Journal of the Meteorological Society of Japan

EARLY ONLINE RELEASE

This is a PDF of a manuscript that has been peer-reviewed and accepted for publication. As the article has not yet been formatted, copy edited or proofread, the final published version may be different from the early online release.

This pre-publication manuscript may be downloaded, distributed and used under the provisions of the Creative Commons Attribution 4.0 International (CC BY 4.0) license. It may be cited using the DOI below.

The DOI for this manuscript is

DOI:10.2151/jmsj.2025-018

J-STAGE Advance published date: February 27, 2025

The final manuscript after publication will replace the preliminary version at the above DOI once it is available.

Tropical cyclone track and intensity predictions in the western North Pacific basin using Pangu-Weather and JMA initial conditions

Munehiko Yamaguchi[1,2*], Yasutaka Ikuta[1], Kosuke Ito[2,3], and Masaki Satoh[2, 4]

1: Meteorological Research Institute, Japan Meteorological Agency
2: Typhoon Science and Technology Research Center, Yokohama National University
3: Disaster Prevention Research Institute, Kyoto University
4: Atmosphere and Ocean Research Institute, The University of Tokyo

* Corresponding author address: Munehiko Yamaguchi, Department of Applied Meteorology Research, Meteorological Research Institute, Japan Meteorological Agency, 1-1 Nagamine, Tsukuba, Ibaraki 305-0052, Japan, Email: myamagu@mri-jma.go.jp, Tel: +81-1-29-853-8627, Fax: +81-1-29-855-7240

Abstract

Tropical cyclones (TCs) are a threat to coastal regions in countries and areas situated in the tropics to, at times, mid-latitudes, and their threat is expected to escalate due to factors like global warming and urbanization. This emphasizes imperative need that warnings based on accurate and reliable forecasts be delivered to those who need them in order to prevent or mitigate TC impacts effectively. While conventional Numerical Weather Prediction (NWP) models have traditionally dominated TC forecasting at short to medium range lead times (i.e., up to two weeks), the emergence of Artificial Intelligence (AI) models, i.e., Machine Learning (ML) models trained on global reanalysis, has raised the possibility of such models competing and thus supplementing NWP models. Here, we examine the potential of ML models in operational TC forecasting, comparing them with conventional NWP models. The ML model used in this study is Pangu-Weather and TC forecasts by this ML model are compared with those from the operational global NWP model at the Japan Meteorological Agency, especially focusing on the track. All 64 named TCs for a period of 2021 to 2023 in the western North Pacific basin are verified. Results indicate that the ML forecasts exhibit smaller position errors compared to the NWP model, alleviate the westward bias around Japan, and retain its forecast accuracy for TCs with unusual

paths, offering potential operational utility. Another benefit would be the ability to deliver forecast results to forecasters quicker than before, since the ML model's forecast takes less than a minute. Meanwhile, challenges such as forecast bust cases and TC intensity, which are also present in NWP models, persist. A proposed way to utilize ML models at current operational systems would be to add ML-based track forecasts as one independent member of consensus forecasts.

Keywords tropical cyclone, tropical cyclone track, operational forecasting, artificial intelligence, machine learning

1. Introduction

Tropical cyclones (TCs) are among the most intense atmospheric phenomena, representing a significant threat, particularly to coastal regions in countries and areas situated in the tropics and extending into the mid-latitudes. They can cause great losses of life and property, and have intense social and economic impacts due to strong winds, heavy precipitation, and storm surge. The threat posed by TCs is expected to intensify due to global warming (e.g., Knutson et al. 2019, 2020, Lee et al.

2020), while urbanization, characterized by high concentration of population and wealth in urban areas (United Nations 2019), presents a significant challenge that the impact of TC landfall in such areas would become enormous (Blake et al. 2013, Normile 2019). As exemplified by the Early Warnings for All initiative (EW4All, World Meteorological Organization [WMO] 2022) led by the United Nations, it is essential that warnings based on accurate and reliable forecasts be delivered in a timely manner to those who need them in order to prevent or mitigate the impacts of TCs.

Among various aspects of TC forecasts, the track is particularly important or fundamental. Getting the winds, precipitation, and storm surge associated with TCs right requires a good track forecast. In general, the accuracy of TC track predictions by numerical weather prediction (NWP) models has improved across all TC basins worldwide, and this can be confirmed, for example, by the inter-comparison study conducted by the Working Group on Numerical Experimentations (WGNE) since 1991 (Yamaguchi et al. 2017). The backgrounds of this improvement include the advancement in NWP systems including the development of NWP models and data assimilation systems, the enhancement of observational networks, and the use of advanced supercomputers. Meanwhile, recent studies such as Conroy et al. (2023) and

Landsea and Cangialosi (2018) point out that the rate of improvement in the accuracy of TC track predictions appears to be slowing down, at least for shorter lead times, where we may be approaching theoretical limits.

In the context of diminishing improvement rate in the accuracy of TC track predictions, a new innovation of weather forecasting by Artificial Intelligence (AI) models, which are Machine Learning (ML) models trained on global reanalysis and often called data-driven models, has emerged (e.g., Bi et al. 2022, 2023; Lam et al. 2022, 2023, Chen et al. 2023a,b). Predictions by ML models have been demonstrated to be as accurate as or more accurate than the state-of-the-art physics-based models (i.e., conventional NWP models) such as the Integrated Forecasting System (IFS) of the European Centre for Medium-Range Weather Forecasts (ECMWF). In the examination of TC forecasting, ML models have also showed smaller position errors than those of IFS though the predictions of TC intensity tend to be weaker than those of NWP models and the best track (Bouallègue et al. 2024). TC track forecasts at operational centers are currently based generally on the outputs from NWP models, but the recent improvement in TC track predictions by ML models is remarkable. Therefore, it is important to conduct forecast experiments using ML models and evaluations across numerous TC cases to determine how ML models can be utilized in operational TC forecasts in the future.

When considering the operational use of ML models for TC track forecasts, it is insufficient to verify the forecast tracks for specific cases (i.e., case studies). Thus, in this study, we conduct forecast experiments for many TC cases and compare the TC track forecasts by an ML model with those of an NWP model. This enables us to deepen our understanding of the characteristics of the track forecasts made by ML models and to highlight the differences from predictions made by NWP models. In this study, forecast experiments are conducted for TCs in the western North Pacific basin. The TCs verified are all named TCs in 3 years from 2021 to 2023. The ML model used is Pangu-Weather (Bi et al. 2022, 2023), and its forecast results are compared to those from the Global Spectral Model of the Japan Meteorological Agency (JMA/GSM, JMA 2023, 2024). This study is characterized by its focus on TCs in the western North Pacific, the verification conducted on a large number of cases covering all named TCs in that basin over a three-year period, and the use of operational global NWP model initial conditions instead of reanalysis data as the initial conditions for the ML model.

This paper is organized as follows. Section 2 describes the methodology and data used in this study. Section 3 presents the results of the forecast experiments by the ML model. Section 4 presents a summary of this study.

2 Methodology and data

This study compares two types of TC track forecasts; one is from Pangu-Weather initiated with JMA/GSM initial conditions (hereafter referred to as PNG-W) and the other is from the operational JMA/GSM (hereafter referred to as GSM). To explore the possibility of utilizing ML-based TC forecasts at JMA, it is necessary to run the ML model from initial conditions that are available in a stable and timely manner. Thus, we select the JMA/GSM initial conditions, which are analysis fields created in real time for the initial conditions of JMA/GSM rather than long-term reanalysis data, to initiate PNG-W in this study.

The PNG-W model used in this study is the pre-trained model available online (https://github.com/198808xc/Pangu-Weather). It is trained on the ECMWF Reanalysis 5 (ERA5, Hersbach et al. 2020) dataset with a horizontal resolution of 0.25 x 0.25 degrees in longitude and latitude, spanning the training period of 39 years from 1979 to 2017. It should be noted that no fine-tuning of the PNG-W model involving the JMA/GSM analysis fields or other data are applied. The initial conditions for PNG-W are 5 variables (geopotential, temperature, specific humidity, zonal and meridional winds) at 13 pressure levels (1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, and 50 hPa) and 4 surface variables (mean sea level pressure, temperature at 2 m, zonal and meridional winds at 10 m) with the same horizontal resolution of 0.25 x 0.25 degrees in longitude and latitude.

