JGR Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2024JH000481

Key Points:

- Pangu-Weather's enhanced environmental and tropical cyclone (TC) track forecasting may stem from superior initial conditions (ICs)
- Pangu-Weather significantly underestimates and predicts a weaker physical structure of TCs
- Pangu-Weather's TC intensity and structure forecasts show low sensitivity to ICs

Correspondence to:
N. Wu and X. Shi,
wunaigeng@hotmail.com;
shixm@ust.hk

Citation:
Shi, Y., Hu, R., Wu, N., Zhang, H., Liu, X., Zeng, Z., et al. (2025). Comparison of AI and NWP models in operational severe weather forecasting: A study on tropical cyclone predictions. *Journal of Geophysical Research: Machine Learning and Computation*, *2*, e2024JH000481. https://doi.org/10.1029/2024JH000481

Received 18 OCT 2024
Accepted 13 MAY 2025

Comparison of AI and NWP Models in Operational Severe Weather Forecasting: A Study on Tropical Cyclone Predictions

Yang Shi[1], Rong Hu[1], Naigeng Wu[1], Hualong Zhang[1], Xinhang Liu[1], Zhilin Zeng[1], Jing Zhu[1], Pucheng Han[1], Cong Luo[1], Hongyan Zhang[1], Jie He[2], and Xiaoming Shi[3]

[1]Guangdong Meteorological Observatory, Marine Weather Forecast Center of South China Sea, Guangzhou, China,
[2]Guangzhou Institute of Tropical and Marine Meteorology, China Meteorological Administration, Guangzhou, China,
[3]Division of Environment and Sustainability, Hong Kong University of Science and Technology, Hong Kong, China

Abstract Data-driven artificial intelligence weather prediction (AIWP) models show great potential in weather forecasts, facilitating paradigm shift of prediction from a deductive to an inductive inference. However, this shift raises concerns regarding the performance of the AIWP models in severe weather forecasting. Tropical cyclones (TCs) are one of the most typical cases of severe weather prediction. In this study, we compare Western Pacific TCs in 2023 produced by the AIWP model, Pangu-Weather, with those generated by numerical weather prediction (NWP) models, specifically the European Center for Medium-Range Weather Forecasts (ECMWF) and the National Centers for Environmental Prediction (NCEP), in the operational context. We analyze the impact of different initial conditions (ICs) on AIWP models, representative by Pangu-Weather, in TC forecasting. Our analysis includes statistical evaluation of forecast skill related to TC activity, track, intensity, and a case study on the physical structure of a TC. The Pangu-Weather model exhibits superior forecast skills compared to the NWP model regarding TC tracks and environmental variables within TC activity domains, particularly at longer forecast lead times. However, the overly smooth forecasts of Pangu-Weather and the coarse-resolution ICs with reduced information of TCs potentially lead to the underestimation of intensity and a weakened dynamic-thermodynamic structure of TCs. Also, Pangu-Weather shows low sensitivity to ICs concerning TC structure and intensity. Hybrid models combining physical processes with data-driven approaches may enhance AIWP performance for severe weather forecasting.

Plain Language Summary Artificial intelligence weather prediction (AIWP) models, such as Pangu-Weather, have introduced significant advancements in weather forecasting. Pangu-Weather has demonstrated superior statistical forecasting skills compared to traditional numerical weather prediction (NWP) models, prompting interest in its performance in forecasting severe weather. In 2023, Guangdong Meteorological Observatory successfully integrated the Pangu-Weather model into its operations. After 1 year of operational data accumulation, we find that Pangu-Weather outperforms traditional NWP models in statistical forecast skills of medium-range forecasting. However, its performance in predicting the intensity and structure of severe weather events, such as tropical cyclones (TCs), is inadequate, due to significant underestimation of both intensity and physical structure of TCs. Additionally, employing higher-quality ICs may enhance Pangu-Weather's forecasting of TC tracks and environmental variables within the TC activity area, but have limited effects on TC intensity and structure. These discrepancies may be related to the coarse-resolution ICs with reduced information of TCs, the bias in model's training data and model's inherent characteristic of converging to average values. Our study enhances the understanding of AIWP models in severe weather forecasting and provides a research foundation for future application and refinement of AIWP models in operational forecasting.

1. Introduction

Over more than a century of development, numerical weather prediction (NWP) has become an indispensable core component of operational weather forecasts (Bauer et al., 2015). Modern NWP models use data assimilation to integrate observations and solve physics-based governing equations (e.g., partial differential equations). Through iterative integration and the estimation of sub-grid scale physical processes, these models forecast future atmospheric conditions (Bjerknes, 1904; Charney et al., 1950; Richardson et al., 1922). Despite significant advancements in the forecasting capabilities, the accuracy of the NWP models remains limited by two primary aspects. First, the inherently nonlinear nature of the physical equations governing atmospheric motions may not be accurately represented by numerical methods for discretizing solutions, which inevitably introduces systematic errors. Second, parameterization schemes used to represent subgrid-scale physical processes also contribute to these systematic errors. Furthermore, traditional NWP faces escalating challenges as model resolution increase and the volume of new observational data expands. These challenges primarily involve the extensive computational resources required for subgrid-scale parameterization, the integration of multi-layer physical and chemical processes, and the assimilating of large-scale, multi-source observational data (Bauer et al., 2015).

In recent years, data-driven artificial intelligence weather prediction (AIWP) models have shown great potential in weather forecasts (Dueben & Bauer, 2018; Scher, 2018; Weyn et al., 2019, 2020). Numerous studies have claimed that AIWP models can match or even surpass state-of-the-art operational NWP model, European Center for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS) (Bi et al., 2023; Chen K., 2023; Keisler, 2022; Lam et al., 2023; Pathak et al., 2022). For instance, FourCastNet (Pathak et al., 2022) utilizes the Adaptive Fourier Neural Operator with a powerful vision transformer backbone network to enhance the accuracy of solving nonlinear partial differential equations in high-resolution models. Similarly, Pangu-Weather (Bi et al., 2023) integrates Earth-specific transformer architectures and position encoding techniques, effectively incorporating prior knowledge of meteorological variables and absolute geographical locations to improve the model's capability in capturing spatial correlations. Additionally, Pangu-Weather also employs a hierarchical temporal aggregation algorithm to reduce cumulative errors. GraphCast (Lam et al., 2023) leverages a graph neural network architecture to map input data on the raw latitude-longitude grid into multi-grid learning features, thereby facilitating efficient propagation of both local and remote information with minimal message-passing steps. Overall, AIWP models offer enhanced flexibility in simulating nonlinear relationships and demonstrate significant advantages in accuracy and computational efficiency compared to conventional NWP models.

The forecasts of AIWP are learned from previous weather conditions, shifting the prediction from a deductive to an inductive inference (Ben Bouallègue et al., 2024). However, this shift raises concerns of the limited sample size of severe weather events within the training data sets. Moreover, all regression algorithms, including deep neural networks, tend to eventually converge to average values (Bi et al., 2023). This “over-smoothing” effect can lead to the loss of critical information, reduce prediction accuracy, and hinder the ability to capture nonlinear patterns in the data (Zhang et al., 2024). These drawbacks in AIWP models may pose challenges for severe weather forecasting, which attracts more public interest than the forecasting of mild weather.

Case studies are essential for understanding the specificity and physical structure of severe weather forecasting in AIWP (Charlton-Perez et al., 2024; Magnusson L., 2023). Also, grasping the generality and statistical characteristics of the forecast performance of AIWP needs statistical analysis of data accumulated for a period of time (Ben Bouallègue et al., 2024). Previous studies mainly focus on utilizing the fifth-generation ECMWF reanalysis (ERA5) data set as the initial conditions (ICs) for the AIWP models, comparing their outputs to those of state-of-the-art NWP models. However, the reanalysis data sets like ERA5 are not available in real-time and cannot be used for operational weather forecasting. Therefore, how does AIWP model perform in severe weather forecasting in the real operational context? There remains a research gap concerning the influence of varying ICs on the performance of the severe weather forecasting in the real operational context. Specifically, how do these conditions affect the dynamic and thermodynamic structure, as well as other multi-dimension features of severe weather?

Pangu-Weather is the first AIWP model to claim forecast accuracy surpassing the traditional leading operational NWP model. It is notable for not only scoring well on statistical analysis of forecast skills but also excellent performance in TC track forecasting. Since 2023, meteorological agencies such as the ECMWF and China Meteorological Administration (CMA) have started to incorporate the open-source Pangu-Weather model into their operations. Therefore, this study evaluates the performance of AIWP models, represented by Pangu-Weather, in the operational context, focusing on TC forecasting. We systematically compare the performance of the Pangu-Weather model against NWP models in real-time forecasting operations, assessing the sensitivity of models to different ICs. The potential and limitations of data-driven AIWP models in TC forecasting are comprehensively analyzed across various dimensions, including TC tracks, intensity and dynamic-thermodynamic structure.

Table 1
Information of Selected Northwestern Pacific Typhoons (≥17 m/s) in This Study

ID	Name	Maximum wind speed (m/s)	Min pressure (hPa)	Maximum grade (Wang et al., 2007)
2,302	Marwar	68	905	Super TY
2,303	Guchol	40	960	TY
2,304	Talim	40	960	TY
2,305	Doksuri	62	915	Super TY
2,306	Khanun	52	935	Super TY
2,307	Lan	55	930	Super TY
2,309	Saola	62	915	Super TY
2,310	Damrey	30	980	STS
2,311	Haikui	35	935	Super TY
2,312	Kriogi	30	980	STS
2,313	Yun-yeung	20	995	TS
2,315	Bolaven	68	900	SuperTY
2,316	Sanba	25	988	STS

Note. According to Wang et al. (2007), the grade of TCs in China is determined by their maximum wind speed, with the thresholds for Super TY, STY, TY, STS and TS are ≥51, 41.5–50.9, 32.7–41.4, 24.5–32.6 and 17.2–24.4 m/s, respectively.

The remainder of this paper is organized as follows. Section 2 introduces the data sets and evaluation methods employed in this study. Section 3 presents the spatio-temporal distributions of statistical forecast skills of environmental field elements from the Pangu-Weather and NWP models within the domain of TC activities. Section 4 introduces the characteristics of the forecasted TC tracks and intensity. Section 5 analyzes the horizontal and vertical structures of TCs, and Section 6 gives main conclusions and discussions.

2. Data and Methods

2.1. Data

The Pangu-Weather model evaluated in this study is based on its open-source code (Bi et al., 2023). This model employs 39 years (1979–2017) of the ERA5 data for training and incorporates a hierarchical temporal aggregation algorithm with the principle of minimizing iterations, producing hourly forecasts for the next 10 days. The Pangu-Weather model incorporates multiple temporal resolutions with different time step configurations (1-hr, 3-hr, 6-hr, and 24-hr versions). We select different time step version with the minimum required iterations for each lead time to control error propagation. The forecast performances of the Pangu-Weather model is compared with the operational NWP models, such as European Center for Medium-Range Weather Forecasts Integrated Forecasting System (ECMWF-IFS) and National Centers for Environmental Prediction Global Forecast System (NCEP-GFS). For comparative analysis with NWP models, ECMWF-IFS and NCEP-GFS analysis data and ERA5 reanalysis data are used as ICs for the Pangu-Weather model, resulting in forecast fields labeled PANGU-ECMWF, PANGU-NCEP and PANGU-ERA5, respectively. Forecasts are initialized twice daily at 0000 UTC and 1200 UTC, with a horizontal resolution of 0.25° × 0.25° and 13 vertical pressure layers, involving five upper-air variables and four surface variables.

