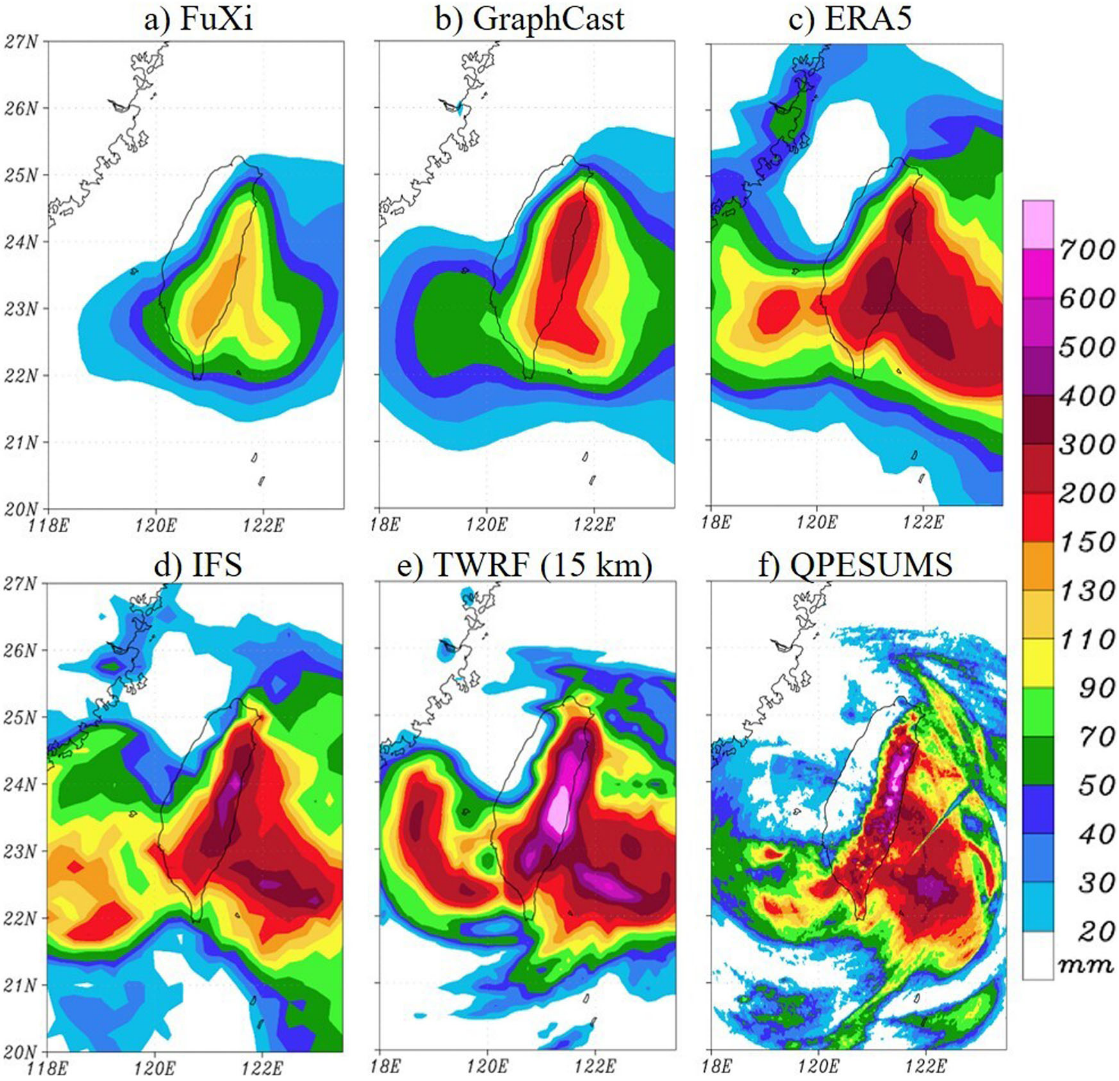

Fig. 8 | Estimated and predicted 48 h accumulated rainfalls on Taiwan and its vicinity during the passage of Typhoon Haikui, with an initial time at 12 UTC 2 September 2023. Predicted rainfalls (mm) of (**a**) FuXi, (**b**) GraphCast, (**c**) ERA5, (**d**) IFS, (**e**) TWRF (15 km), and estimated rainfalls (mm) of (**f**) QPESUMS.