The GSM used in this study is an operational global NWP model at JMA. In 2021 and 2022, it utilized the spectral triangular truncation 959 with a reduced Gaussian grid system (TL959), corresponding to 0.1875 x 0.1875 degrees in longitude and latitude (JMA 2023). In 2023, a quadratic and reduced Gaussian grid system (TQ959) was adopted, corresponding to 0.125 x 0.125 degrees in longitude and latitude (JMA 2024). In the vertical layers, 128 stretched sigma pressure hybrid levels are used with a model top of 0.01 hPa throughout the verification period in this study. The horizontal resolution of PNG-W is 0.25 x 0.25 degrees in longitude and latitude, so the JMA/GSM fields are interpolated horizontally using bilinear interpolation to match this resolution.

TC track data, i.e., TC position and intensity (minimum sea level pressure), are created from the outputs of mean sea level pressure fields for both PNG-W and GSM. We adopt a tracking method used in the WGNE inter-comparison study (Yamaguchi et al. 2017). A minimum pressure location in the mean sea level pressure field is defined as the central position of a TC. A surface-fitting technique is employed so that the central position is not necessarily on a grid point of the mean sea level pressure fields. First, the locations of pressure minimum points that could be the potential center of the TC are identified from the mean sea level pressure field at each forecast time. The mean sea level pressure at the minimum point must be at least 2 hPa lower than the average mean sea level pressure within a circle of 1000 km radius centered at that point. Additionally, the mean sea level pressure at the minimum point must be the lowest within a circle of 500 km radius from that point. The initial TC central position is defined as the closest point within a 500 km radius from the analyzed TC central position, based on the best-track data, among the candidate points mentioned above. The TC central position at time T + 6 h is defined within a 500 km radius from the initial TC central position. After this, the TC central position is defined within a 500 km radius from the point that is determined by linearly extrapolating the last two positions. The TC tracking ends when appropriate candidate points do not exist.

The TCs verified in this study are named TCs in the western North Pacific basin from 2021 to 2023. The number of named TCs in 2021, 2022, and 2023 are 22, 25, and 17, respectively, so the total number of TCs verified in this study is 64. For these 64 TCs, we evaluate the forecast results up to 5-days ahead, using all forecasts initialized at 0000 and 1200 UTC. For the TC tracking and evaluation, the JMA best track data is used.

3 Results

3.1 Mean position errors

Figure 1 shows the mean position errors of GSM and PNG-W and the number of verification samples for 1- to 5-day forecasts. The mean position errors of PNG-W are smaller than those of GSM throughout the forecast times considered and the differences between them are statistically significant at all five forecast times based on the two-sided 95 % confidence interval (Student's t-test). The improvement rate of the 1- to 5-day forecasts is 8, 19, 18, 15, and 9 %, respectively. As the rate of improvement in operational TC track forecasting has been declining in recent years, especially for short-term forecasts (Conroy et al. 2023, Landsea and Cangialosi 2018), the magnitude of these improvements would be very attractive when considering utilizing them for operational purposes.

3.2 Forecast bust case

The accuracy of the mean position errors of TCs is evident from the verification result shown in Fig. 1. On the other hand, when examining individual forecast cases, there are instances characterized by large TC position errors, known as forecast bust. In this subsection, we focus on cases where the track forecast errors from GSM is particularly large and investigate how the ML model forecasts those particular cases. Typhoon No. 11 in 2023 (HAIKUI), which moved westward over the southern ocean of Japan and made landfall over Taiwan, is a case where not only GSM, but also global NWP models from ECMWF, the U.S. National Centers for Environmental Prediction (NCEP), and the Met Office in the United Kingdom (UKMO) tended to forecast its track further northward than observed. As JMA’s operational TC track forecasts are primarily based on a consensus of the track predictions by the 4 global NWP models mentioned above (i.e., ECMWF, JMA, NCEP and UKMO), the average position error for JMA's 5-day track forecast for Typhoon HAIKUI exceeded 1000 km.

Figure 2a shows the forecasts of GSM and PNG-W initialized at 1200 UTC on August 30, 2023. GSM forecasts a northwestward movement of HAIKUI, while PNG-W forecasts a westward movement, more comparable to the best track. However, when looking at the forecasts by PNG-W initialized 12 hours before and after (Figs. 2b,c), the continuous westward motion is not forecast as in the forecast initialized at 1200 UTC on August 30, 2023 (Fig. 2a). These results suggest that although there is an initial time when PNG-W forecasts the westward movement of HAIKUI, it would be difficult for forecasters at operation to consistently forecast the westward movement of HAIKUI even if they use PNG-W's forecast results at operations because the forecasts change significantly depending on the initial times. As shown in Fig. 1, the accuracy of the track predictions using PNG-W is generally high; however, this does not imply that instances of forecast bust, where the forecast track is significantly off, will disappear. Similar "flip-flop" issue was also observed in a previous study working on Super Typhoon SAOLA in 2023 (Chan et al. 2024).

3.3 Bias in forecast TC positions

What cases, then, does PNG-W improve the track forecasts over GSM? When we look at each forecast case verified in this study, we notice that in many cases the slow bias of GSM after recurvature has improved. Figures 3a,b show the examples of such cases. Figure 4 is a mean bias map of the track forecasts of GSM and PNG-W. The figure illustrates the average direction and magnitude of the errors in the forecast positions relative to the observed positions. This is created using all 3-day forecasts of GSM and PNG-W verified in this study (i.e., all 64 TCs are considered). The westward bias seen in GSM around Japan, which would be associated with the slow bias after recurvature in the context of the steering flow concept, generally improves in PNG-W. This reduction in bias around Japan would be one of the valuable outcomes for forecasters who closely monitor TCs approaching or making landfall over Japan.

Then, we examine the position errors by separating them into along- and cross-track directions to further understand the characteristics of the track forecasts of GSM and PNG-W. Figure 5 shows the results of calculating the track forecast errors in the along- and cross-track directions for GSM and PNG-W for every 24 hours from the 24- to 120-hour forecasts. The along-track direction is calculated from the observed position at the time of the verification and the 6 hours prior to the verification, and the cross-track direction is orthogonal to that direction. Positive (negative) values in the along-track direction verification indicate that the track forecasts have a fast (slow) bias, and positive (negative) values in the cross-track direction verification indicate that they have a bias to the right (left) relative to the along-track direction. The verification results in the along-track direction show that the slow bias seen in GSM is improved in PNG-W. However, looking at the 120-hour forecast, PNG-W has a rather fast bias. The verification results in the cross-track direction show little difference between PNG-W and GSM. These results are consistent with Liu et al. (2024) that showed that the Pangu-Weather model gives the accuracy of predictions for largescale circulation and TC tracks.

Next, we examine the along- and cross-track directions by separating the verification samples by TC motion directions, which we define to be given by the along-track direction. Figure 6 shows the verification results when the direction of TC motion, ϑ, is in the first ($0° \leqq \vartheta \leqq 90°$, hereafter referred to as Q1) and second ($90° \leqq \vartheta \leqq 180°$, Q2) quadrants, respectively. Note that $\vartheta = 0°$, 90°, 180°, and 270° correspond to East, North, West, and South directions, respectively. The number of verification samples for the Q1 (Q2) direction at 24, 48, 72, 96, and 120 hours is 160 (311), 137 (235), 112

(162), 84 (121), and 66 (84), respectively. The verification in the Q2 direction does not reveal any major difference between GSM and PNG-W. On the other hand, the verification in the Q1 direction shows that PNG-W has a reduced slow bias and a reduced bias on the left side of the motion direction compared to GSM. The verification of the Q1 direction is expected to include many cases where TCs move eastward after recurvature in the western North Pacific basin, so the reduction of the slow bias is consistent with the bias maps seen in Fig. 4.

Finally, we perform the same verification, but for different motion speeds. Figure 7 shows the verification results when the TC motion speed, v, is $v < 10$ km/h (slow motion speed), $10 \leqq v < 20$ km/h (medium motion speed), and $20 \leqq v$ km/h (fast motion speed). The number of verification samples for the slow (medium, fast) motion speed at 24, 48, 72, 96, and 120 hours is 111 (214, 174), 92 (163, 135), 73 (119, 98), 63 (88, 63), and 47 (64, 45), respectively. The verification in the fast motion speed subgroup shows that PNG-W has a reduced slow bias and a reduced bias on the left side of the motion direction compared to GSM. The verification of the fast speed motion is expected to include cases where TCs move along the westerly jet after recurvature, so the reduction of the slow bias here is also consistent with the bias map seen in Fig. 4.