The ECMWF-IFS analysis data used for ICs has horizontal resolutions of 0.2° × 0.2° (upper levels) and 0.1° × 0.1° (surface). These data are linearly interpolated to 0.25° × 0.25° resolution before input into the Pangu-Weather model. The NCEP-GFS ICs are obtained from the NCEP online data set, which features a horizontal resolution of 0.25° × 0.25° and 25 vertical levels. The ERA5 data also has a horizontal resolution of 0.25° × 0.25° and encompass 37 vertical pressure levels.

Since May 2023, the Pangu-Weather model has been operational at the Guangdong Meteorological Observatory. In this research, we evaluate the TCs with official designation above 17 m/s occurring in the Northwest Pacific during May–October 2023 (Table 1). Typhoons No. 2308 and No. 2314 are excluded from evaluation due to data loss caused by network and server failures. Following Zhang (2018), the TC center is defined by the minimum of geopotential height at 850 hPa, and the maximum 10-m wind speed of a TC is determined as the highest 10-m wind speed within a 500 km radius from the TC center.

Real-time TC information from CMA severed as the observation reference for evaluating forecasted TC tracks and intensity. Due to sparse marine observations and the limited spatial and temporal coverage of satellite data, ERA5 data is used as a reference for evaluating the spatial distribution of errors, as well as the horizontal and vertical structures of TCs. In the meantime, since the Pangu-Weather model is trained on ERA5 data, the performances are evaluated to match an independent benchmark-the FNL (Final) Operational Global Analysis data (NCEP, 2000), to ensure the robustness of the results.

2.2. Evaluation Metrics

Root-mean-square error (RMSE) and anomaly correlation coefficient (ACC) are used to assess the forecasting performance of models. The RMSE measures the discrepancy between forecasts and observations, as shown in Equation 1 (Rasp & Thuerey, 2021).

$$\text{RMSE} = \frac{1}{N_{\text{forecasts}}} \sum \sqrt{\frac{1}{n} \sum_{k=1}^{n} L(j)\left(y_k - o_k\right)^2} \tag{1}$$

where (y_k, o_k) denotes the kth of n pairs of target and predicted values. $N_{\text{forecasts}}$ represents the total number of forecasts. n denotes the total number of grids or observational stations from a single forecast. $L(j)$ is the weight function of latitude j.

$$L(j) = \frac{\cos(\text{lat}(j))}{\frac{1}{N_{\text{lat}}} \sum_{j}^{N_{\text{lat}}} \cos(\text{lat}(j))} \tag{2}$$

where N_{lat} denotes the total number of grids those have the same latitude j.

Anomaly correlation coefficients (ACC) is used to evaluate the model's performance by comparing forecasted anomalies to observed anomalies. The formula for calculating ACC is as follows (Rasp & Thuerey, 2021):

$$\text{ACC} = \frac{\sum_i L(j)\, y'_{(i,j)}\, o'_{(i,j)}}{\sqrt{\sum_i L(j) {y'_{(i,j)}}^2} \sqrt{\sum_i L(j)\, {o'_{(i,j)}}^2}} \tag{3}$$

where i and j represents the grid index along the same latitude and longitude respectively. $y'_{(i,j)}$ and $o'_{(i,j)}$ is the anomaly of the forecast and observation. The 30-year climatological mean from 1991 to 2020 was used as the baseline for anomaly calculation.

3. Statistical Evaluation of Forecast Skills

In this section, the overall statistical forecasting skills of Pangu-Weather and NWP models are compared within the domain of TCs' activities in the Northwest Pacific (100°E–150°E, 0°N–50°N).

3.1. Temporal Variations of Forecast Skills

To evaluate the forecasting performance of the Pangu-Weather and NWP models on the forecasts of TC-related environment fields, we select eight variables associated with the dynamic-thermodynamic structure of the TCs at the middle and lower levels. Given that the lower troposphere is influenced by various factors such as boundary layer processes and underlying surface conditions, which make the state of the atmosphere in these layers more complex and variable, it better reflects the forecast performance and accuracy of NWP and Pangu-Weather models. Among the selected eight selected variables, the low-level thermodynamic field is crucial as it reflects the dynamic characteristics and temperature structure of the TC's at lower levels. Meanwhile, the geopotential height of 850 hPa, mean sea level and u, v component of 10-m wind are used for TC positioning and intensity estimation, respectively. Therefore, an assessment of these parameters can more accurately reflect the differences between the NWP and Pangu-Weather model in TC forecasting.

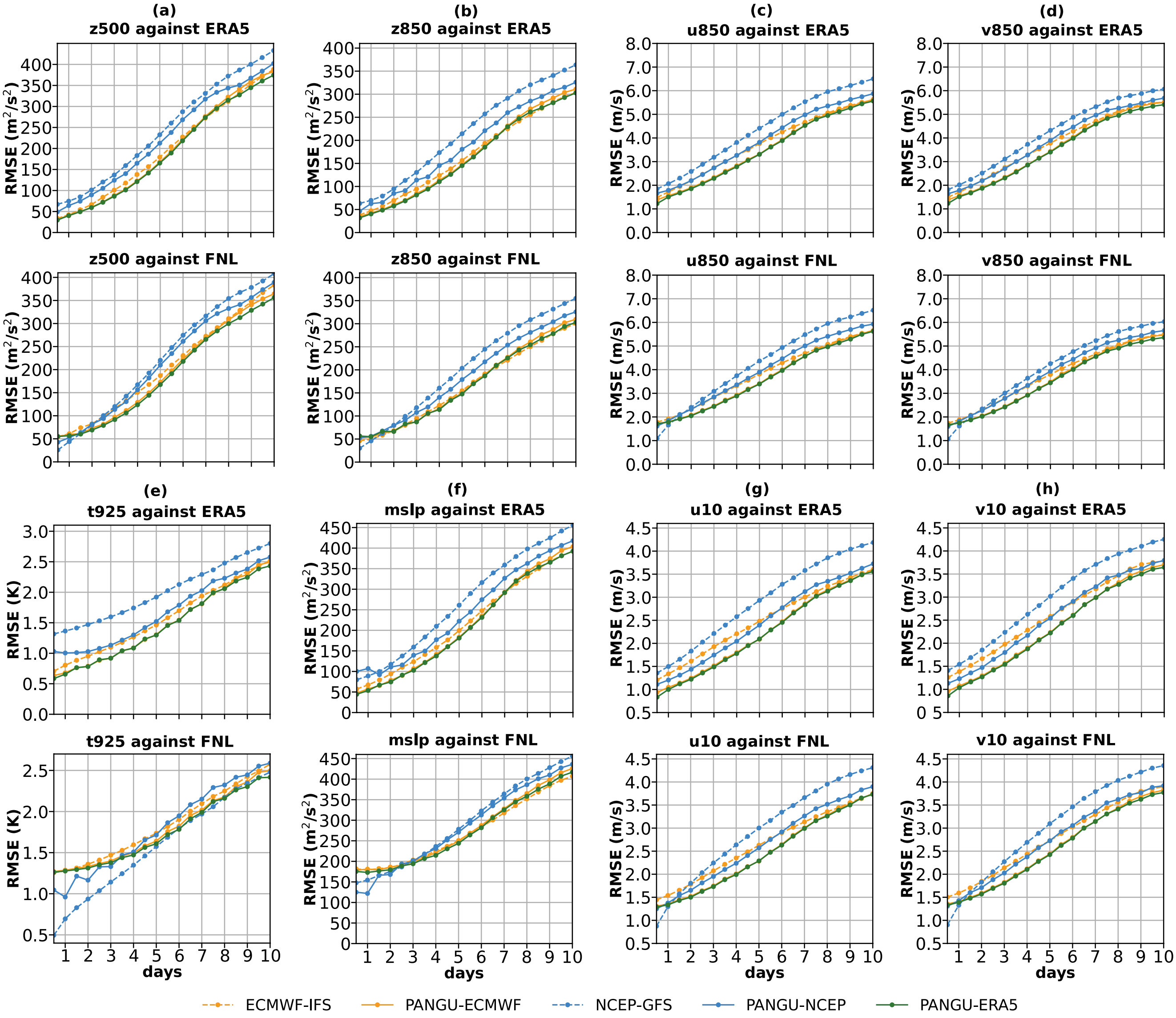

Figure 1. Root Mean Square Error of the various element forecasts from the Pangu-Weather and numerical weather prediction models averaged from May–October 2023 against ECMWF reanalysis and FNL.

Figure 1 illustrates the variations in RMSE over lead times against the two observation references respectively. The results indicate that forecasts at most lead times from the Pangu-Weather model demonstrate higher performance socres from May to October comparied to the NWP model with the same ICs, except for elements at the first one to two lead times and temperature at lower level of troposphere from 925 hPa to surface against FNL. Among the forecasts from the Pangu-Weather model driven by different ICs, PANGU-ERA5 exhibits the smallest errors across nearly all lead times, followed by PANGU-ECMWF, while PANGU-NCEP shows the largest errors. The performance of PANGU-ECMWF is comparable to that of PANGU-ERA5.

When using FNL as the observational reference, the NCEP-GFS demonstrates the best performance for all elements at the first one to two lead times. Nevertheless, as the lead time increases, the RMSE of NCEP-GFS rapidly escalates, eventually making it the model with the largest errors at the subsequent lead times. This phenomenon may be attributed to the fact that both NCEP-GFS and FNL are derived from the same model system, resulting in nearly identical fields at initial time. With the lead time increases, the error growth rate of NCEP-GFS is significantly larger than other models, causing it to quickly become the model with the greatest RMSE. In the meantime, NCEP-GFS appeared to be the best for temperature at 925 hPa. This is inconsistent with the

