https://doi.org/10.1038/s44387-025-00037-3

Enhancing tropical cyclone track and intensity predictions with the OWZP-Transformer model

Zihao Lin[1], Jung-Eun Chu[1] ✉ & Yoo-Geun Ham[2]

Tropical cyclones (TCs) pose significant risks in coastal region; however, traditional numerical models have struggled with forecasting TC intensity, especially rapid intensification (RI). Here we develop an OWZP-Transformer model which incorporates 15 input variables including the Okubo-Weiss-Zeta Parameter (OWZP) as the structural parameter to predict TC track, intensity, and 24 h future intensity change simultaneously over the western North Pacific (WNP). The OWZP-Transformer achieves competitive performance for track prediction and has a 25.1–37.8% improvement in intensity and 37.4–54.8% improvement in 24 h intensity change prediction compared to existing models. It successfully identifies over 60% RI events from the test dataset during 2020–2023. We further perform ablation experiments to evaluate the impact of each predictor category on model performance and investigate the relative contribution of input variables using two explainable feature importance methods. This study highlights OWZP-Transformer as a reliable tool for enhancing both accuracy and efficiency of TC predictions.

Tropical cyclones (TCs) are the most devastating natural disaster, leading to significant casualties and economic damage in coastal regions. Approximately one-third of the world's TCs originate in the western North Pacific (WNP) region, which makes WNP coastal regions vulnerable to TCs. These coastal regions pose a significant risk to local communities and infrastructure[1,2]. Given the potential exacerbation of TC impacts due to climate change, which could lead to more costly events, the necessity for accurate and timely predictions of TC track and intensity has become more critical than ever.

Significant efforts have focused on identifying and forecasting TC tracks and intensity. Modeling approaches can be classified across three categories: statistical, dynamic, and statistical-dynamic[3]. Statistical models use a regression analysis to predict TC tracks based on reanalysis or simulation data. They are computationally efficient and flexible in handling various factors[4], but they struggle to produce reliable predictions in areas with sparse data[5]. Dynamical models, also known as numerical weather prediction models, simulate TC movement using numerical methods based on systems of partial differential equations that express atmospheric dynamics and physics[6]. However, they are computationally intensive and highly sensitive to initial conditions[3]. Statistical-dynamic models first use dynamical models to predict key environmental fields, which are then used as predictors in statistical techniques to quantify uncertainty and identify key factors for TC forecasts. This hybrid approach achieves a balance between the accuracy of TC track and intensity prediction and computational efficiency by leveraging the strengths of both methods. However, they have relied on linear assumptions, which may not accurately reflect nonlinear interactions between variables.

Despite the improvement of predicting TC tracks over the past several decades[7], forecasting TC intensity continues to pose substantial challenges[8]. Particularly, the prediction of rapid intensification (RI) is less skillful than other aspects of TC forecasting. RI is characterized by an increase in intensity of 30 kt (1 kt ≈ 0.5144 m s^{-1}) or more within a 24 h period[9]. In general, strong TCs experience RI in their development[10]. Some studies have indicated that RI is linked to environmental conditions such as vertical wind shear[11], structural variables such as vortical hot towers and radial eddy momentum transport[12], and internal structures of TCs[13]. However, the physical mechanisms driving TC intensity changes remain unclear, which hinders the development of accurate mathematical models to describe the processes explicitly.

Recently, deep learning (DL) has facilitated its application in forecasting the track and intensity of TCs. The layered architecture inherent in DL model allows for the uncovering of latent patterns within datasets using the backpropagation algorithm[14]. DL models require relatively modest computational resources and thus making them suitable for practical applications, offering a compelling balance between accuracy and efficiency. DL models that were built on extensive reanalysis datasets[15] exceeded those