intensity prediction, attributed to its higher resolution and minimal initial intensity bias. Surprisingly, FengWu exhibited the poorest intensity prediction skill. A more detailed examination of typhoon prediction for individual cyclones highlights diversification in model performance, indicating potential for future improvements.

The seasonal average of the 7-day prediction of the WPSH for the five MLWP models all indicates weak biases. A case study of Typhoon Haikui demonstrated close relationship between the predicted track and the position of the WPSH. Haikui was a mostly westward-moving typhoon that made landfall in Taiwan, causing significant rainfall in countries it encountered. The challenge presented in the early stage of Haikui led to largest track errors for some models, with IFS showing the largest among them. While these models can closely approximate the typhoon's track and some models generate reasonably accurate rainfall patterns, their resolution may not suffice for regional applications. Consequently, there remains a necessity for high-resolution regional models to offer detailed meteorological information, encompassing not only rainfall predictions but also variables such as temperature extremes. It is evident that MLWP models are progressing towards higher resolution predictions, whether through enhanced training techniques or downscaling predictions[64,65]. Another consideration for future development is the adoption of additional data within the existing reanalysis data. Further rigorous verifications of these models are still required.

The pace of development in MLWP is truly remarkable, with new systems emerging rapidly. Oskarsson et al.[66] have notably introduced a regional model based on GraphCast for regional weather forecasting around the Nordic area. However, a significant challenge in daily short-term prediction arises from the lack of high-resolution regional reanalysis data for training purposes. To address the challenge of insufficient resolution in ERA5 for TC intensity prediction, the FengWu group has developed the Multi-modal multi-Scale Causal AutoRegressive model (MSCAR)[67]. This innovative approach combines satellite images for TCs with ERA5 reanalysis data allowing for the extraction of causal relationships across these multi-modal datasets to enable global TC intensity autoregressive forecasting. The results of MSCAR show promising short-term performance, indicating a new pathway for the development of high-resolution AI/ML models in the fields of extreme weather predictions.

In October 2023, ECMWF launched its own AI prediction system[68], AIFS (where "I" denotes both AI and IFS), marking a significant advancement in the field. AIFS utilizes Graph Neural Networks technology and shares the same grid structure as IFS. The current iteration of AIFS boasts 13 vertical levels and a 0.25° resolution[69], with preliminary results indicating its superior performance compared to IFS.

As with conventional NWP models, MLWP models are continually evolving, with efforts focused on extending forecast lead times into subseasonal, to seasonal prediction, and ultimately climate predictions. A

Table 3 | Brief descriptions of five MLWP models

Model	Algorithm	Vertical pressure levels (excluding surface)	Frequency Options (h)	Time scale (days)	Variables (short name of ECMWF grib2 Table)	
					On pressure Levels (in addition to u, v, t, z)	At surface or specific height (in addition to msl, 10 u, 10 v, 2t)
Pangu-Weather	3D Earth-specific transformer (Swin Transformer v1 with Earth-specific positional bias)	13	a) 1 b) 3 c) 6^+ d) 24	-----	q	------
FourCastNet v2	Vision Transformer with Spherical Fourier Neural Operators	13	6	-----	r	100 u, 100 v, sp, tcwv
GraphCast	Muti-Mesh Graph neural network	a) 13^+ b) 37	6	-----	q, w	tp*
FuXi	U-Transformer (Swin Transformer v2 with cube embedding)	13	6	0 ~ 5^+ 6 ~ 10^+ 11 ~ 15	r	tp*
FengWu	Transformer based on Cross model fusion	a) 13^+ b) 37	6	-----	q	-----

$^+$Used in this study.