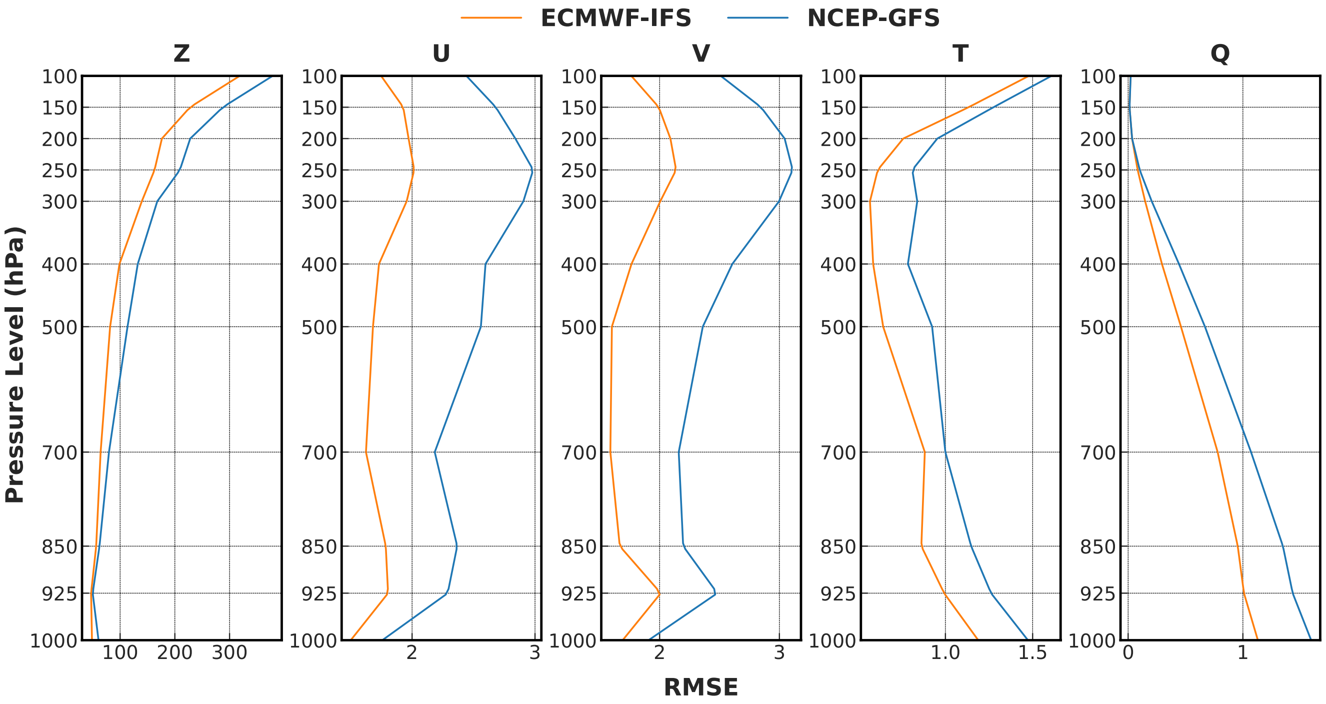

Figure 2. Root Mean Square Error of analysis data from ECMWF-IFS and NCEP-GFS against radiosonde observations from China mainland during May–October 2023.

conclusions when using ERA5 data as observation references. Researchers have found that the ERA5 reanalysis data is more reliable than FNL in humidity and temperature at lower level of troposphere under complex weather conditions, and has significant advantages over FNL and NCEP reanalysis data in terms of boundary layer height, surface parameters and precipitable water vapor (Chen et al., 2021; Guo et al., 2021; Meng et al., 2022). Therefore, the results against ERA5 is more reliable for temperature at 925 hPa.

Preliminarily, we made a simple comparison between the operational analysis data from ECMWF-IFS and NCEP-GFS with soundings observations from China mainland during May-October 2023 in Figure 2. The evaluation results show that ICs from ECMWF-IFS outperform those from NCEP-GFS. This demonstrates that the performance of the Pangu-Weather model in environment field may be related to the quality of its ICs.

Moreover, environmental fields evaluated against ERA5 and FNL (with horizontal resolution of 0.25° × 0.25°) are closely tied to large-scale atmospheric systems. Consequently, the Pangu-Weather model demonstrates superior performance compared to NWP and exhibits higher sensitivity to ICs in large-scale circulation fields.

Furthermore, we analyzed the ACC variations between the predicted and observed anomaly fields as a function of lead times, as shown in Figure 3. A higher ACC indicates a more accurate representation of the spatial patterns of anomalies. Previous studies have indicated that an ACC greater than 0.6 suggests that the model possesses forecast skill (Bauer et al., 2015; K. Chen et al., 2023; Chen, Du, et al., 2023; Chen, Zhong, et al., 2023; Jolliffe & Stephenson, 2003). In Figure 3, the Pangu-Weather model demonstrates superior performance in forecasting spatial distribution characteristics of anomaly fields compared to the corresponding NWP models for most variables, except for specific humidity and temperature at lower level of troposphere from 925 hPa to surface against FNL. Similar to the RMSEs, the PANGU-ERA5 achieves the highest ACC scores at almost all lead times, followed by the PANGU-ECMWF, with the lowest score for the PANGU-NCEP.

3.2. Spatial Distribution of Forecast Errors

Figure 4 presents the spatial distribution of 10-m averaged wind speeds of ERA5 and the RMSE values for 10-m wind speed forecasts from each model at the lead times of 24, 48 and 72 hr against ERA5, averaged over the period from May to October 2023 without dates of observed TCs above 17 m/s. The averaged 10 m wind speeds from ERA5 of the same period is also included as a reference in Figure 4. The averaged spatial

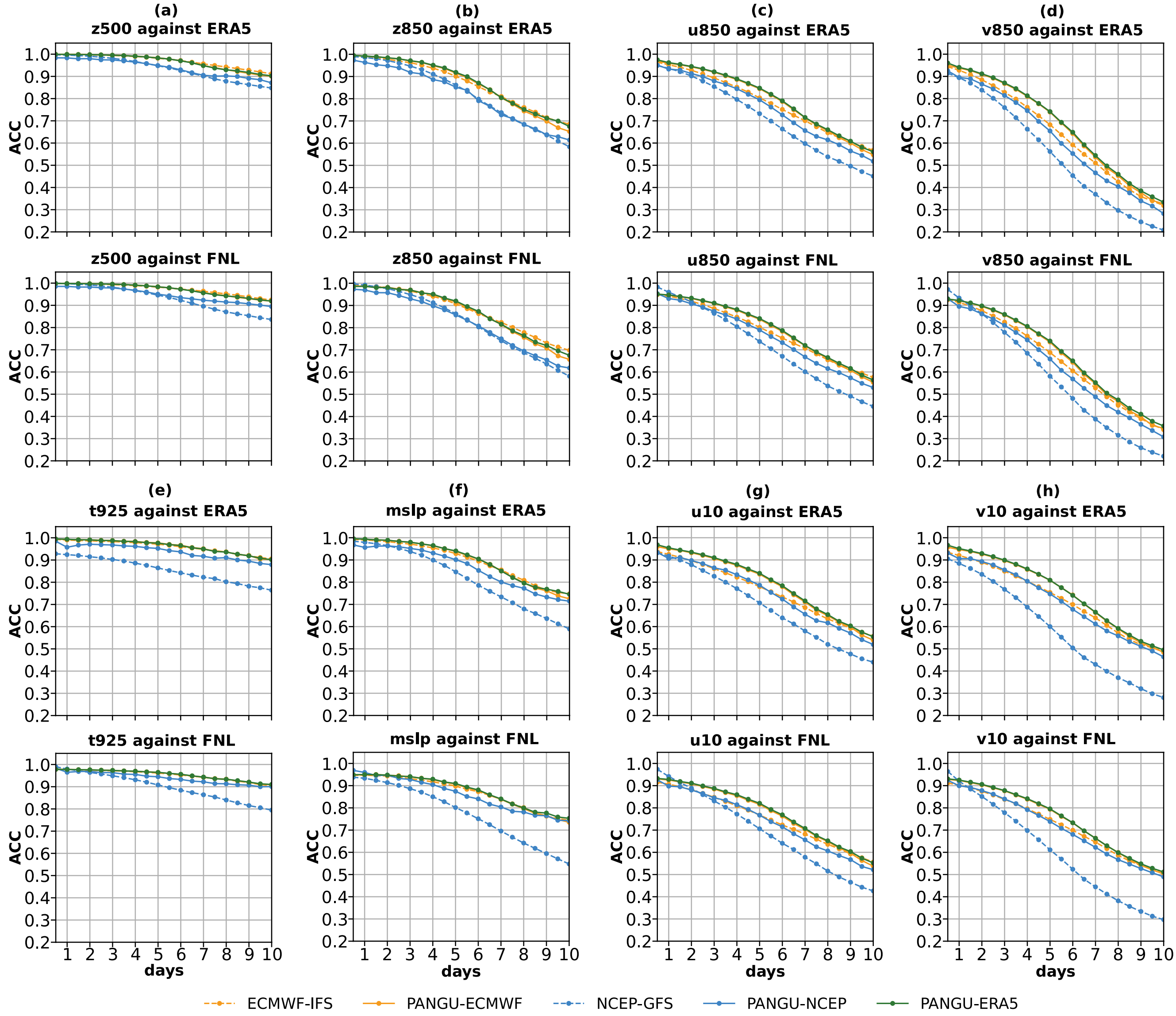

Figure 3. Same as Figure 1, but for the anomaly correlation coefficients (ACC).

distribution of ERA5's 10 m wind speeds can be represented by both the Pangu-Weather and NWP models. In regions with lower wind speeds in ERA5, the RMSE of the models' forecasts is relatively smaller, whereas in regions with higher wind speeds in ERA5, the RMSE of models' forecasts is relatively larger. Additionally, the forecast error of the Pangu-Weather model is smaller than that of the NWP model with the same ICs. Furthermore, the differences between the Pangu-Weather and NWP models become increasingly pronounced with longer forecast lead times.

ICs significantly influence the forecasts of both the Pangu-Weather and NWP models. Specifically, forecasts from PANGU-ECMWF and ECMWF-IFS demonstrate lower RMSE compared to those from PANGU-NCEP and NCEP-GFS, respectively. This is possibly related to the higher quality of ICs provided by ECMWF-IFS in comparison to NCEP-GFS, and it may also be associated with the fact that the Pangu-Weather model was trained on the ERA5 data set.

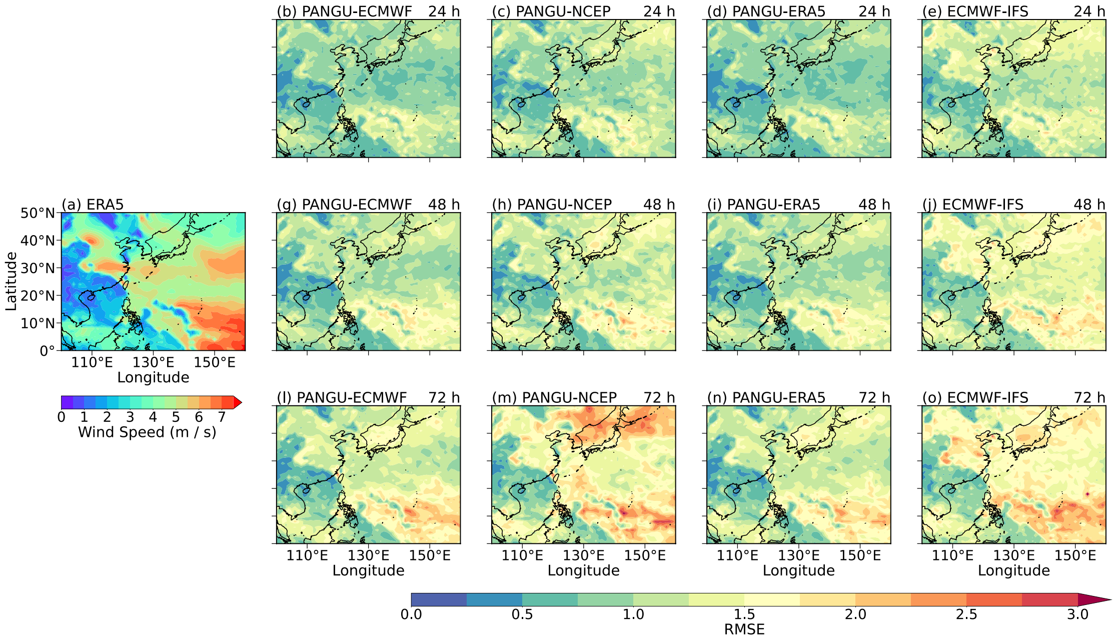

Figure 4. 10-m wind speeds of ECMWF reanalysis (a) and Root Mean Square Error of 10-m wind speed forecasts at the leading times of (b–f) 24 hr, (g–k) 48 hr and (l–p) 72 hr, all averaged from May to October 2023.