[1]School of Energy and Environment, City University of Hong Kong, Hong Kong, China. [2]Department of Environmental Planning, Graduate School of Environmental Studies, Seoul National University, Seoul, South Korea. ✉e-mail: jungeun.chu@cityu.edu.hk

of traditional numerical weather prediction models. For reference, numerical errors of the 24 h track forecast in 2019 ranged from 87.6 to 107.3 km according to an analysis of operational forecast from five agencies: Japan Meteorological Agency (JMA), China Meteorological Administration (CMA), Joint Typhoon Warning Center (JTWC), Korea Meteorological Administration (KMA), and Hong Kong Observatory (HKO). The mean absolute error (MAE) for 24 h intensity prediction from numerical models, converted to 10-minute averages as per WMO guideline, was approximately 5.9 to 10.5 m s^{-1} [16]. Note that the performance of numerical weather prediction models varies among operational agencies, due to different techniques used to estimate the position and intensity of a TC. Additionally, it is important to acknowledge that while DL models are often perceived as black boxes due to their opaque internal workings, ongoing efforts—such as the use of heat-map techniques or other Explainable Artificial Intelligence (XAI) algorithms— have focused on building their trustworthiness[17].

Previous research on applying deep learning to TC forecasting has mostly focused on track prediction using various models and input datasets. An Long Short-Term Memory (LSTM) model was developed to predict TC tracks using extensive observation data from 1949 to 2012[18]. Their model demonstrated improved prediction accuracy for short-term forecasts, with mean errors of 46.0 km for 6 h predictions and 105.7 km for 24 h predictions, outperforming traditional methods. A Generative Adversarial Network (GAN) with the velocity fields incorporated alongside satellite images was implemented for the next 6 h track prediction[19], demonstrating that the error was reduced by 27.7% from 95.6 km to 69.1 km. This enhancement was particularly evident in forecasting sudden westward or northward track changes. More complex model structures have also emerged. More recently, a DL approach that integrates auto-encoder (AE) and Gated Recurrent Unit (GRU) networks was proposed, incorporating environmental factors[20]. This DL approach demonstrated 15% reduction in averaged absolute position errors in 24 h track forecast than numerical weather prediction models and 13–27% better than other deep learning models. To leverage both 2D and 3D features for predicting TC tracks, a spatiotemporal model named Attention-based Multi GonvGRU (AM-ConvGRU) was developed[21]. This model integrated Residual Channel Attention Blocks (RCAB) to select high-response isobaric planes and a Multi-ConvGRU module to extract large-scale spatial features. The AM-ConvGRU model significantly outperformed other state-of-the-art models, with error reductions of up to 20% in path prediction for TCs in the WNP basin. To improve the forecasting skill of TC tracks with better understanding the mechanisms, a Bidirectional Gate Recurrent Unit (BiGRU) model was developed[22]. This approach prioritized significant temporal data points, leading to better performance in medium- and long-term forecasts. The model outperformed several other state-of-the-art models, achieving a MAE reduction of 12.1% for latitude and 17.6% for longitude compared to models without attention mechanisms.

Compared to track prediction, the mechanisms governing intensity, particularly the rapid intensification (RI) of TCs, remain less understood, making this aspect of forecasting challenging. Therefore, a comprehensive data-based DL model is necessary to extract insights from large datasets[23]. Pan et al. [23] utilized a Recurrent Neural Networks (RNN) to predict TC intensity, leveraging an extensive dataset from the WNP region. Their model, designed to harness the power of DL, achieved a 24 h prediction error of 5.1 m s^{-1}, which was a significant improvement over traditional dynamical model. Yuan et al.[24] employed an LSTM model to forecast TC intensity using a rolling forecast method on large-scale TC datasets. Their model demonstrated superior performance, particularly in long-term predictions, with a MAE improvement of 6% to 10% across various lead times ranging from 24 to 120 h.