*6 h accumulated.

notable advantage of the MLWP models is their ability to generate large numbers of ensemble members by perturbing the initial fields in the current state, all while benefiting from their extremely high computational efficiency. We also advocate for the utilization of multi-model ensembles in addition to single-model ensembles, with direct applications including the use of multi-model ensemble predictions as boundary conditions for high-resolution regional models. Given that MLWP systems heavily rely on data, collaboration between the AI and meteorological communities is essential to enhance the prediction of extreme weather events by leveraging high-resolution and reliable localized data.

Methods

In this study, we conduct an independent evaluation of the following prominent MLWP models: Pangu-Weather, FCN2, GraphCast, FuXi, and FengWu (Table 3), all with horizontal resolution of 0.25°. While some models have multiple versions with different numbers of vertical levels, we use a 13-level configuration and 6 h time step for all models. We conducted our own simulations for each model using the codes published by model developers, available on arXiv (http://arxiv.org). ERA5 reanalysis are used as the initial condition in our simulations, which is the training data for all five MLWP models. The ML algorithm (backbone) used in these five MLWP models are briefly summarized in Table 3.

We compare independently performances of the five global MLWP systems in a region covering East Asia and western North Pacific. This is the area for which CWA is responsible for daily operation of weather predictions. The area also covers the region of most rigorous tropical cyclone (TC) or typhoon activities. Our evaluation period spans from June to November 2023, during which predictions of 11 typhoon cases will be assessed, with three short-lived ones excluded from the analysis. We conduct simulations of each model with identical initial conditions from ERA5 and the forecasts are also verified against ERA5 reanalysis at the valid time. The forecast lead time is 168 h (7 days) for the simulation.

The evaluation matrix includes RMSE, ACC of representative atmospheric variables, and predicted track and intensity errors. The ERA5 reanalysis are used as verification for the forecast fields.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Received: 11 April 2024; Accepted: 11 September 2024;
Published online: 28 September 2024

References

1. Turing, A. M. Computing machinery and intelligence. *Mind* **49**, 433–460 (1950).
2. Samuel, A. L. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.* **3**, 210–229 (1959).
3. Bar-Sinai, Y., Hoyer, S., Hickey, J. & Brenner, M. P. Learning data-driven discretizations for partial differen tial equations. *Proc. Natl. Acad. Sci.* **116**, 15344–15349 (2019).
4. Kochkov, D. et al. Machine learning–accelerated computational fluid dynamics. *Proc. Natl. Acad. Sci.* **118**, e2101784118 (2021).
5. Gröenquist, P. et al. Deep learning for post-processing ensemble weather forecasts. *Philos. Trans. Roy. Soc.* **A379**, 20200092 (2021).
6. Hess, P. & Boers, N. Deep learning for improving numerical weather prediction of heavy rainfall. *J. Adv. Model. Earth Syst.* **14**, e2021MS002765, https://doi.org/10.1029/2021MS002765 (2022).
7. Rojas-Campos, A. et al. Postprocessing of NWP precipitation forecasts using deep learning. *Wea. Forecast.* **38**, 487–497 (2023).
8. Dueben, P. D. & Bauer, P. Challenges and design choices for global weather and climate models based on machine learning. *Geosci. Model Dev.* **11**, 3999–4009 (2018).
9. Weyn, J. A., Durran, D. R. & Caruana, R. Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data. *J. Adv. Model. Earth Syst.* **11**, 2680–2693 (2019).
10. Weyn, J. A., Durran, D. R. & Caruana, R. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *J. Adv. Model. Earth Syst.* **12**, e2020MS002109 (2020).
11. Rasp, S. & Thuerey, N. Data-driven medium-range weather prediction with a Resnet pretrained on climate simulations: A new model for weather bench. *J. Adv. Model. Earth Syst.* **13**, 2020MS002405 (2021).
12. Keisler, R. Forecasting global weather with graph neural networks. *arXiv*, 2202.07575v1, https://doi.org/10.48550/arXiv.2202.07575 (2022).
13. de Burgh-Day, C. O. & Leeuwenburg, T. Machine learning for numerical weather and climate modeling: A review. *Geosci. Model Dev.* **16**, 6433–6477 (2023).
14. Pathak, J. et al. 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 (2022).
15. Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).

16. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2010.11929v2, https://doi.org/10.48550/arXiv.2010.11929 (2020).
17. Li, Z., et al & Anandkumar Fourier neural operator for parametric partial differential equations. https://doi.org/10.48550/arXiv.2010.08895 (2020).
18. J. Guibas, M. Mardani, et al. Adaptive Fourier Neural Operators: Efficient token mixers for transformers. *arXiv*, 2111.13587v2, https://doi.org/10.48550/arXiv.2111.13587 (2021).
19. Bonev, B., et al. Spherical Fourier Neural Operators: Learning stable dynamics on the sphere. *arXiv*, 2306.03838v1, https://doi.org/10.48550/arXiv.2306.03838 (2023).
20. Chen, L., Du, F., Hu, Y., Wang, Z. & Wang, F. SwinRDM: integrate swinRNN with diffusion model towards high-resolution and high-quality weather forecasting. *arXiv*, 2306.03110v1, https://doi.org/10.48550/arXiv.2306.03110 (2023).
21. Bi, K., et al. Pangu-Weather: A 3D high-resolution model for fast and accurate global weather forecast, *arXiv*, 2211.02556v1, https://doi.org/10.48550/arXiv.2211.02556 (2022).
22. Liu, Z., et al. Swin transformer: hierarchical vision transformer using shifted windows. *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), virtual* 10012–10022 (2021).
23. Lam, R. et al. Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).
24. Chen, L. et al. FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Clim. Atmos. Sci.* **6**, 190 (2023).
25. Liu, Z. et al. Swin transformer v2: Scaling up capacity and resolution. *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, pp. 11999–12009 (2022).
26. Chen, K. et al. FengWu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv*, 2304.02948v1, https://doi.org/10.48550/arXiv.2304.02948 (2023).
27. Schaul, T., Quan, J., Antonoglou, I. & Silver, D. Prioritized experience replay. *arXiv*,1511.05952v4, https://doi.org/10.48550/arXiv.1511.05952 (2015).
28. Bouallegue, Z. B. et al. The Rise of Data-Driven Weather Forecasting A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context. *BAMS* **105**, E864–E883 (2024).
29. Charlton-Perez, A. J. et al. Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciarán. *npj Clim. Atmos. Sci.* **7**, 93 (2024).
30. Schultz, M. G. et al. Can deep learning beat numerical weather prediction? *Philos. T. Roy. Soc. A.* **379**, 20200097 (2021).
31. Mass, C. F. & Kuo, Y.-H. Regional real-time numerical weather prediction: Current status and future potential. *Bull. Am. Meteor. Soc.* **79**, 253–263 (1998).
32. Wu, W., Lynch, A. H. & Rivers, A. Estimating the uncertainty in a regional climate model related to initial and lateral boundary conditions. *J. Clim.* **18**, 917–933 (2005).
33. Hsiao, L.-F., Chen, D.-S., Hong, J.-S., Yeh, T.-C. & Fong, C.-T. Improvement of the numerical tropical cyclone prediction system at the Central Weather Bureau of Taiwan: TWRF (Typhoon WRF). *Atmosphere* **11**, 657 (2020).
34. Hsiao, L.-F. et al. Blending of global and regional analyses with a spatial filter: Application to typhoon prediction over the western north Pacific Ocean. *Wea. Forecast.* **30**, 754–770 (2015).
35. Leutbecher, M. & Palmer, T. N. Ensemble forecasting. *J. Comput. Phys.* **227**, 3515–3539 (2008).
36. Magnusson, L. et al. ECMWF activities for improved hurricane forecasts. *Bull. Am. Meteorol. Soc.* **100**, 445–458 (2019).
37. Bishop, C. M. Pattern Recognition and Machine Learning. Springer New York 2006 (2006).
38. Hagedorn, R., Doblas-Reyes, F. J. & Palmer, T. N. The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. *Tellus* **57A**, 219–233 (2005).
39. Kirtman, B. et al. The North American multimodel ensemble: Phase-1 seasonal-to-interannual prediction; Phase-2 toward developing intraseasonal prediction. *Bull. Am. Meterol. Soc.* **95**, 585–601 (2014).
40. Wang, B., Xiang, B. & Lee, J.-Y. Subtropical high predictability establishes a promising way for monsoon and tropical storm predictions. *Proc. Natl. Acad. Sci. USA* **110**, 2718–2722 (2013).
41. George, J. E. & Gray, W. M. Tropical cyclone recurvature and noncurvature as related to surrounding wind-height fields. *J. Appl. Meteor.* **16**, 34–42 (1977).
42. Madden, R. A. & Julian, P. R. Description of global-scale circulation cells in the tropics with a 40–50 day period. *J. Atmos. Sci.* **29**, 1109–1123 (1972).
43. Madden, R. A. & Julian, P. R. Observations of the 40–50-day tropical oscillation - a review. *Mon. Wea. Rev.* **122**, 814–837 (1994).
44. Sobel, A. H. Tropical weather. *Nat. Educ. Knowl.* **3**, 2 (2012).
45. Mapes, B. E. & Houze, R. Cloud clusters and superclusters over the oceanic warm pool. *Mon. Wea. Rev.* **121**, 1398–1415 (1993).
46. Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J. & Neumann, C. J. The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone data. *Bull. Am. Meterol. Soc.* **91**, 363–376 (2010).
47. Knapp, K. R., Diamond, H. J., Kossin, J. P., Kruk, M. C. & Schreck, C. J. I. International Best Track Archive for Climate Stewardship (IBTrACS) Project, Version 4. NOAA National Centers for Environmental Information, https://doi.org/10.25921/82ty-9e16 [access date: July 2023] (2018).
48. Velden, C. & Leslie, L. The basic relationship between tropical cyclone intensity and the depth of the environmental steering layer in the Australian region. *Wea. Forecast.* **6**, 244–253 (1991).
49. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18**, 50–60 (1947).
50. Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bull.* **1**, 80–83 (1945).
51. Holland, G. J. *The Global Guide to Tropical Cyclone Forecasting*. Geneva: World Meteorological Organization. (Ed.) (1993).
52. Emanuel, K. A. Tropical cyclones. *Annu. Rev. Earth Planet. Sci.* **31**, 75–104 (2003).
53. Hsiao, L.-F., Liou, C.-S., Yeh, T.-C. & Guo, Y.-R. A vortex relocation scheme for tropical cyclone initialization in Advanced Research WRF. *Mon. Wea. Rev.* **138**, 3298–3315 (2010).
54. Hsiao, L.-F. et al. Application of WRF 3DVAR to operational typhoon prediction in Taiwan: Impact of outer loop and partial cycling approaches. *Wea. Forecast.* **27**, 1249–1263 (2012).
55. Lee, C.-S., Huang, L.-R., Shen, H.-S. & Wang, S.-T. A climatology model for forecasting typhoon rainfall in Taiwan. *Nat. Hazards* **37**, 87–105 (2006).
56. Tsai, H., Lu, K., Elsberry, R. L., Lu, M. & Sui, C. Tropical Cyclone–like Vortices Detection in the NCEP 16-Day Ensemble System over the Western North Pacific in 2008: Application and Forecast Evaluation. *Wea. Forecast.* **26**, 77–93 (2011).
57. Elsberry, R. L., Tsai, H.-C., Chin, W.-C. & Marchok, T. P. ECMWF Ensemble Forecasts of Six Tropical Cyclones That Formed during a Long-Lasting Rossby Wave Breaking Event in the Western North Pacific. Atmosphere 2024, **15**, 610 (2024).
58. Chang, P.-L. et al. An Operational Multi-Radar Multi-Sensor QPE System in Taiwan. *Bull. Amer. Meteor. Soc*. **102**, E555–E577 (2021).
59. Goodfellow, I., Bengio, Y. & Courville, A. Deep Learning. *MIT Press*, Chapter 7 (2016).
60. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).