4. TC Track and Intensity Forecasts

4.1. TC Track Forecast Errors

Various statistics of the errors of TC track forecasts from the models are compared, as shown in Figure 5. In terms of the mean error of forecasted TC tracks, the PANGU-ERA5 and the PANGU-ECMWF outperform the ECMWF-IFS at all forecast lead times, except for the 12-hr lead time. The advantage of the Pangu-Weather models becomes more pronounced as the lead time increases, especially at the 120-hr, where the difference in mean errors exceeds 50 km. Within 60 hr lead time, the ECMWF-IFS has a smaller median track errors compared to the PANGU-ECMWF, indicating a greater number of samples with smaller errors. Thus, the ECMWF-IFS has higher practical reference usability for TC track forecasts within 60 hr. However, beyond the 72 hr, the PANGU-ECMWF exhibits smaller median and mean track errors than the ECMWF-IFS. Furthermore, the dispersion of track errors supports similar conclusions, that is, beyond the 72-hr leading time, the error range of the track forecasts from the ECMWF-IFS widens and exceeds that of the PANGU-ECMWF.

In Table 2, we present a comprehensive comparison of the Mean Absolute Error (MAE) in track forecasts across models, along with their relative improvement margins against the ECMWF-IFS benchmark. To assess the statistical significance of differences in MAE of TC track forecasts between models, we implemented the bootstrap significant test proposed by Gilleland (2020a, 2020b). Similar with the conclusions in Figure 1, PANGU-ECMWF and PANGU-ERA5 show short term deficit at initial 12 hr forecast lead time, with significant higher MAE (ΔMAE = +12.9% and +15.9%, respectively) than ECMWF-IFS ($p < 0.05$). The performance crossover at forecast lead time larger than 36 hr, statistically significant MAE reductions merge (PANGU-ECMWF: −3.2%, PANGU-ERA5: −3.1%, $p < 0.05$). At 72 hr, error reduction amplifies to −9.2% (PANGU-ECMWF) and −7.4% (PANGU-ERA5). And PANGU-ECMWF achieves peak performance with −16.2% MAE reduction, demonstrating superior handling of steering flow interactions.

Case studies yield further insight into the differences in track forecasts between the Pangu-Weather and the NWP models (Figure 6). The Pangu-Weather model exhibits smaller track deviations and higher stability in long-term forecasts compared to the NWP model with the same ICs. For instance, affected by the steering flow of western Pacific subtropical high and terrain effects, the track forecasts of Typhoon Doksuri (No. 2311, Figure 6a) present considerable challenges. The Pangu-Weather model successfully predicts the movement of Typhoon Doksuri passing through the Bashi Channel and stably moving toward the northeast part of the South China Sea south of

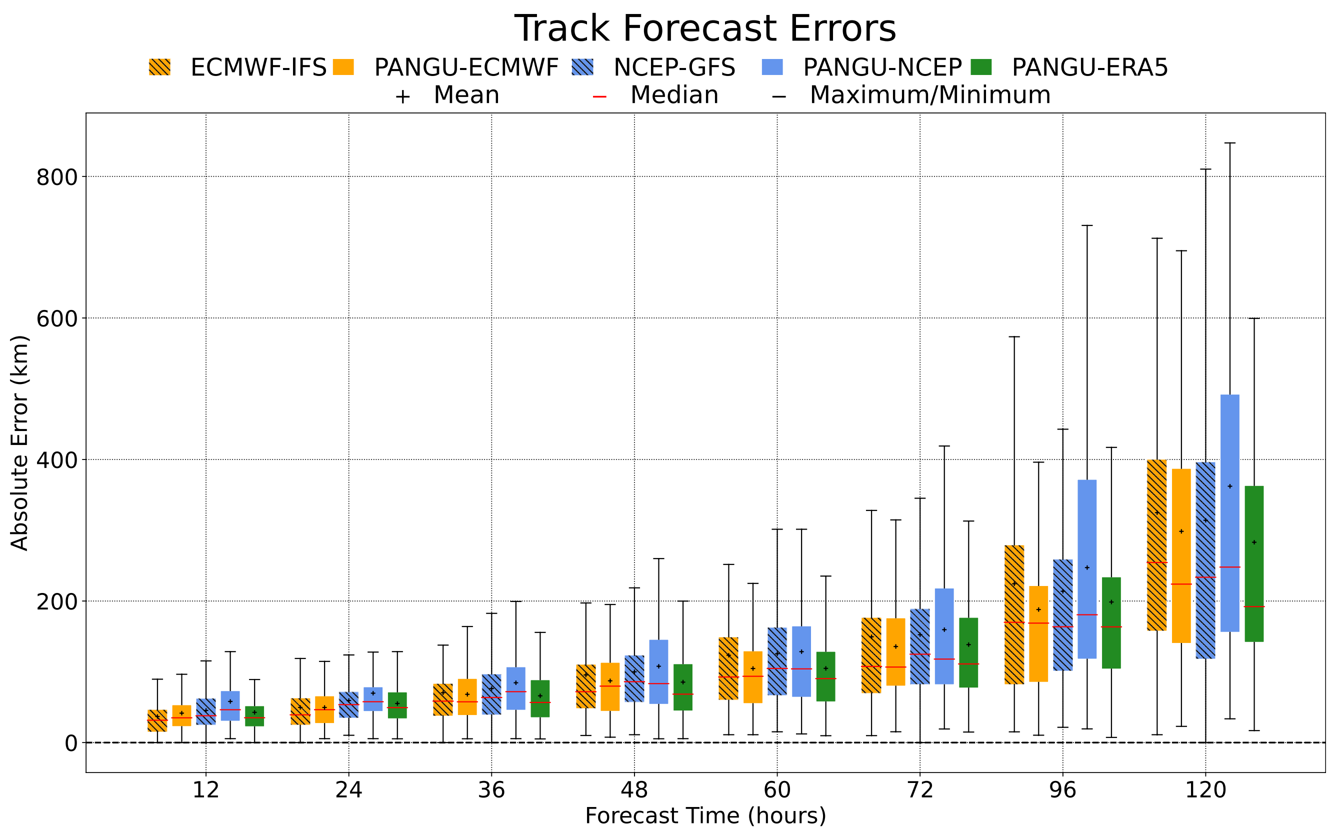

Figure 5. Box plots of the errors of the Northwest Pacific tropical cyclone track forecasts in 2023. "+", "–", the distance between the upper and lower sides of the rectangular boxes, and the upper and lower boundaries of the leading lines on the boxes represent the mean value, median value, dispersion, maximum value and minimum value of the errors.

Taiwan much earlier than the NWP models with the same ICs. Similarly, as shown in Figure 6b, in the forecasts of Typhoon Haikui, all models initially skewed the TC track forecasts northward. However, both PANGU-ECMWF and PANGU-ERA5 made clear southward adjustments after 72 hr. The predicted landfall locations demonstrate a

Table 2
The Mean Absolute Error of Tropical Cyclone Track Forecasts for the Pangu Model and Numerical Weather Prediction Models is Presented, With ECMWF-IFS Serving as the Benchmark Model

LEADTI-MES (hour)	12 hr	24 hr	36 hr	48 hr	60 hr	72 hr	84 hr	96 hr	108 hr	120 hr
ECMWF-IFS	36.8	49.7	70.5	95.6	123.4	149.5	184.0	224.4	263.9	324.9
PANGU-ECMWF	41.6**	49.6	68.2**	87.3**	104.7**	135.7**	161.4**	188.0**	241.5**	298.5**
	12.9%	−0.2%	−3.2%	−8.7%	−15.2%	−9.2%	−12.3%	−16.2%	−8.5%	−8.1%
PANGU-ERA5	42.7**	55.2	66.1**	85.4**	104.8**	138.4**	165.3**	198.6**	244.7**	283.1**
	16.0%	11.2%	−6.2%	−10.6%	−15.0%	−7.4%	−10.2%	−11.5%	−7.3%	−12.9%
PANGU-NCEP	59.2**	72.6**	84.5**	107.8	128.5	159.5	196.1	247.3	302.2	362.2
	60.7%	46.1%	19.9%	12.8%	4.1%	6.7%	6.6%	10.2%	14.5%	11.5%
NCEP-GFS	45.4**	59.4**	76.4	100.0	125.5	152.5	184.7	213.8**	255.1**	314.1*
	23.3%	19.6%	8.4%	4.6%	1.8%	2.0%	0.4%	−4.7%	−3.3%	−3.3%

Note. Only the MAE (km) of track forecast is annotated for ECMWF-IFS, while the other models display their MAE (km) and improvement percentages (%) relative to ECMWF-IFS below their respective track forecast errors. In the table, * indicates that the difference in track forecast MAE between the model and ECMWF-IFS is statistically significant at the 90% confidence level, and ** denotes significance at the 95% confidence level. The gray-shaded models indicate those used as benchmark models for calculating the significance evaluation of TC track and intensity.

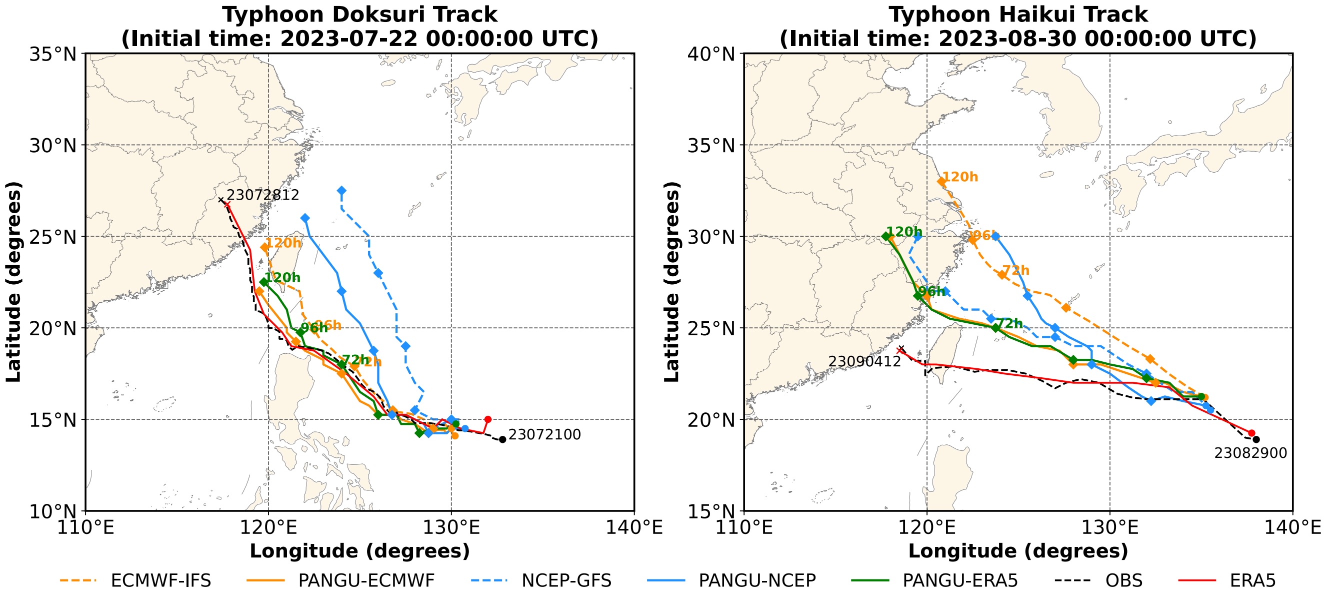

Figure 6. Observed and ECMWF reanalysis (ERA5) tracks of Typhoons Doksuri and Haikui and corresponding forecasts from different models. The black dashed linedenotes the tropical cyclone (TC) track observations from CMA, and the red solid line denotes the TC track derives from the ERA5 data set. Both black dashed and red solid lines are marked with a solid circle and a "×" representing the first and last observations, respectively. The colored dotted lines denote the track forecasts from each model, and the diamond symbols represent the forecast positions at the leading times of 24–168 hr at an interval of 24 hr.

southward shift from approximately 30°N in ECMWF-IFS forecasts to near 26°N. This adjustment corresponds to a nearly 50% reduction in track error compared to both CMA's operational analysis and ERA5 reanalysis data.