3.4 Forecasts for unique TC tracks

Some may argue that NWP models are more accurate for TCs with peculiar paths (e.g., TCs that suddenly change direction or take a looping path) because their forecasts are based on the laws of dynamics and physics under any given circumstance. Then, we examine the track forecasts for five TCs that took peculiar paths during the 3-year period from 2021 to 2023. These five TCs are Typhoons No. 6 (IN-FA) and No. 8 (NEPARTAK) in 2021, No. 11 (HINNAMNOR) in 2022, and No. 6 (KHANUN) and No. 9 (SAOLA) in 2023.

Figures 8a,b,c,d,e show the track forecasts of GSM and PNG-W when the TCs suddenly changed their motion direction or took a circular path during the forecast period. As the figures clearly show, the ML model is generally able to capture abrupt changes in the track and the circular path as well as the NWP model. There is a case where the abrupt changes in the track is not well forecast by the ML model as shown in Fig. 8b.

However, it is true with the NWP model and it does not seem that the ML model only forecasts badly. To confirm that the ML model is at least not worse overall than the NWP model for track predictions of TCs with unusual tracks, we conduct a verification of position errors using the entire forecast tracks over the lifetimes of the five individual TCs. As Fig. 9 shows, the ML model has smaller position errors than the NWP model throughout the forecast times. Thus, it seems unlikely that ML models are less proficient than NWP models for TCs that take an unusual path.

3.5 Consistency of consecutive forecasts

Forecasters issue TC forecasts on a routine basis when TCs are present in the area of responsibility, and the forecast frequency increases when TCs approach or make landfall. The temporal consistency of the TC forecasts is one of the forecaster's concerns in the forecasting process. Thus, it is important to understand how much the forecast locations of TCs tend to change as the initial conditions change, whether in ML or NWP models.

Then, we investigate the extent to which forecast locations change relative to previous forecasts. Figure 10 shows box plots evaluating how far the latest forecast position is compared to the forecast position with the initial time ΔT hours ago for every 24 hours from the 24- to 120-hour forecasts, with ΔT being verified at 12, 24, 36, and 48 hours.

Since this study uses 12-hourly forecasts (i.e., forecasts initialized at 0000 and 1200 UTC), for the verification of 3-day forecasts with ΔT = 12 hours, for example, the distance between the 72-hour forecast position at a certain initial time and the 84-hour forecast position with the initial time 12 hours earlier is calculated. Smaller values on the Y-axis indicate less variation in the forecast TC positions across consecutive forecasts.

With the exception of ΔT = 12, the forecast positions of TCs in PNG-W tend to change less than those in GSM. At ΔT = 24, 36, and 48, PNG-W shows statistically significant continuity of forecast TC positions compared to GSM in the 24- and 48-hour forecasts. These results suggest that ML models may provide more stable forecasts compared to NWP models, especially at short lead times. On the other hand, at ΔT = 12, the forecast positions of TCs from GSM tend to show less variation compared to those from PNG-W, but the tendency is not statistically significant. In order to be more robust regarding the consistency of consecutive forecasts by ML and NWP models, it is important to increase the number of verification cases and also to incorporate verification using other ML models.

3.6 Intensity forecasts

Although the main focus of this study is the verification of TC track forecasts, we briefly discuss the verification results of the intensity forecasts. Figures 11a,b are the mean absolute error and bias of the intensity forecasts in terms of the central pressure (hPa). The intensity forecast errors of PNG-W are larger than those of GSM throughout the forecast times, which is in consistent with previous studies such as Bouallègue et al. (2024) and He and Chan (2024). The bias of PNG-W is highly positive, indicating weaker TC intensity compared to GSM and to observations.

The effectiveness of PNG-W in forecasting TC tracks has been demonstrated in this study, but it has larger errors than GSM with respect to intensity forecasts. This would be partly due to the limitations inherent in ERA5. ERA5 has a horizontal resolution of

0.25 x 0.25 degrees in longitude and latitude which is too coarse to resolve the inner core structures of TCs. To more accurately predict TC intensity, there would be two possible approaches: either using higher-resolution training data or developing specialized ML models that can address the resolution limitations and mitigate the intensity bias.

A possible approach to leverage the advantages of ML models for TC track forecasting while mitigating intensity forecast bias could include the following method. Statistical dynamical models such as the Statistical Hurricane Intensity Prediction Scheme (SHIPS, DeMaria and Kaplan 1994, DeMaria et al. 2014) and the Typhoon Intensity Forecasting scheme based on SHIPS (TIFS, Yamaguchi et al. 2018) are implemented at operational centers including the US National Hurricane Center and JMA. In such statistical dynamical models, environmental parameters that are predictors for the models are computed along the forecast track. Thus, by calculating environmental parameters used in SHIPS and TIFS based on forecast fields from ML models, it would be expected that the forecast accuracy of intensity forecasts improves (in this case, since outputs from dynamical models are not used, the term "statistical-dynamical model" may not be appropriate).

3.7 Computational time

ML models offer an advantage in terms of the production time as the computational cost to run ML models is quite low. In JMA's operational system, for example, it takes about 19 minutes, with 484 *Intel Xeon 8160* CPUs totaling 11616 physical cores, from the start of a GSM job to output the forecast results for the next 5 days (this time does not include time for data assimilation or post-processing such as TC tracking). Meanwhile, the computation of PNG-W up to 5-day ahead takes less than a minute using a single NVIDIA A100. This indicates that the forecast results from the ML model would be available about 18 minutes earlier than GSM. For forecasters busy with operational work, this time difference may be valuable.

4 Summary

In this study, we evaluated the accuracy of TC track forecasts using an ML model by comparing its predictions with those from an NWP model. Using Pangu-Weather as the ML model, forecast experiments were conducted for all 64 named TCs in the western North Pacific basin from 2021 to 2023, and the results were compared with those of

JMA/GSM, a conventional global NWP model operated at JMA. The JMA/GSM initial conditions are used to initiate the ML and the NWP models.

First, the accuracy of the track forecasts by the ML model exceeds that of the NWP model. The improvement rates of the ML model over the NWP model for the 1- to 5-day forecasts are 9, 19, 18, 15, and 9 %, respectively. Considering the decrease in the improvement rates of track forecasts by NWP models, these values are not insignificant. In addition, the ML model is found to be as good as or better than the NWP model at forecasting TCs with unusual paths. However, these results do not imply that the ML model is a panacea, and cases of forecast busts, such as that observed with Typhoon No. 11 in 2023 (HAIKUI), can still occur in the ML model.

Second, the ML improves track forecasts over the NWP model by reducing the slow bias, particularly after recurvature, corresponding to a reduction in the westward bias around Japan. When examining the position errors in the along- and cross-track directions, the ML shows improvements in the along-track direction, especially for TCs moving eastward or at fast speeds. The ML has the advantage that it has an implicit bias-correction as it had the chance to correct the model when comparing to the true state (i.e., analysis fields) during the model training period. As a result, it would be able to effectively reduce the bias. Regarding the temporal consistency of TC forecast positions, the ML model generally provides more stable forecasts compared to the NWP model, especially at shorter lead times, though further verification with additional cases and ML models is necessary to confirm the robustness of these results.

Although the main focus of this study was TC track forecasting, we also examined the intensity forecasts. We observed that the intensity forecasts by the ML model were weaker than the NWP model and the best track, as shown in Bouallègue et al. (2024). This would be primarily due to the limitations inherent in ERA5 whose horizontal resolution is too coarse to resolve the inner core structures of TCs.

A proposed way to utilize ML models at current operational systems would be to add the ML-based track forecasts as one independent member of the consensus forecasts. In the consensus, one might take advantage of the ML model's good performance and put a larger weight on it. Alternatively, one could consider putting a larger weight on the ML model in the post-recurvature track forecasts, taking into account its ability to reduce slow bias. The creation of optimal consensus forecasts is a topic of our next study.