61. Ng, A. Y. Feature selection, L1 vs. L2 regularization, and rotational invariance. In Proceedings of the Twenty-first International Conference on Machine Learning (ICML) (2004).
62. Li, C.-H., Berner, J., Hong, J.-S., Fong, C.-T. & Kuo, Y.-H. The Taiwan WRF ensemble prediction system: Scientific description, model-error representation and performance results. *Asia-Pac. J. Atmos. Sci.* **56**, 1–15 (2020).
63. Chen, D.-S. et al. Impacts of assimilating dual-Doppler radar-retrieval winds on the TWRF predictions of Typhoon Nesat (2017). *Terr. Atmos. Ocean. Sci.* **32**, 649–667 (2021).
64. Mardani, M. et al. Residual Diffusion Modeling for Km-scale Atmospheric Downscaling. *arXiv*, 2309.15214v3, https://doi.org/10.48550/arXiv.2309.15214 (2024).
65. Han, T. et al. FengWu-GHR: Learning the Kilometer-scale Medium-range Global Weather Forecasting. arXiv:2402.00059v1, https://doi.org/10.48550/arXiv.2402.00059 (2024).
66. Oskarsson, J., Landelius, T. & Lindsten, F. Graph-based neural weather prediction for limited area modeling. *arXiv*, 2309.17370v2, https://doi.org/10.48550/arXiv.2309.17370 (2023).
67. Wang, X., et al Global tropical cyclone intensity forecasting with multi-model multi-scale casual autoregressive model. *arXiv*, 2402.13270v1, https://doi.org/10.48550/arXiv.2402.13270 (2024).
68. Alexe, M. et al. ECMWF unveils alpha version of new ML model. Accessed 13 October 2023, https://www.ecmwf.int/en/about/media-centre/aifs-blog/2023/ECMWF-unveils-alpha-version-of-new-ML-model (2023).
69. Lang, S. et al. AIFS-ECMWF's data-driven forecasting system. arXiv preprint arXiv:2406.01465 (2024).