Previous studies have strived to predict both track and intensity simultaneously through DL modeling. Tong et al.[25] developed a convolutional LSTM (ConvLSTM) model tailored for predicting the short-term track and intensity of TCs. This model effectively captured both temporal and inter-parameter correlations from the input data and showed better accuracy over traditional LSTM models, particularly in predicting tropical cyclone intensity, with a reduction in RMSE by 19.6% in intensity and by 15.6% in track for 24 lead hours. Na et al.[26] developed an RNN model utilizing an Echo State Network (ESN) to predict TC development and characteristics. The model achieved MAE of 32.73 km for track, 3.84 kt for intensity, and 3.12 mbar for MSLP, outperforming traditional numerical methods over 6 to 24 h forecast periods. Huang et al.[27] developed a Multi-Modal Spatial-Temporal Network (MMSTN) that integrates past 72 h trajectory and intensity data from 1949 to 2016 to perform short-term forecasts of track and intensity. Evaluated on the CMA Best Track Dataset of 2017–2019, MMSTN achieved average 6 h track, pressure, and wind errors of 27.57 km, 1.69 hPa, and 0.95 m/s, respectively, outperforming state-of-the-art deep learning models and the official agency method in most short-term prediction metrics. Jiang et al.[5] introduced a Transformer model specifically tailored for short-term TC track and intensity predictions. Compared to traditional DL approaches like LSTM, this model demonstrated reduced MAEs in TC track by 15.8–27.0% and intensity by 6.0–17.4%.

As for RI prediction, there have been a few research achievements to date. Yang et al.[28] first developed an LSTM model for RI prediction, achieving a Brier skill score improvement of 0.09 to 0.11 compared to operational models in the Atlantic and Eastern North Pacific. Xu et al.[29] introduced a spatio-temporal model named a Spatial Attention Fusing NETwork (SAF-Net), which effectively integrated both domain-expert knowledge (2D-TSDK) and data-driven knowledge (3D-TSDK) for TC structural to predict 24 h intensity changes. Specifically, the model achieved a lower MAE ranging from 3.94 to 4.78 m s^{-1} across different years, compared to traditional numerical approaches. Kim et al. [30] significantly enhanced the application of deep learning in predicting TC intensity by developing the DeepTC model, a CNN tailored for forecasting 24 h intensity changes and RI events in the western Pacific. The DeepTC model demonstrated superior performance compared to operational forecasts, reducing the MAE by approximately 8.9% to 10.2% and increasing the coefficient of determination (R^2) by about 31.7% to 35%. This several results highlighted its effectiveness in improving prediction accuracy. However, previous studies still have the following limitations. Firstly, they typically predict tracks, intensity, or RI separately, lacking a model that simultaneously predicts all three important parameters. Secondly, none of the existing models have achieved 6 h intensity and 24 h intensity prediction errors well below 2 m s^{-1} and 4 m s^{-1}, respectively. Lastly, not many studied have investigated the contributions of various factors to the model's prediction outcomes.

This study aims to develop a DL model that is efficient, accurate, and capable of simultaneously predicting TC track, intensity, and RI events in WNP region. Since our model incorporates an important parameter representing a low-deformation vorticity, known as Okubo–Weiss–Zeta parameter (OWZP)[31], in addition to the multi-head self-attention mechanism of the Transformer, we will refer to our model as OWZP-Transformer (see Methods). We incorporate 15 physical parameters categorized into four distinct groups: basic factors, gradient factors, environmental factors, and structural factors (Table 1), and conduct ablation experiments to quantify the influence and contribution of each category to the model's predictive performance. Lastly, we employ two distinct yet related DL interpretability methods—DeepLIFT[32] and DeepLiftShap[33]—to analyze the contributions of various input factors to the model's prediction outcomes. By leveraging these techniques, we aim to provide a transparent understanding of the model's decision-making process, demonstrating the significance of these factors in enhancing predictive performance.