Another advantage may be that forecasts from ML models are available earlier than from NWP models. In the framework of this study, the ML-based forecasts are available approximately 20 minutes earlier. This availability advantage will be significant when it comes to ensemble forecasts.

It is typical for operational centers to produce their TC track forecasts with a consensus approach using multiple NWP models (Conroy et al. 2023). This means that all agencies tend to have similar forecast results since NWP model results are basically available via the Global Telecommunication System known as GTS, the Internet, etc. The new innovation of ML-based forecasting has the potential to change this international standard of adopting the consensus of major NWP model outputs, and it is likely that each operational center will have its own characteristics in the future depending on how it utilizes ML-based forecasts.

Finally, while we evaluated the potential of ML models for operational TC forecasting in this study, we do not intend to claim that the existence of NWP models or their development is unnecessary. Rather, the opposite is true. Reanalysis data are still needed to train ML models, and this is where NWP models and related techniques such as data assimilation are essential. Thus, further development of NWP systems will be important to improve overall forecast accuracy and to improve forecast accuracy on a continuous basis.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 23K26359 and 24K00703. This study was also supported in part by the Moonshot R&D Grant JPMJMS2282-02 from the Japan Science and Technology Agency and JSPS Core-to-Core Program (grant number: JPJSCCA20220001). We used the pre-trained Pangu-Weather model available at https://github.com/198808xc/Pangu-Weather (https://doi.org/10.5281/zenodo.7678849). The authors thank Dr. Hao-Yan Liu for discussions of initial phase of this study.

Data availability

The Pangu-Weather model is available at https://github.com/198808xc/Pangu-Weather. The datasets of JMA/GSM are operationally provided via the Japan Meteorological Business Support Center (http://www.jmbsc.or.jp/en/index-e.html) and are freely available for research purposes.

References

1. Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2022: Pangu-Weather: A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast, preprint at arXiv:2211.02556.

2. Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2023: Accurate medium-range global weather forecasting with 3D neural networks, *Nature*, **619**, 533–538.

3. Blake, E. S., T. B. Kimberlain, R. J. Berg, J. P. Cangialosi, and J. L. Beven, 2013: Tropical cyclone report hurricane sandy (AL182012) 22 – 29 October 2012, *National Hurricane Center*.

4. Bouallègue, B. Z., and Coauthors, 2024: The rise of data-driven weather forecasting: A first statistical assessment of machine learning-based weather forecasts in an operational-like context. *Bull. Amer. Meteor. Soc.*, https://doi.org/10.1175/BAMS-D-23-0162.1, in press.

5. Chan, P. W., Y. H. He, and Y. S. Lui, 2024: Forecasting Super Typhoon Saola and its effects on Hong Kong, *Weather*. (In Print)

6. Chen, K., T. Han, J. Gong, L. Bai, F. Ling, J.-J. Luo, X. Chen, L. Ma, T. Zhang, R. Su, U. Ci, B. Li, X. Yang, and W. Ouyang, 2023a:FengWu: Pushing the Skillful Global

Medium-rangeWeather Forecast beyond 10 Days Lead, preprint at 10.48550/arXiv.2304.02948.

7. Chen, L., X. Zhong, F. Zhang, Y. Cheng, Y. Xu, Y. Qi, and H. Li, 2023b: FuXi: a cascade machine learning forecasting system for 15-day global weather forecast, *npj Climate and Atmospheric Science*, **6**, 1–11.

8. Conroy, A., H. Titley, R. Rivett, X. Feng, J. Methven, K. Hodges, A. Brammer, A. Burton, P. Chakraborty, G. Chen, L. Cowan, J. Dunion, and A. Sarkar, 2023: Track forecast: Operational capability and new techniques - Summary from the Tenth International Workshop on Tropical Cyclones (IWTC-10), *Tropical Cyclone Research and Review*, **12**, 64-80.

9. DeMaria, M., and J. Kaplan, 1994: A statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic Basin. *Weather and Forecasting,* **9**, 209-220.

10. DeMaria, M., C. Sampson, J. Knaff, and K. Musgrave, 2014: Is Tropical Cyclone Intensity Guidance Improving? *Bull. Amer. Meteor. Soc.*, **95**, 387–398.

11. Dunion, J. P., C. Davis, H. Titley, H. Greatrex, M. Yamaguchi, J. Methven, R. Ashrit, Z. Wang, H. Yu, A.-C. Fontan, A. Brammer, M. Kucas, M. Ford, P. Papin, F. Prates, C. Mooney, A. Kruczkiewicz, P. Chakraborty, A. Burton, M. DeMaria, R. Torn, and J. L.

Vigh, 2023: Recommendations for improved tropical cyclone formation and position probabilistic Forecast products, *Tropical Cyclone Research and Review*, **12**, 241-258.

12. He Y. H., and P. W. Chan, 2024: Challenges in the Forecasting of Severe Typhoon Koinu in 2023, *Atmosphere*, **15**, 31.
13. Hersbach H, and Coauthors, 2020: The ERA5 global reanalysis. *Q. J. R. Meteorol Soc.* **146**, 1999–2049.
14. JMA, 2023: Outline of the operational numerical weather prediction at the Japan Meteorological Agency. Appendix to WMO Numerical Weather Prediction Progress Rep., Japan Meteorological Agency, Tokyo, Japan, 252 pp. [Available online at http://www.jma.go.jp/jma/jma-eng/jma-center/nwp/outline2023-nwp/index.htm]
15. JMA, 2024: Outline of the operational numerical weather prediction at the Japan Meteorological Agency. Appendix to WMO Numerical Weather Prediction Progress Rep., Japan Meteorological Agency, Tokyo, Japan, 262 pp. [Available online at http://www.jma.go.jp/jma/jma-eng/jma-center/nwp/outline2024-nwp/index.htm]

16. Knutson, T., S. J. Camargo, J. C. L. Chan, K. Emanuel, C.-H. Ho, J. Kossin, M. Mohapatra, M. Satoh, M. Sugi, K. Walsh, and L. Wu. 2019: Tropical cyclones and climate change assessment: Part I: Detection and attribution. *Bull. Amer. Meteor. Soc*., **100**, 1987–2007.

17. Knutson, T., S. J. Camargo, J. C. L. Chan, K. Emanuel, C.-H. Ho, J. Kossin, M. Mohapatra, M. Satoh, M. Sugi, K. Walsh, and L. Wu, 2020: Tropical cyclones and climate change assessment: Part II: Projected response to anthropogenic warming. *Bull. Amer. Meteor. Soc.,* **101**, E303–E322.

18. Lam, R., A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, A. Pritzel, S. Ravuri, T. Ewalds, F. Alet, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, J. Stott, O. Vinyals, S. Mohamed, and P. Battaglia, 2022: GraphCast: Learning skillful medium-range global weather forecasting, preprint at 10.48550/arXiv.2212.12794.

19. Lam, R., A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, O. Vinyals, J. Stott, A. Pritzel, S. Mohamed, and P. Battaglia, 2023: Learning skillful medium-range global weather forecasting, *Science*, **382**, 1416–1421.

20. Landsea, C. W., and J. P. Cangialosi, 2018: Have We Reached the Limits of Predictability for Tropical Cyclone Track Forecasting?. *Bull. Amer. Meteor. Soc.*, **99**, 2237–2243.

21. Lee, T.-C., T. R. Knutson, T. Nakaegawa, M. Ying, E. J. Cha, 2020: Third assessment on impacts of climate change on tropical cyclones in the Typhoon Committee Region – Part I: Observed changes, detection and attribution, *Tropical Cyclone Research and Review*, **9**, 1-22.

22. Liu, H.-Y., Z.-T. Tan, Y. Wang, J. Tang, M. Satoh, L. Lei, J.-F. Gu, G.-Z. Nie, and Q.-Z. Chen, 2024: A hybrid machine learning/physics-based modeling framework for two-week extended prediction of tropical cyclones. *J. Goephys. Res.*: Machine Learning and Computation. https://doi.org/10.1029/2024JH000207

23. Normile, D., 2019: Deadly typhoon forces Japan to face its vulnerability to increasingly powerful storms. *Science*, doi: 10.1126/science.aaz9495.