Regarding the sensitivity to ICs, the PANGU-ECMWF and the PANGU-ERA5 consistently demonstrate smaller mean and median errors at all forecast lead times compared to the PANGU-NCEP. As depicted in Figure 6, the TC track forecasts of the PANGU-ECMWF and PANGU-ERA5 closely align with both the observations and ERA5 data set, which is consistent with the findings from the assessment of environmental variables in Section 3.2. This suggests that higher-quality ICs in the Pangu-Weather model may contribute to reduce forecast errors, highlighting the model's high sensitivity to ICs and underscoring that uncertainties of ICs are crucial factors impacting model performance.

The movement of TCs is governed by larg-scale interactions across synoptic and planetary-scale systems including he subtropical high, monsoon flows, the Intertropical Convergence Zone (ITCZ), cross-equatorial air currents and etc. As demonstrated in Section 3.1, the Pangu-Weather model exhibits high forecast skills for environment variables, which are closely tied to large-scale atmospheric systems. So, the Pangu-Weather model demonstrates high forecast skills in TC track prediction, primarily due to its advanced capability in simulating key environmental variables, which reflect the characteristics of synoptic and planetary-scale circulation systems.

4.2. TC Intensity Forecast Errors

Figure 7 presents the box plots of the errors associated with TC intensity forecasts. In the following analysis, the TC intensity is defined by the maximum 10-m wind speeds near the TC center. The result indicate that all the models generally underestimate TC intensity; however, the NCEP-GFS model demonstrates the smallest absolute values for both mean and median errors. In contrast, the Pangu-Weather models, regardless of ICs, significantly underestimate TC intensity, with mean and median errors exceeding 10 m s^{-1}. This underestimation is approximately 30% lower than the median and mean observed values of 33 and 34.7 m s^{-1}, respectively. Moreover, the degree of underestimation tends to increase with the forecast lead time. Within the different ICs of the Pangu-Weather models, the intensity errors exhibit similar patterns across most lead times, suggesting that these forecasts are not sensitive to the variations of ICs. Additionally, the error dispersion in intensity forecasts is greater for the Pangu-Weather models than that for NWP models at all forecast lead times, suggesting lower reliability in TC intensity forecasts of the Pangu-Weather model than those of NWP models.

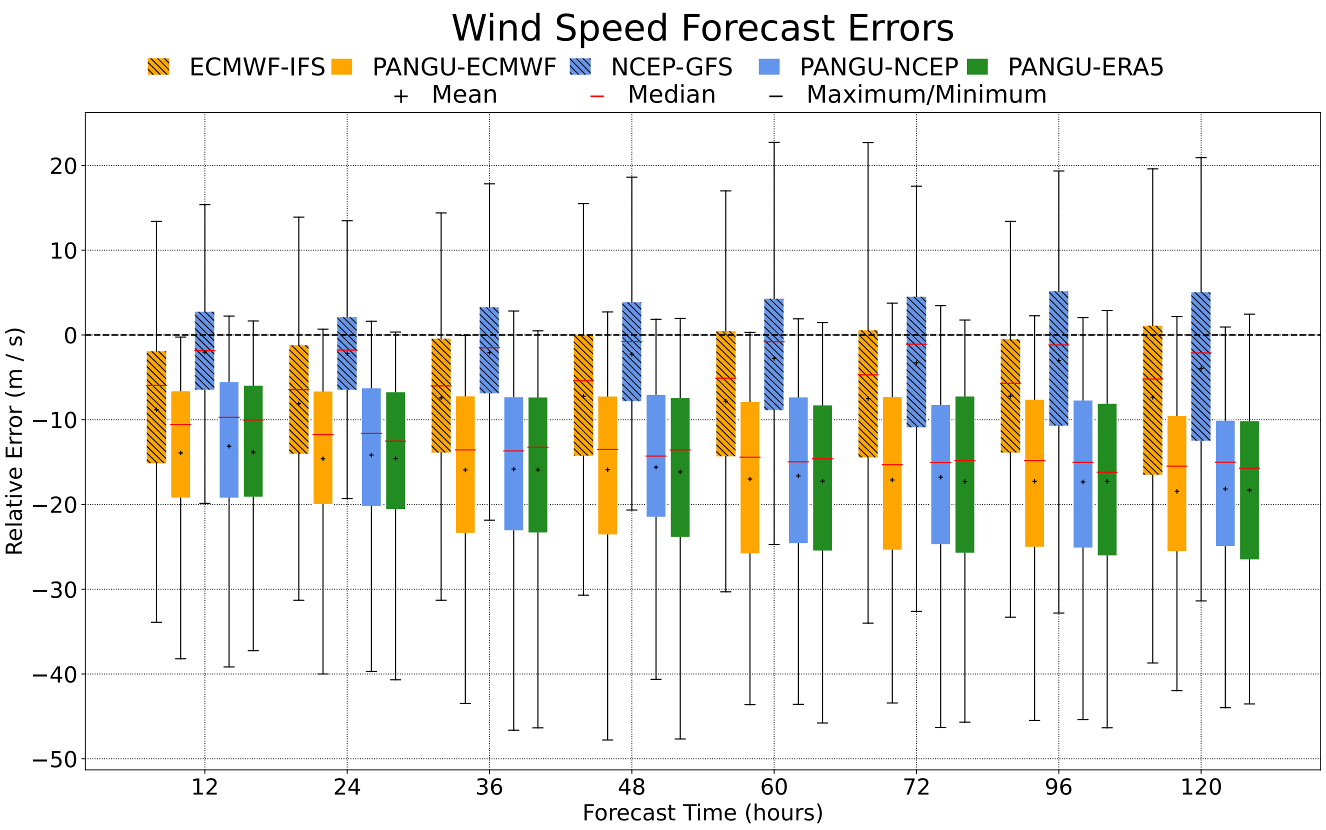

Figure 7. Same as Figure 5, but for the errors of tropical cyclone intensity forecasts.

Table 3 compares the intensity MAE of various models relative to the GFS model as the benchmark. The NWP models maintain consistently smaller MAEs in TC intensity across all forecast lead times, with the MAE of NCEP-GFS being significantly lower than those of ECMWF-IFS ($p < 0.05$). The MAE of the Pangu-Weather models with different ICs comparied with NCEP-GFS is significantly larger ($p < 0.05$) across all forecast lead times, reaching its peak deviation of 142.2% at 36 hr (PANGU-ECMWF). Notably, the differences in MAE of TC intensity forecast among Pangu-Weather models driven by different ICs are relatively minor.

Table 3
The Mean Absolute Error of Tropical Cyclone Intensity Forecasts for the Pangu Model and Numerical Weather Prediction Models is Presented, With NCEP-GFS Serving as the Benchmark Model

LEADTI-MES (hour)	12	24	36	48	60	72	84	96	108	120
NCEP-GFS	6.0	6.1	6.5	6.7	7.4	7.8	8.2	8.9	9.3	9.0
ECMWF-IFS	9.7**	9.5**	9.2**	9.4**	10.2**	10.0**	10.7**	10.2**	9.8**	10.5**
	62.1%	55.2%	42.3%	40.5%	37.1%	28.2%	31.3%	14.4%	5.3%	17.4%
PANGU-ECMWF	13.8**	14.4**	15.7**	15.6**	16.5**	16.5**	17.1**	16.8**	17.2**	17.5**
	−130.2%	−135.5%	−142.3%	−133.0%	−122.2%	−112.5%	−109.3%	−88.7%	−84.8%	−95.5%
PANGU-ERA5	13.7**	14.4**	15.7**	15.8**	16.8**	16.6**	17.0**	16.7**	17.3**	17.4**
	129.2%	134.5%	141.7%	135.9%	125.2%	113.8%	108.9%	87.8%	86.5%	93.9%
PANGU-NCEP	13.0**	14.0**	15.7**	15.3**	16.2**	16.2**	16.5**	16.6**	17.1**	16.9**
	118.2%	128.6%	142.0%	128.2%	117.6%	108.6%	102.2%	87.0%	84.4%	88.8%

Note. Only the MAE (m/s) of track forecast is annotated for NCEP-GFS, while the other models display their MAE (m/s) and improvement percentages (%) relative to NCEP-GFS below their respective track forecast errors. In the table, * indicates that the difference in intensity forecast MAE between the model and NCEP-GFS is statistically significant at the 90% confidence level, and ** denotes significance at the 95% confidence level. The gray-shaded models indicate those used as benchmark models for calculating the significance evaluation of TC track and intensity.

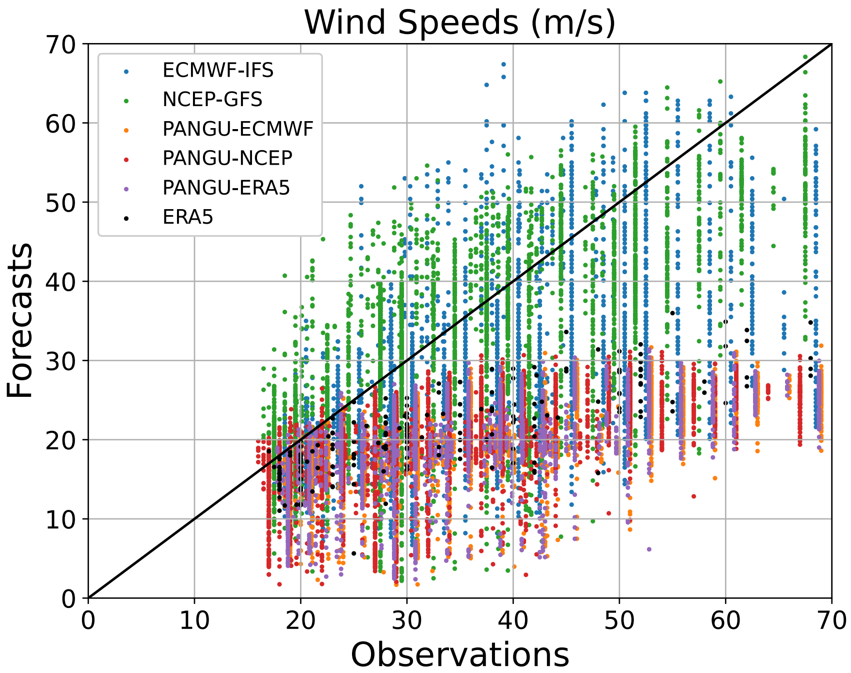

Figure 8. Scatter plots of observed versus forecasted 10-m wind speeds (m s^{-1}) for TCs in Table 1 with observed intensities ≥17 m s^{-1}, aggregated across all initialization times and forecast lead times. Note that to avoid superposition of data from different models, the values of scattered dots are added to a random number within the range from −1 to 1, except for the ECMWF reanalysis.