Results

Performance evaluation of OWZP-Transformer

First, we evaluate the performance of the OWZP-Transformer in predicting TC latitude, longitude, intensity, minimum sea level pressure (MSLP), and 24 h intensity change (dU_RI) using CMA best track and ERA5 reanalysis data from 1980 to 2023 (Table 2). Data from 1980 to 2019 are used for training and validation, while data from 2020 to 2023 are reserved for testing (Table S1). Table 2 displays the overall prediction performance evaluated using MAE, RMSE, R^2 and bias on the test dataset. The MAE for the OWZP-

Table 1 | Input scalar variables for DL model

Category	Variable	Description	Processing
Basic factors	Lon	Longitude (°)	min-max scaling
	Lat	Latitude (°)	
	Intensity	Maximum 2-min average sustained wind speed (m s^{-1})	
	MSLP	Minimum sea level pressure (hPa)	
	dUmax24	Intensity change over past 24 hours (m s^{-1})	
Gradient factors	Trans_speed	Translation speed (km h^{-1})	
	Mov_angle	Moving angle (rad)	
	U_steering	The zonal steering flow component (m s^{-1}) within 200–850 hPa layer	
	V_steering	The meridional steering flow component (m s^{-1}) within 200–850 hPa layer	
Structural factors	OWZ_{850}	Maximum OWZ parameter at 850 hPa ($\times 10^{-6} s^{-1}$)	10°×10° region maximum & min-max scaling
	OWZ_{500}	Maximum OWZ parameter at 500 hPa ($\times 10^{-6} s^{-1}$)	
Scale environmental factors	VWS_{200_850}	Average vertical wind shear (200 hPa minus 850 hPa, m s^{-1})	10°×10° region average & min-max scaling
	RH_{925}	Average relative humidity at 925 hPa (%)	
	RH_{700}	Average relative humidity at 700 hPa (%)	
	SH_{925}	Average specific humidity at 925 hPa (g kg^{-1})	

Processing indicates the process of converting either scalar or 2-dimensional data into a standardized scalar value.

Table 2 | OWZP-Transformer overall performance on test dataset

Evaluation matrix	Track (km)	Latitude (°)	Longitude (°)	Intensity (m s^{-1})	MSLP (hPa)	dU_RI (m s^{-1})
MAE	97.06	0.46	0.78	1.42	2.88	3.12
RMSE	125.50	0.60	1.15	1.82	3.93	3.77
R^2		0.9959	0.9966	0.9817	0.9722	0.8660
Bias		-0.43	-0.64	-0.91	0.28	-2.82

Variables evaluated using root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2) and bias. dU_RI represents the prediction of 24 h intensity change, while all other variables correspond to 6 h predictions.

Transformer ranges from 0.46 to 0.78°, which is equal to 97.06 km using Haversine formula for track predictions, 1.42 m s^{-1} for intensity, and 2.88 hPa for MSLP. Additionally, the MAE for 24 h intensity change is 3.12 m s^{-1}. This difference between intensity and dU_RI arises because the intensity prediction for the next time step (6 h ahead) relies heavily on its autocorrelation, whereas the dU_RI prediction involves next 24 h (next 4 time steps), which depends more on the interactions with other physical factors. The R^2 values are higher than 0.97 for most predicted variables, except for dU_RI, which has an R^2 of ~0.87. This performance is comparable with operational forecasts for TC tracks and superior for TC intensity, which generally has MAE higher than 4 m s^{-1}. The bias values for both track and intensity are negative, indicating that the model exhibits a slight systematic underestimation of TC intensity. Based on an analysis of the 56 TCs in the test dataset, 94.6% of TCs show a southward bias in latitude and 85.7% show a westward bias in longitude, indicating that the track bias is directionally southward in latitude and westward in longitude. Overall, the prediction results are promising especially for TC intensity prediction, highlighting the OWZP-Transformer's strong multi-task prediction capabilities for TCs.