24. United Nations, 2019: World urbanization prospects: the 2018 revision (ST/ESA/SER.A/420), *United Nations*.

25. World Meteorological Organization, 2022: EARLY WARNINGS FOR ALL, *World Meteorolgcal Organization*, 55pp.

26. Yamaguchi, M., J. Ishida, H. Sato, and M. Nakagawa, 2017: WGNE intercomparison of tropical cyclone forecasts by operational NWP models: A quarter-century and beyond. *Bull. Amer. Meteor. Soc*., **98**, 2337–2349.

27. Yamaguchi, M., H. Owada, U. Shimada, M. Sawada, T. Iriguchi, K. Musgrave, and M. Demaria, 2018: Tropical cyclone intensity prediction in the western North Pacific basin using SHIPS and JMA/GSM. *SOLA,* **14**, 138-143.

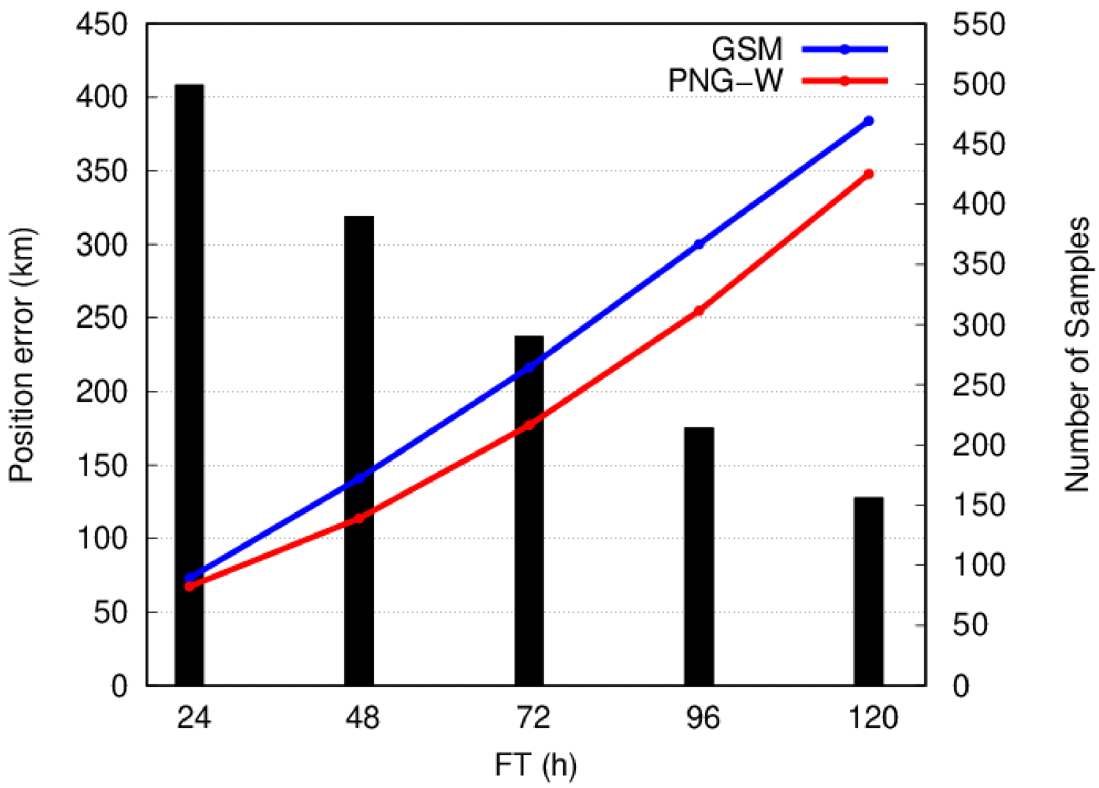

Figure 1. Mean position error of track forecasts of GSM (blue) and PNG-W (red) (km, y-axis on the left). Y-axis on the right represents the number of samples, shown by the black bars. X-axis is the forecast times from 24 to 120 hours. The TCs verified here are all named TCs from 2021 to 2023 (64 TCs in total).

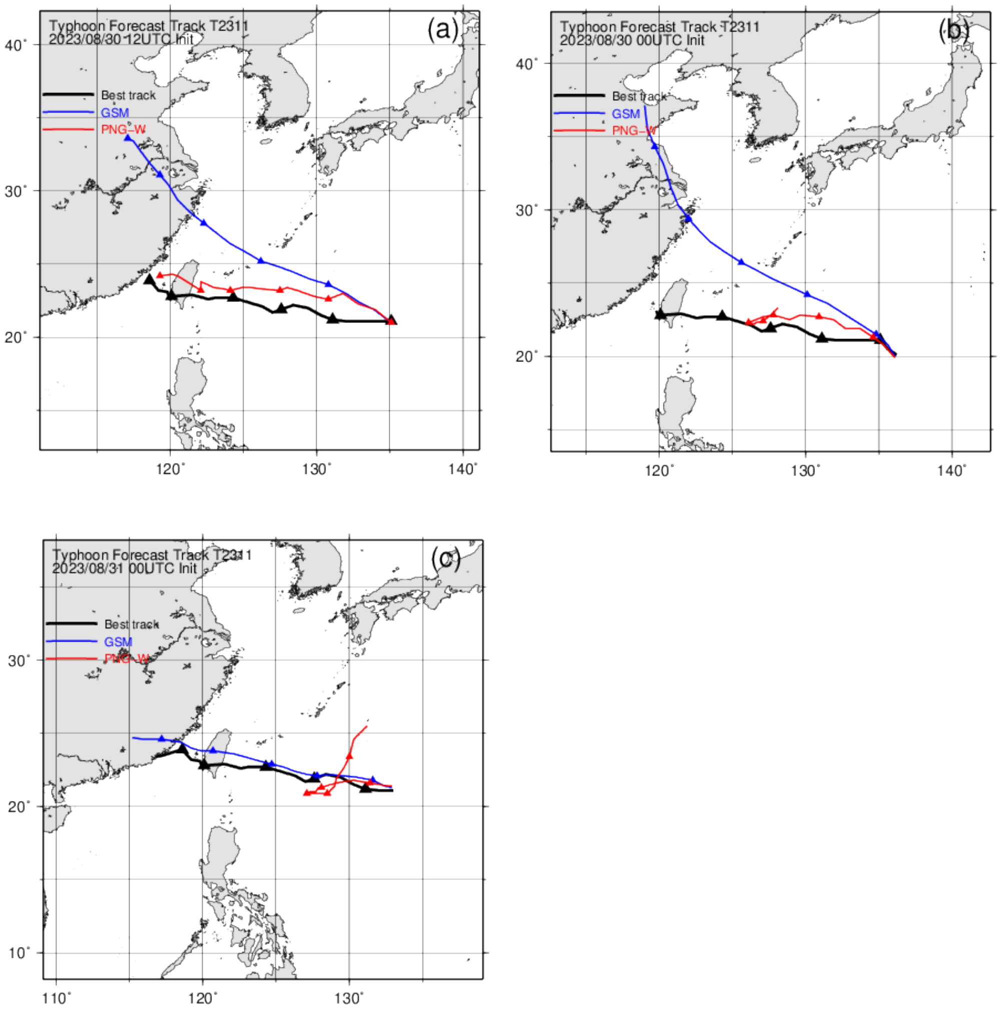

Figure 2. Track forecasts by GSM (blue) and PNG-W (red) for Typhoon HAIKUI. The initial times of the forecast are (a) 1200 UTC of 30 August 2023, (b) 0000 UTC of 30 August 2023, and (c) 0000 UTC of 31 August 2023, respectively. The best track is shown in black. The triangles are plotted every 24 hours at the time of 1200 UTC.