Figure 8 analyzed the scatter distribution of observed TC intensity against real-time forecasts (including the ERA5) from both the Pangu-Weather and NWP models. The results reveal that for TCs with wind speeds exceeding 18 m s^{-1}, the intensity forecasts from the Pangu-Weather models are lower than observed values. Notably, as observed maximum 10-m wind speeds increases, the corresponding growth rate of forecasted wind speeds in the Pangu-Weather models slows. For TCs with maximum 10-m wind speeds exceeding 51 m s^{-1} (force 16), the Pangu-Weather model caps its forecast at 30 m s^{-1} (force 11), while the ERA5 data caps at 35 m s^{-1}. In contrast, for TCs exceeding 18 m s^{-1}, NWP model forecasts align more closely with the observed values. When considering TCs at or above the intensity grade of TS (Wang et al., 2007), the Pangu-Weather model distinctly underestimates TC intensity compared to NWP models, further underscoring the lower reliability in TC intensity forecasts.

5. TC Structure Forecasts

5.1. TC Horizontal Wind Structure

Typhoons Mawar and Doksuri exhibit long lifespans (over 7 and 4 days, respectively), show stable track forecasts, and remain distant from land for the majority of their duration. These characteristics facilitate a detailed analysis of the horizontal and vertical structure of wind fields. We define the northeast, northwest, southwest, and southeast quadrants as Quadrants 1st to 4th, respectively, centered on the TCs. Figure 9 illustrates the anomalies of maximum 10-m wind speed of TCs in 4 quadrants during different intensity phases, ranging from STY to SuperTY (Wang et al., 2007). The distribution of wind speed anomalies across quadrants is evaluated to compare the performance of the Pangu-Weather and NWP models to assess their capabilities in forecasting the horizontal wind structure of TCs.

Negative wind speed anomalies indicate that the maximum wind speed in a specific quadrant is lower than the average across all quadrants. For instance, in Figure 9a, the NCEP-GFS shows negative wind speed anomalies in the second and third quadrants, indicating that the maximum wind speed in these quadrants is lower than that of other quadrants, with the second quadrant displaying the lowest wind speed.

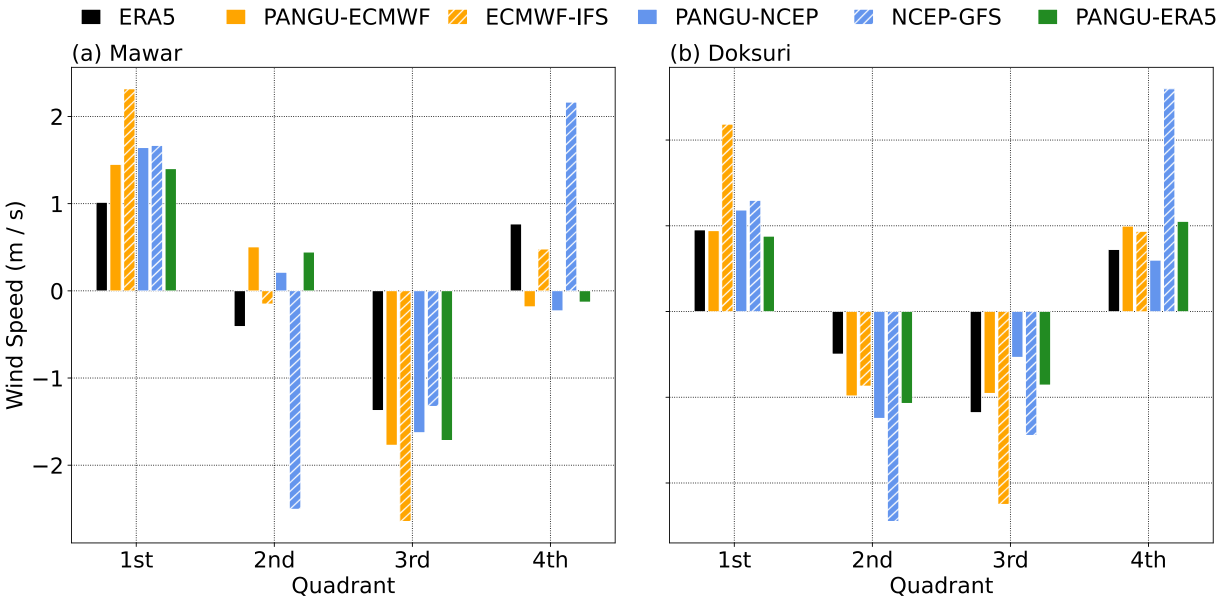

Figure 9. Anomalies of the maximum wind speed of (a) Typhoon Mawar and (b) Typhoon Doksuri in different quadrants within range of 250 km during the periods of severe typhoon to super typhoon from the ECMWF reanalysis data, Pangu-Weather model and numerical models.

The Pangu-Weather model and the ECMWF-IFS exhibit similar anomaly distributions across all quadrants, which are consistent with ERA5 data. The highest and lowest value of the maximum 10-m wind speed of the TCs are found in the first and third quadrants, respectively. Conversely, the NCEP-GFS indicates the maximum and minimum values of wind speed in the fourth and second quadrants, respectively, which diverges from the forecasts of other models. These findings suggest that the maximum 10-m wind speed distribution forecasted by the Pangu-Weather model aligns more closely with ERA5.

5.2. Warm Core and Vertical Wind Structure of TCs

The lateral inflow of water vapor in the boundary layer is a primary energy source for TCs. The release of latent heat from water vapor condensation forms the warm core of a TC, thereby influencing its intensity (Chen & Ding, 1979). Taking Typhoon Mawar (2302) as an example, we analyzed the vertically integrated water vapor flux during its intensification from STY to SuperTY (Figure 10). All models show a continuous increase in water vapor flux from 1200 UTC on May 24 to 1200 UTC on May 26, peaking at 700–900 $\mathrm{kg \cdot m^{-2}\ s^{-1}}$ at 0000 UTC on May 27, followed by a fluctuating decline. The similarity in integrated water vapor flux among the models suggests consistent water vapor transport conditions near the TC in both Pangu-Weather and NWP models.

To investigate the warm core and the vertical wind structure, we analyze vertical cross-sections of temperature anomalies and wind speeds along the line connecting the TC center to the location of the maximum 10-m wind speed of the TC. The average vertical distribution of temperature anomalies for Typhoon Mawar during its intensification from STY to SuperTY reveals that all models successfully capture the warm core structure in the upper and middle troposphere within the typhoon eye. However, discrepancies between NWP models and ERA5 data indicate a more pronounced warm core in the NWP forecasts, with significant positive temperature anomalies between 250 and 50 hPa, and negative values at 600 hPa. Especially, the differences of temperature anomalies between NCEP-GFS and ERA5 in upper troposphere exceed 4 K in the eye region, underscoring a stronger warm core forecast in NWP models compared to ERA5. Conversely, the Pangu-Weather model underestimate the warm core, exhibiting three centers of negative temperature anomalies at 150, 400 and 600 hPa in the eye region, indicating relatively weaker warm core intensity and an overly smooth distribution of temperature anomalies.

Figure 11 illustrates the vertical distribution of wind speeds around the TC center. All models effectively capture the calm wind area in the typhoon eye and the symmetrical the high wind speed band centered over the eye, as well as the rapid decrease of wind speed with increasing radius outside the TC vortex region. The wind speed

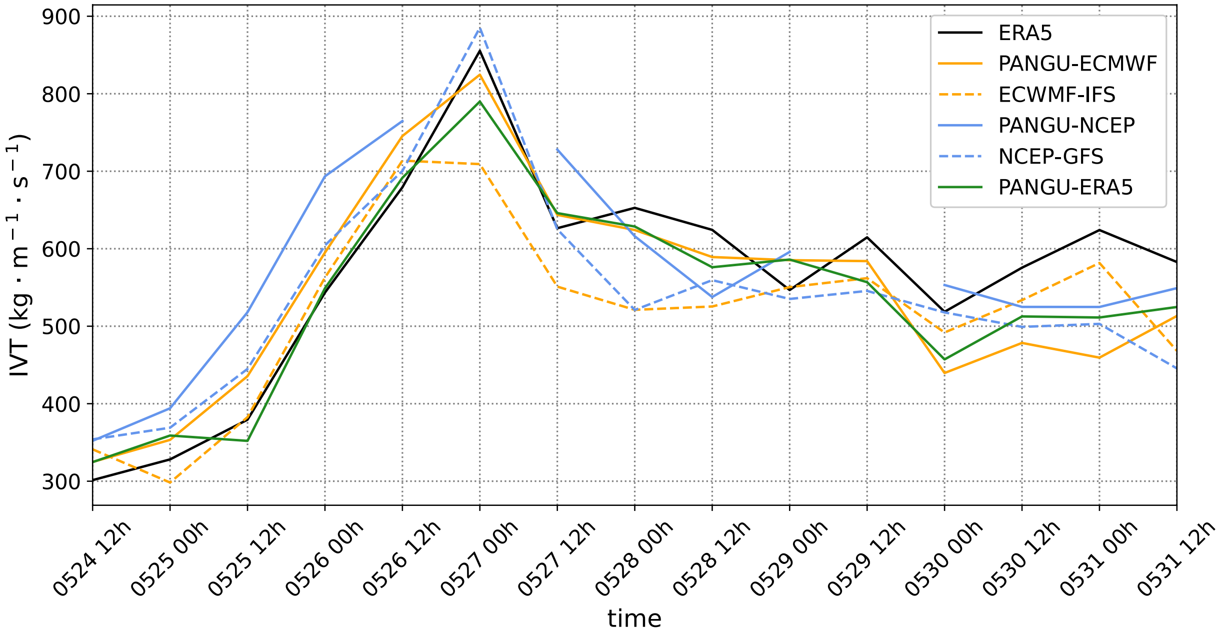

Figure 10. Vertically integrated water vapor flux in the rectangle area (10° × 10°) around the center of Typhoon Mawar during its period of severe–super typhoon. All Pangu-Weather models evaluated in this study are from real-time forecasts, and the discontinuous of PANGU-NCEP is caused by the unstable connection to initial conditions data set download website, where history data set are not available.