Acknowledgements

The research group of this study appreciates deeply the generality and open minds of the model developers for Pangu-Weather, FCN2, GraphCast, FuXi, and FengWu for providing the codes and instructions in the open domain, allowing us the opportunity to understand and use these models. Five MLWP models are available through the following repositories: Pangu-Weather: https://github.com/198808xc/Pangu-Weather. FourCastNet v2: https://github.com/ecmwf-lab/ai-models-fourcastnetv2. GraphCast: https://github.com/google-deepmind/graphcast. FuXi: https://github.com/tpys/ai-models-fuxi. FengWu: https://github.com/OpenEarthLab/ai-models-fengwu. We acknowledge the support from the CWA Numerical Information Division, which provided essential computing and storage infrastructure critical for the study. Furthermore, we are grateful to ECMWF for providing the ERA5 reanalysis and IFS forecast fields, which contributed to the research endeavors. Finally, we appreciate two reviewers for their constructive comments and suggestions, which helped us improve the quality of the study and the manuscript.

Author contributions

C.C.L., K.H., M.S.P., D.S.C., P.L.C., and L.F.H. designed research. C.C.L. produced forecasts of MLWP models. C.C.L. and K.H. wrote the code for data analysis. M.S.P., K.H., C.C.L., and D.S.C. analyzed the data. M.S.P., K.H., C.C.L., P.L.C., and L.F.H. wrote the draft manuscript. C.P.C manages this project. C.T.F., J.S.H., K.C.L., C.R.C., and H.C.K. participated in the discussion and suggested changes to the manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Kathryn Hsu.