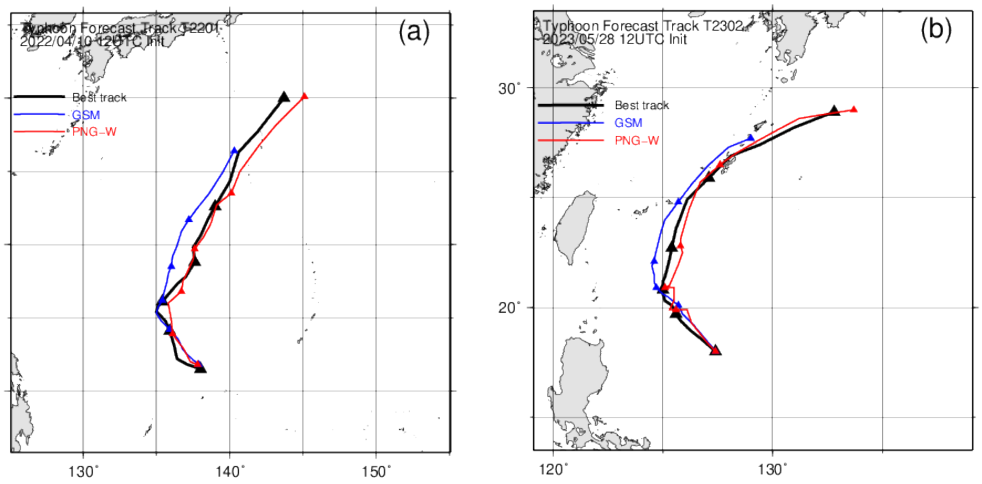

Figure 3. Same as Figure 2, but (a) for Typhoon MALAKAS (Typhoon No. 1 in 2022), initialized at 1200 UTC of 10 April 2022, and (b) for Typhoon MAWAR (Typhoon No. 2 in 2023), initialized at 1200 UTC of 28 May 2023.

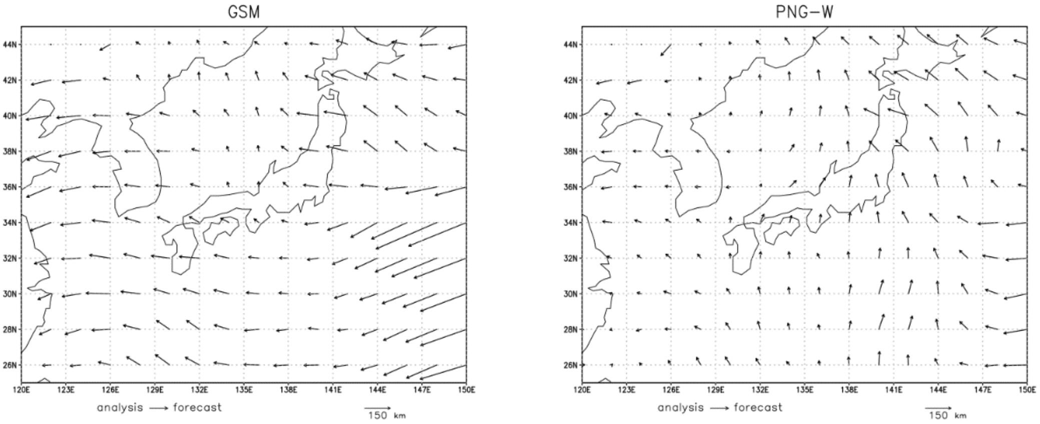

Figure 4. Mean bias of track forecasts by GSM (left) and PNG-W (right). The forecast time verified is 72 hours. The arrow shows the direction of the bias and the length of the arrow shows the magnitude of the bias (see legend on the figures). The TCs verified here are all named TCs from 2021 to 2023 (64 TCs in total).

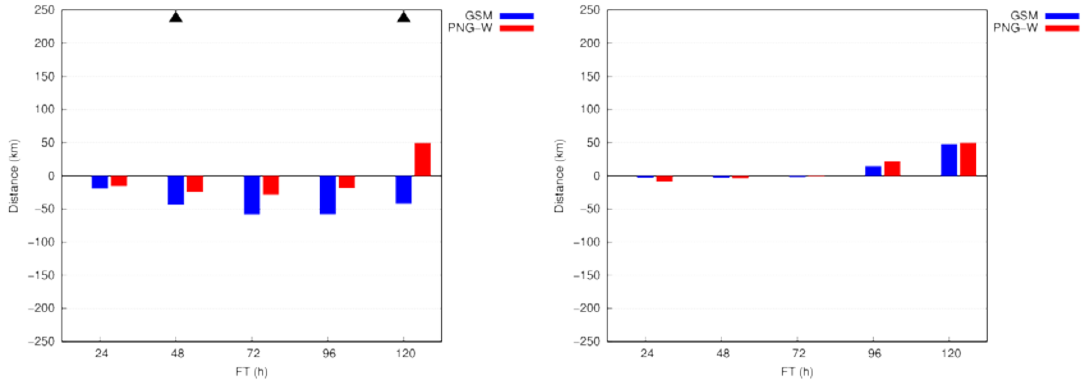

Figure 5. Mean position error of track forecasts in the (left) along- and (right) cross-track directions of GSM (blue) and PNG-W (red). X-axis is the forecast times from 24 to 120 hours. The black triangles represent that the difference between GSM and PNG-W are statistically significant based on the 2-sided 95 % confidence interval (Student's t-test). The TCs verified here are all named TCs from 2021 to 2023 (64 TCs in total).

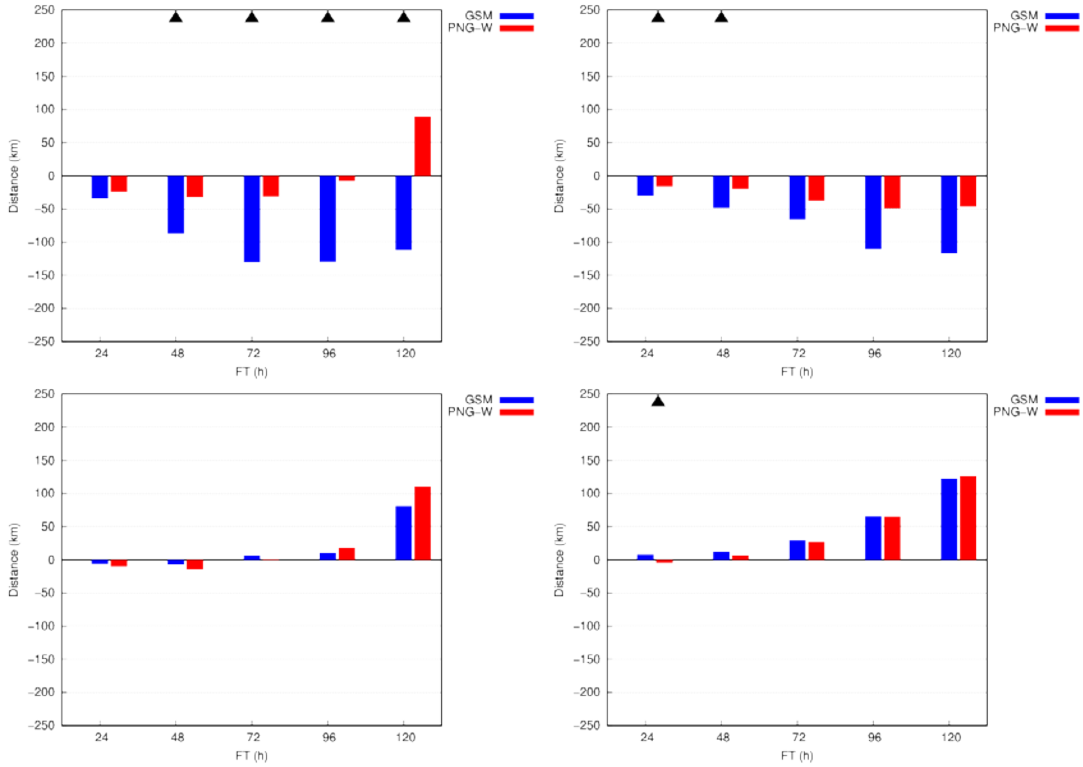

Figure 6. Same as Fig.5, but (top left) and (top right) for the along- and cross-track direction errors when the direction of TC motion, ϑ, is in the first quadrant ($0^\circ \leqq \vartheta \leqq 90^\circ$), respectively, and (bottom left) and (bottom right) for the along- and cross-track direction errors when ϑ is in the second quadrant ($90^\circ \leqq \vartheta \leqq 180^\circ$), respectively.