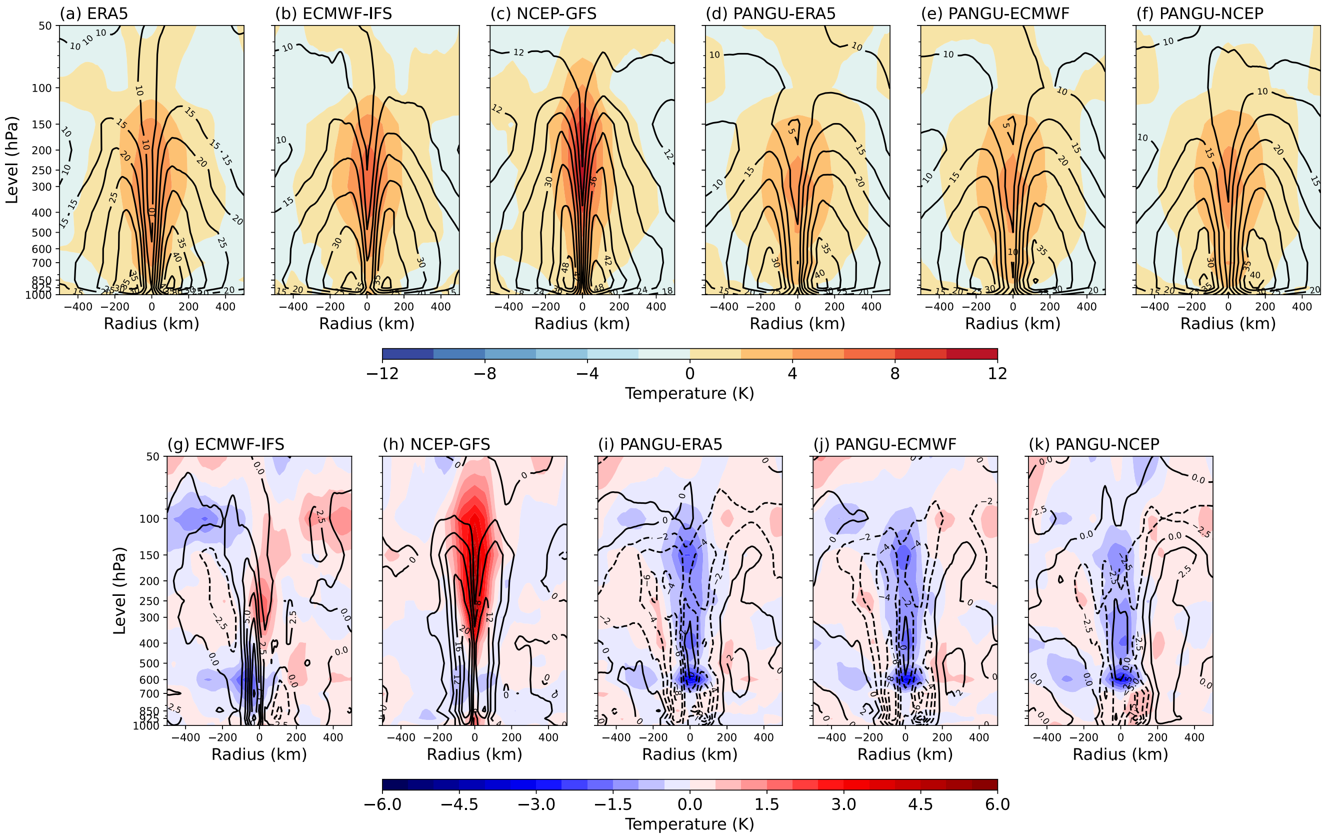

Figure 11. (a–f) Average vertical distributions of the wind speed (contours) overlaid with atmospheric temperature anomalies (shaded) during Typhoon Mawar period of severe typhoon to super typhoon, and (g–k) anomalies of the wind speed (contours) and temperature (shaded) from the model forecasts relative to the ECMWF reanalysis data.

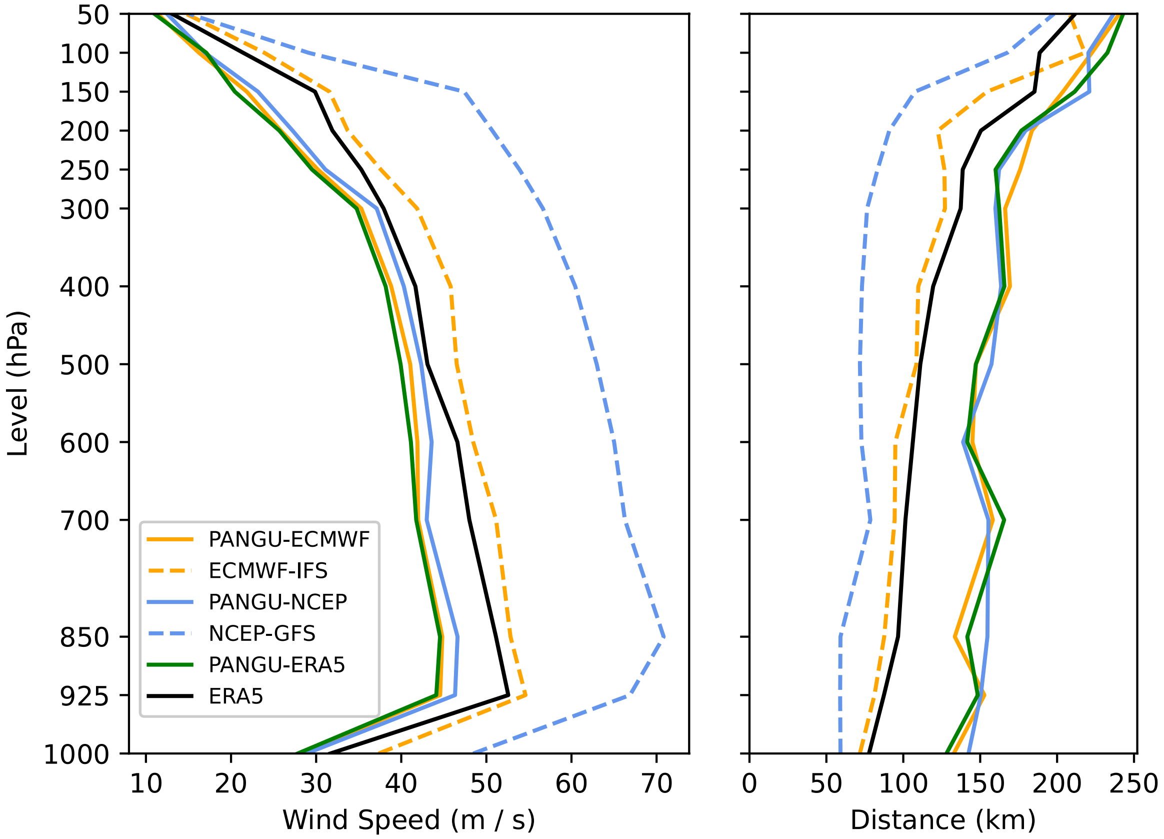

Figure 12. Average maximum wind speed profiles (left panel) and the distance of the max wind speeds (right panel) of Typhoon Mawar during its period of severe typhoon to super typhoon within 250 km. The red triangles denote the observed average intensity of this typhoon during this period.

differences between the forecasts and ERA5 data in Figure 11 demonstrates that the Pangu-Weather model (including the PANGU-ERA5) presents a positive wind speed anomaly in the eye region and negative wind speed anomalies in the TC vortex region, reflecting an overly smooth wind speed distribution similar to the temperature anomalies. Furthermore, the impact of different ICs on the Pangu-Weather model shows negligible differences in both temperature anomalies and wind speeds, which is consistent with the model's insensitivity to ICs in TC intensity forecasts, as discussed in Section 4.2.

The above analysis indicates that, under similar water vapor transport conditions, the Pangu-Weather model predicts a weaker warm core structure and vertical distribution of horizontal winds in the vortex region comparing to NWP models. Furthermore, the Pangu-Weather model exhibits reduced sensitivity to initialization fields in forecasting the vertical structure of TCs.

From Figure 11, notable wind speed deviations from the Pangu-Weather model are present at various levels within the TC vortex region. Specifically, the PANGU-ECMWF shows wind speed deviations of over 8 m s^{-1} lower than the ERA5 data between 700 and 850 hPa, while the deviation is only 4 m s^{-1} near 200 hPa. Further analyis comparing the average vertical distribution of maximum wind speed for Typhoon Mawar during its intensification (Figure 11) reveals a similar one-peak structure among models, with peaks at 850–925 hPa. All models are able to capture the horizontal maximum wind speed within the 850–925 hPa layer, and the horizontal maximum wind speed gradually decreasing as altitude increases. Across all vertical levels, the NCEP-GFS exhibited the highest wind speeds, followed by the ECMWF-IFS and ERA5. The Pangu-Weather model exhibits the lowest maximum wind speed values at all vertical levels, even lower than those of ERA5. And the maximum wind speeds of the Pangu-Weather model with different ICs remained relatively consistent.

Figures 11 and 12 demonstrate that the Pangu-Weather model exhibits weaker forecasting capability for the vertical structure of TCs, with maximum wind speeds across all pressure layers within the typhoon structure being consistently lower than those predicted by high-resolution NWP models. This discrepancy may be linked to the resolution of the ICs. The publicly accessible version of the Pangu-Weather model (Bi et al., 2023) is limited to ingesting coarse-resolution ICs (0.25° × 0.25°), despite the original initial data from ECMWF-IFS and NCEP-GFS having native resolutions significantly finer than 0.25° × 0.25° (Bengtsson & Han, 2024; ECMWF, 2024). The upscaling process required to match the model's input specifications substantially degrades fine-scale structural information of TCs, thereby hindering the prediction of detailed TC dynamics. Additionally, Figure 12 reveals that the maximum wind speeds predicted by the Pangu-Weather model at all vertical levels are lower than those in its training data set ERA5 (with a deviation of approximately 10 m/s at 925 hPa). This discrepancy may be attributed to the inherent characteristics of the model. As a regression-based method, Pangu-Weather tends to forecast values closer to the regional mean rather than capturing extreme events (Bi et al., 2023).

6. Conclusions and Discussions

AIWP models have shown great potential in weather forecasting. Although AIWP models have exhibited superior skills in statistical metrics, whether they can describe the physical characters in disaster weather forecasting as traditional NWP is one of the key criteria for assessing the reliability of their results in practical applications. Additionally, we examine how variations in the initial input fields influence the forecast outcomes. This approach allows us to explore the sensitivity and robustness of the model's predictions, providing insights into its behavior under different ICs. This investigation of the influence of different ICs may help us to analyze the sources of forecast errors in the Pangu-Weather model.

1. In our study, we evaluates the performance of the Pangu-Weather model in severe weather forecasting by comparing the forecasts of the Western Pacific TCs above 17 m/s in 2023 made by the Pangu-Weather model with those generated by NWP models in the operational context. Our analysis focuses on the impacts of different ICs on the Pangu-Weather model in forecast skills of environment fields within the domain of TC activities, TC tracks and intensity, and physical structure of TC. In operational forecasting, the Pangu-Weather model demonstrates statistically superior skill to traditional NWP models in capturing large-scale circulation patterns, particularly in forecasts of TC tracks and environment fields within the domain of TC activities. The discrepancy in forecast errors between the Pangu-Weather and NWP models widens as the forecast lead time increases. The advantage of Pangu-Weather in TC track forecasts becomes more pronounced at forecast lead times larger than 36 hr, with track error reductions exceeding 15% compared to ECMWF-IFS at 96-hr lead time. Comparative analysis of NWP ICs reveals that the ECMWF-IFS exhibits superior performance in environment fields, representing large-scale circulation patterns compared to the NCEP-GFS. Consequently, the Pangu-Weather model demonstrates heightened sensitivity to the quality of ICs in large-scale circulation fields, and the performances in the model with different ICs are likely attributed to the quality of the input ICs.
2. Despite its strengths in large-scale circulation prediction, the Pangu-Weather model exhibits notable limitations in forecasting fine-scale structural characteristics of TCs, particularly intensity and warm-core structures. The model's overly smooth outputs systematically underestimate TC intensity, with mean and median errors of approximately 10 $\mathrm{m{\cdot}s^{-1}}$ in maximum 10-m wind speed—most prominently misclassifying SuperTY as STS. While Pangu-Weather can qualitatively capture the warm-core signature and basic vertical thermal structure near the TC center, the vertical structures are weaker than those produced by high-resolution NWP models. In addition, in respect to maximum horizontal wind speed profile, the Pangu-Weather model is able to capture the horizontal maximum wind speed within the 850–925 hPa layer. However, the model significantly underestimates wind speeds in the lower troposphere, compared with NWP models and ERA5. This discrepancy stems from two interrelated factors: (a) The coarse-resolution ICs ingested by the model degrades fine-scale structural information of TCs, and (b) Its regression-based architecture inherently biases predictions toward smoothed temporal-spatial averages, suppressing extreme values associated with rapidly intensifying TCs. Moreover, the model's low sensitivity to ICs may stems from the insufficient resolution of its input fields, which can not resolve the fine-scale structural features of TCs. Consequently, the quality of ICs-whether high or low—exert negligible influence on the accuracy of TC intensity and structural forecasts.