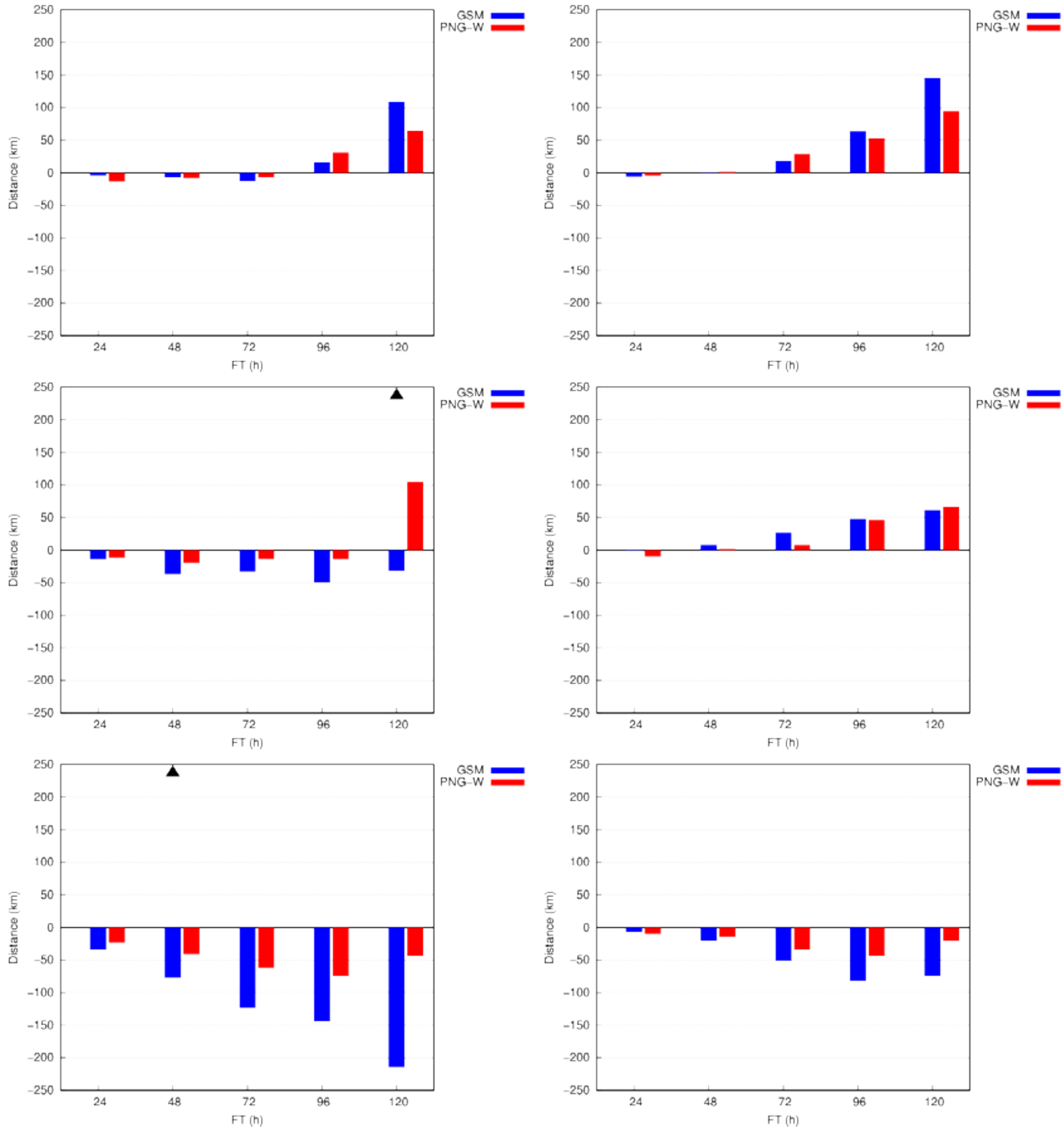

Figure 7. Same as Fig.5, but (top left) and (top right) for the along- and cross-track direction errors when the TC motion speed, v, is $v < 10$ km/h, respectively, (middle left) and (middle right) for the along- and cross-track direction errors when $10 \leqq v < 20$ km/h, respectively, and (bottom left) and (bottom right) for the along- and cross-track direction errors when $20 \leqq v$ km/h, respectively.

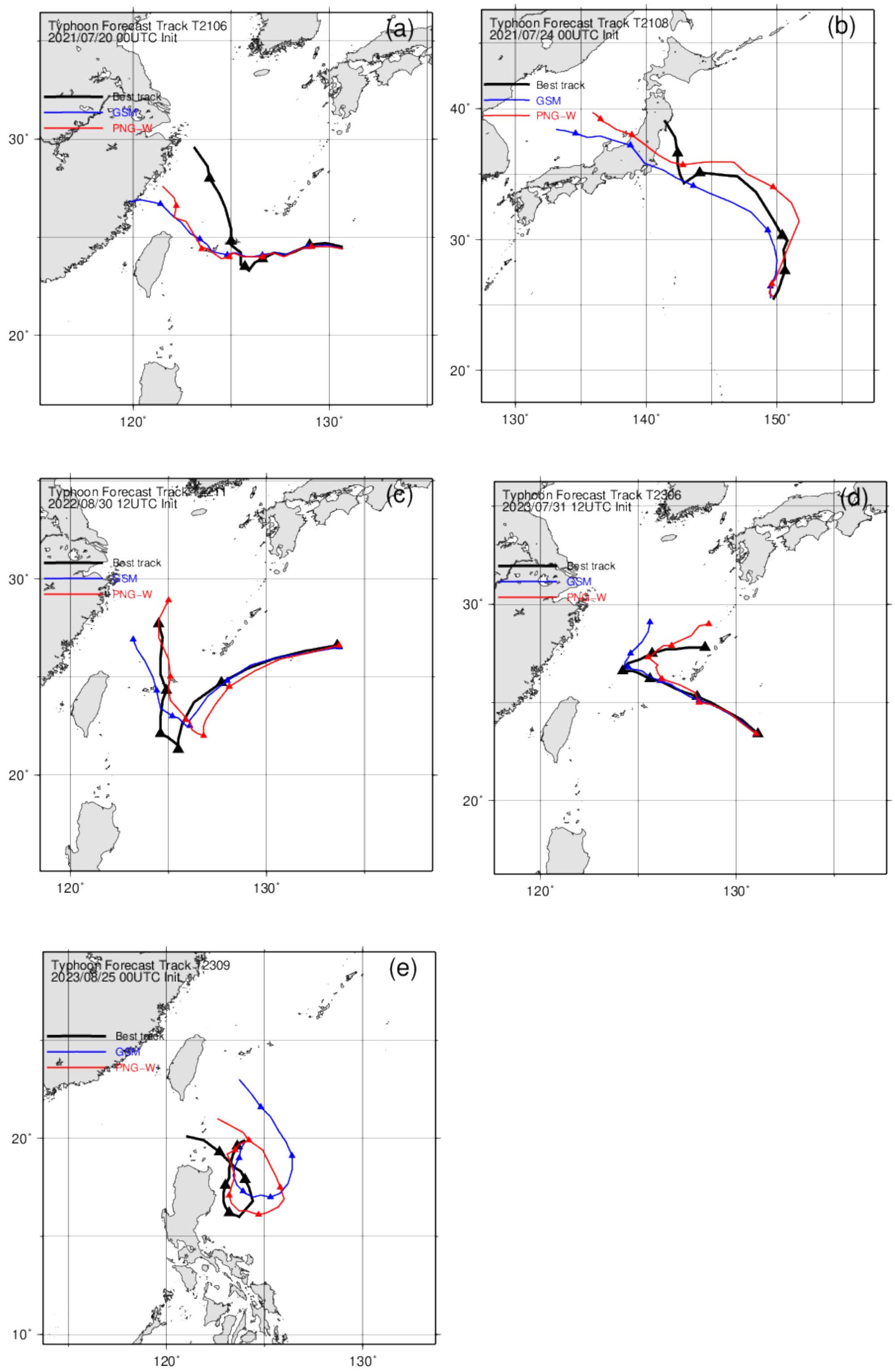

Figure 8. Same as Figure 2, but (a) for Typhoon IN-FA (Typhoon No. 6 in 2021), initialized at 0000 UTC of 20 July 2021, (b) for Typhoon NEPARTAK (Typhoon No. 8 in 2021),

initialized at 0000 UTC of 24 July 2021, (c) for Typhoon HINNAMNOR (Typhoon No. 11 in 2022), initialized at 1200 UTC of 30 August 2022, (d) for Typhoon KHANUN (Typhoon No. 6 in 2023), initialized at 1200 UTC of 31 July 2023, and (e) for Typhoon SAOLA (Typhoon No. 9 in 2023), initialized at 0000 UTC of 25 August 2023.