In summary, the superior forecast skills of Pangu-Weather model in environmental fields and TC track forecasts may be related to the higher-quality ICs, which can effectively leverage its inherent advantages. However, the model exhibits notable limitations in capturing TC vertical structure and intensity evolution. This deficiency is independent of training data or output resolution constraints (Charlton et al., 2024), but likely stems from insufficient representation of mesoscale vortex characteristics in the ICs, particularly the absence of high-fidelity thermal-dynamic profiles critical for resolving convective core dynamics. Moreover, its regression-based architecture inherently biases predictions toward smoothed temporal-spatial averages, suppressing extreme values associated with TC intensity. Such limitations highlight fundamental challenges in AI-driven TC modeling, where physics-agnostic approaches struggle to replicate nonlinear interactions between boundary layer dynamics, diabatic heating, and vortex alignment—processes explicitly resolved by convection-permitting NWP framework. Hybrid models that combine physical processes and data-driven methods, as well as implementing a stacking approach combining general large-scale AIWP models with regional-scale AIWP models optimized for severe weather, might help further improve the accuracy of severe weather forecasting.

One limitation of this study is the focus on TCs above 17 m/s, excluding the weaker cases which may occur in the operational forecasts. Future research could address this by analyzing the weaker TCs to conduct a more comprehensive comparison of the NWP and Pangu-Weather models in operational TC forecasting. In addition, there lacks an in-depth analysis of the reasons why the Pangu-Weather model tend to underestimate the intensity and structure of TCs and their relationship with physical variables. As the primary energy source for TCs is the surface flux of latent and sensible heat from the ocean. Future work should focus on discussing the differences between NWP models and the Pangu-Weather model regarding the surface fluxes of latent and sensible heat from the ocean, with the aim of better understanding the reasons for the large models' underperformance in TCs' intensity forecasting.

Data Availability Statement

Dataset-The forecast data from ECMWF and NCEP was obtained from CMA's internal website. Publicly available data can be accessed through TIGGE (Bougeault et al., 2010; Swinbank et al., 2016). The ERA5 data set is available at (Hersbach et al., 2023). The Real-time TC information from CMA was also obtained from CMA's internal website. The publicly real-time data can be available from (CMA, 2025). *Software*-The Pangu-Weather model evaluated in this article is available at (Bi et al., 2022). The deployment of the Pangu-Weather model, data analysis and the production of all figures and results in this article are using the software Python (Python Software Foundation, 2021).

Acknowledgments

The work described in this paper was supported by National Key Research and Development Program of China (2023YFC3008005), Guangdong Social Development and Technological Collaborative Innovation System Construction Project (2024A1111120024), Youth Innovation Team of China Meteorological Administration (CMA2024QN01) Research Grants Council of Hong Kong SAR, China (HKUST-1301322 and HKUST-16307323), and Special Review Project of China Meteorological Administration (FPZJ2024-094).

References

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956

Ben Bouallègue, Z. Z., Clare, M. C. A., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., et al. (2024). The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context. *Bulletin America Meteorology Society*, *105*(6), E864–E883. https://doi.org/10.1175/BAMS-D-23-0162.1

Bengtsson, L., & Han, J. (2024). Updates to NOAA's unified forecast system's cumulus convection parameterization scheme between GFSv16 and GFSv17. *Weather and Forecasting*, *39*(11), 1559–1570. https://doi.org/10.1175/WAF-D-23-0232.1

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*(7970), 533–538. https://doi.org/10.1038/s41586-023-06185-3

Bi, K., Xie, L., Zhang, H., et al. (2022). 2022 release [Software]. *Pangu-Weather*. https://github.com/198808xc/Pangu-Weather

Bjerknes, V. (1904). Das Problem der Wettervorhersage, betrachtet vom Stanpunkt der Mechanik und der. *Meteorologische Zeitschrift*, *21*, 1–7.

Bougeault, P., Toth, Z., Bishop, C., Brown, D., Chen, D. H., Ebert, B., et al. (2010). The THORPEX interactive grand global ensemble [Dataset]. *Bulletin of the American Meteorological Society*, *91*(8), 1059–1072. https://doi.org/10.1175/2010bams2853.1

Charlton-Perez, A. J., Dacre, H. F., Driscoll, S., Gray, S. L., Harvey, B., Harvey, N. J., et al. (2024). Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of storm ciarán. *npj Climate and Atmospheric Science*, *7*(1), 93. https://doi.org/10.1038/s41612-024-00638-w

Charney, J. G., Fjoertoft, R., & Neumann, J. V. (1950). Numerical integration of the barotropic vorticity equation. *Tellus*, *2*(4), 237–254. https://doi.org/10.1111/j.2153-3490.1950.tb00336.x

Chen, B., Yu, W., Wang, W., Zhang, Z., & Dai, W. (2021). A global assessment of precipitable water vapor derived from GNSS zenith tropospheric delays with ERA5, NCEP FNL, and NCEP GFS products. *Earth and Space Science*, *8*, e2020EA001564. https://doi.org/10.1029/2021EA001796

Chen, K., Han, T., Gong, J., Bai, L., Ling, F., et al. (2023). FengWu: Pushing the skillful global medium-range weather forecast beyond 10 Days lead. *arXiv preprint at arXiv:2304.02948*.

Chen, L., & Ding, Y. (1979). *Introduction to typhoons in the western Pacific* (pp. 156–157). Science Press. (in Chinese).

Chen, L., Du, F., Hu, Y., Wang, F., & Wang, Z. (2023). SwinRDM: Integrate SwinRNN with diffusion model towards high-resolution and high-quality weather forecasting. *arXiv preprint at arXiv:2306.03110*, *37*(1), 322–330. https://doi.org/10.1609/aaai.v37i1.25105

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

China Meteorological Administration. (2025). The real-time typhoon message. Retrieved from http://www.nmc.cn/publish/typhoon/message.html

Dueben, P. D., & Bauer, P. (2018). Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, *11*(10), 3999–4009. https://doi.org/10.5194/gmd-11-3999-2018

ECMWF. (2024). IFS documentation CY49R1 - Part III: Dynamics and numerical procedures. *IFS Documentation CY49R1*. https://doi.org/10.21957/d04fb7a27e

Gilleland, E. (2020a). Bootstrap methods for statistical inference. Part I: Comparative forecast verification for continuous variables. *Journal of Atmospheric and Oceanic Technology*, *37*(11), 2117–2134. https://doi.org/10.1175/JTECH-D-20-0069.1

Gilleland, E. (2020b). Bootstrap methods for statistical inference. Part II: Extreme-value analysis. *Journal of Atmospheric and Oceanic Technology*, *37*(11), 2135–2144. https://doi.org/10.1175/JTECH-D-20-0070.1

Guo, J., Zhang, J., Yang, K., Liao, H., Zhang, S., Huang, K., et al. (2021). Investigation of near-global daytime boundary layer height using high-resolution radiosondes: First results and comparison with ERA5, MERRA-2, JRA-55, and NCEP-2 reanalyses. *Atmospheric Chemistry and Physics*, *21*(22), 17079–17097. https://doi.org/10.5194/acp-21-17079-2021

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., et al. (2023). ERA5 hourly data on single levels from 1940 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. https://doi.org/10.24381/cds.adbb2d47. (Accessed on 2023-06-05 to 2023-11-30).

Jolliffe, I. T., & Stephenson, D. B. (2003). *Forecast verification. A practitioner's guide in atmospheric science* (pp. 128–129). John Wiley and Sons Ltd.

Keisler, R. (2022). Forecasting global weather with graph neural networks. *arXiv preprint at arXiv*:2202.07575v1

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., et al. (2023). GraphCast: Learning skillful medium-range global weather forecasting. *Science*, *382*(6677), 1416–1421. https://doi.org/10.1126/science.adi23

Magnusson, L. (2023). First exploration of forecasts for extreme weather cases with data-driven models at ECMWF. *ECMWF Newsletter, No. 176, ECMWF, Reading, United Kingdom*, 8–9. https://www.ecmwf.int/sites/default/files/elibrary/012023/81379-newsletter-no-176-summer-2023.pdf

Meng, X., Guo, H., Cheng, J., & Yao, B. (2022). Can the ERA5 reanalysis product improve the atmospheric correction accuracy of Landsat series thermal infrared data? *IEEE Geoscience and Remote Sensing Letters*, *19*, 7506805. https://doi.org/10.1109/LGRS.2022.3167388

National Centers for Environmental Prediction. (2000). NCEP FNL operational model global tropospheric analyses. *Continuing from July*, 1999. https://doi.org/10.5065/D6M043C6

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., et al. (2022). FourCastNet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint at arXiv: 2202.11214*.

Python Software Foundation. (2021). 4, 2021 release (version 3.10.0) [Software]. *Python Language Reference*. Retrieved from https://www.python.org/

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, *12*(11). https://doi.org/10.1029/2020MS002203

Rasp, S., & Thuerey, N. (2021). Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for WeatherBench. *Journal of Advances in Modeling Earth Systems*, *13*(2). https://doi.org/10.1029/2020MS002405

Richardson, L. F. (1922). *Weather prediction by numerical process* (p. 236). Cambridge University Press, xii +.

Scher, S. (2018). Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. *Geophysical Research Letters*, *45*(22), 12616–12622. https://doi.org/10.1029/2018GL080704

Swinbank, R., Kyouda, M., Buchanan, P., Froude, L., Hamill, T. M., Hewson, T. D., et al. (2016). The TIGGE project and its achievements. *Bulletin America Meteorology Social*, *97*(1), 49–67. https://doi.org/10.1175/BAMS-D-13-00191.1

Wang, B., Xu, Y., & Bi, B. (2007). Forecasting and warning of tropical cyclones in China. *Data Science Journal*, *6*, 723–737. https://doi.org/10.2481/dsj.6.S723

Weyn, J. A., Durran, D. R., & Caruana, R. (2019). Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2680–2693. https://doi.org/10.1029/2019MS001705

Weyn, J. A., Durran, D. R., & Caruana, R. (2020). Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, *12*(9). https://doi.org/10.1029/2020MS002109

Zhang, X. (2018). A GRAPES-based mesoscale ensemble prediction system for tropical cyclone forecasting: Configuration and performance. *Quarterly Journal of the Royal Meteorological Society*, *144*(719), 478–498. https://doi.org/10.1002/qj.3220

Zhang, X., Xu, Y., He, W., Guo, W., & Cui, L. (2024). A comprehensive Review of the oversmoothing in graph neural networks. *Computer Supported Cooperative Work and Social Computing*, *2012*, 451–465. https://doi.org/10.1007/978-981-99-9637-7_33