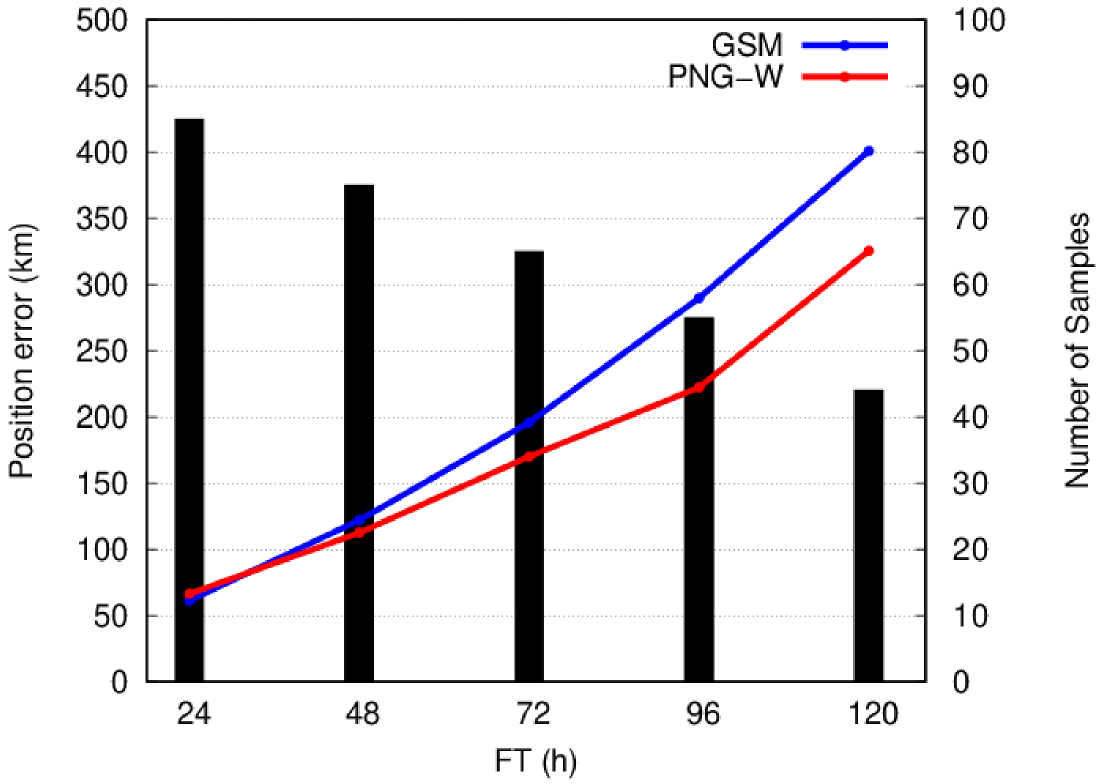

Figure 9. Same as Fig. 1, but the verification is based on the five TCs shown in Fig. 8 only.

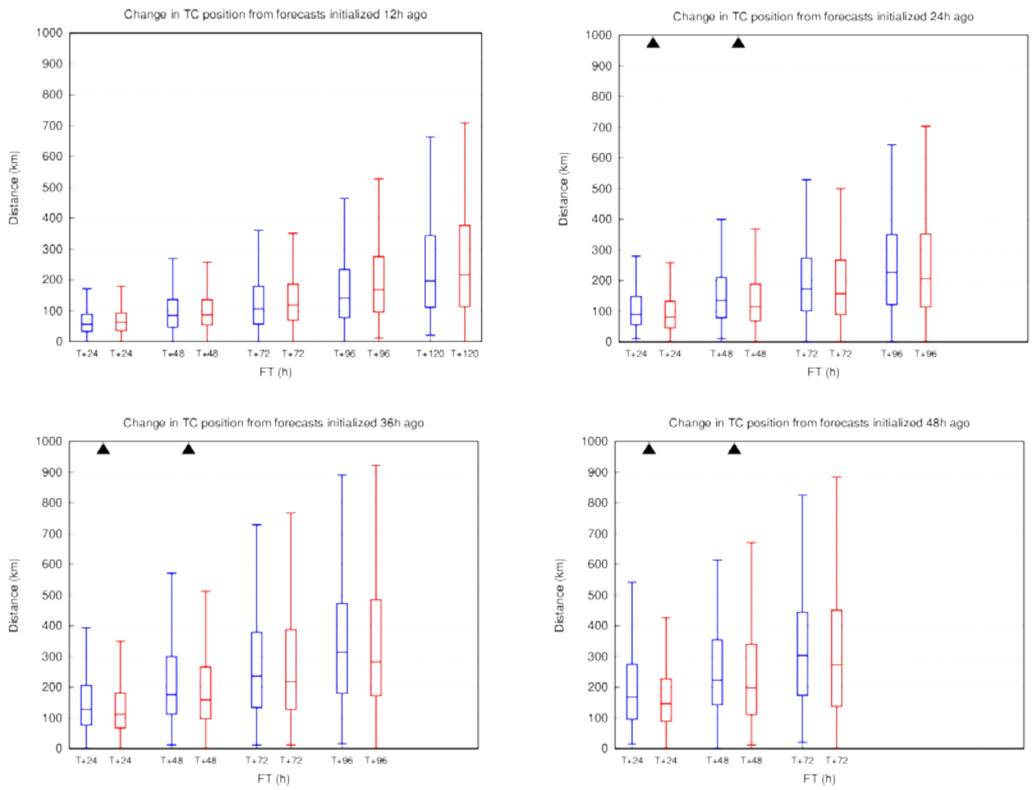

Figure 10. Box plots that show how far the latest forecast TC position is compared to the forecast position with the initial time ΔT hours ago for every 24 hours from the 24- to 120-hour forecasts, with ΔT being verified at (top left) 12, (top right) 24, (bottom left) 36, and (bottom right) 48 hours, respectively. The five sets of the box plots correspond to the verification of 24 to 120 hours forecasts from left to right, with blue representing GSM and red representing PNG-W. The black triangles represent that the difference between GSM and PNG-W are statistically significant based on the 2-sided 95 % confidence interval (Student's t-test). The TCs verified here are all named TCs from 2021 to 2023 (64 TCs in total).

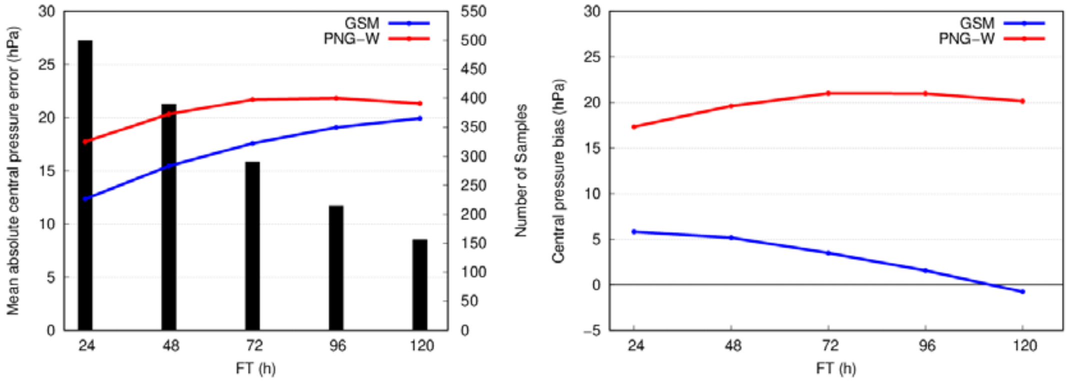

Figure 11. (Left) Mean absolute central pressure error of GSM (blue) and PNG-W (red) (hPa, y axis on the left). Y-axis on the right represents the number of samples, shown by the black bars. (Right) Central pressure bias of GSM (blue) and PNG-W (red) (hPa). X-axis is the forecast times from 24 to 120 hours. The TCs verified here are all named TCs from 2021 to 2023 (64 TCs in total).