To evaluate the performance of OWZP-Transformer relative to existing DL models, we compare our results with those from two recent studies: TCN model[34], which is trained on basic factors using data from 1980 to 2017 for training and validation and from 2018 to 2021 for testing, and Basic-Transformer model[5], which incorporates both basic and gradient factors, likewise trained on the 1980–2017 data and tested on the 2018–2021 data (left in Fig. 1). To ensure a fair and rigorous comparison between models, we retrained the OWZP-Transformer using the historical data from 1980 to 2017 and evaluated its performance on the same testing period (2018–2021) as the other models. Under this consistent testing scenario, the OWZP-Transformer achieved RMSE values of 0.32, 0.80, 1.79, and 3.57 for latitude, longitude, intensity, and MSLP, respectively. These results are indicated by the blue-colored triangles on the left of Fig. 1. Overall, the TCN model demonstrates the weakest performance among the three models. While our OWZP-Transformer shows slightly higher RMSE values than the Basic-Transformer in longitude predictions, it achieves the lowest RMSE for latitude predictions. Nevertheless, the OWZP-Transformer still demonstrates highly competitive performance, as its R^2 values (Table 2) for both latitude and longitude predictions exceed 99%, indicating strong predictive reliability. For the intensity and MSLP, the OWZP-Transformer outperforms the others. Notably, while earlier studies did not achieve intensity errors well below 2 m s^{-1}, even the most recent state-of-the-art model that fuses multi-source satellite observations and explicitly incorporates TC convective structural features such as deviation-angle variance (DAV) derived from infrared brightness temperature symmetry[35] reports a 6 h intensity prediction RMSE still exceeding 2 m s^{-1} (4.48 kt ≈ 2.3 m s^{-1}). In contrast, our model achieves an error value less than 1.9 m s^{-1}, which corresponds to 25.1–37.8% improvement than other DL models in the next 6 h intensity prediction. We believe this improvement is linked to the inclusion of structural factors (i.e., OWZP), underscoring the importance of input parameters that effectively capture the physics of intensity change.

We further compare the performance for 24 h TC intensity changes (right panel of Fig. 1). It provides a MAE comparison of the three different models. The first model is DeepTC[30], which is trained on basic factors, gradient factors and environment factors using data from 2000 to 2018 under a leave-one-year-out cross-validation scheme. In this scheme, each target year is held out for testing, the two preceding years serve as validation, and the remaining years are used for training. The second model is SAF-Net[2]

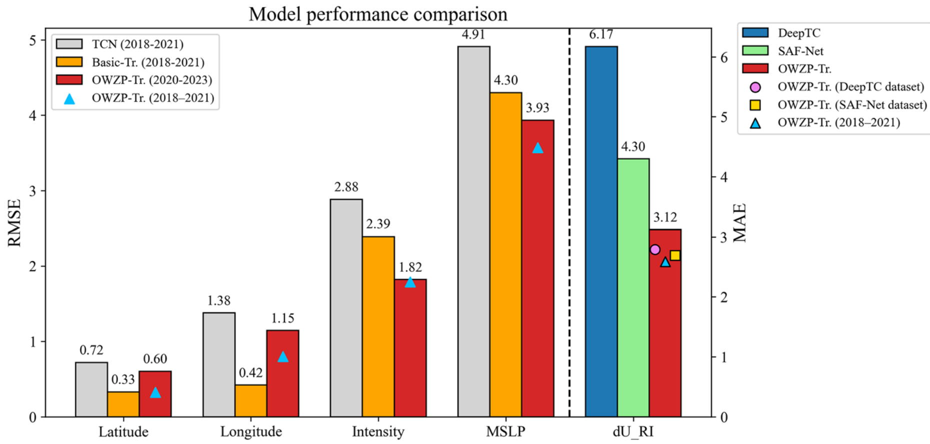

Fig. 1 | Performance of OWZP-Transformer for TC track and intensity predictions compared with other existing DL methods. (Left) RMSE for 6 h latitude (°), longitude (°), intensity (m s^{-1}), and minimum sea level pressure (MSLP, hPa) predictions, in comparison with two existing models, TCN[34] and Basic-Transformer[5]. (Right) MAE for 24 h intensity change (dU_RI, m s^{-1}) prediction, in comparison with DeepTC[30] and SAF-Net[29]. The markers indicate the performance of the OWZP-Transformer when evaluated on the same test data periods as the respective comparison methods.

[9] that is trained on basic factors and gradient factors using data from 2000 to 2014 for training and validation, and evaluated on 2015–2018 as the test period. The third model is our OWZP-Transformer. To ensure a fair comparison, our OWZP-Transformer model was retrained using an identical data period and evaluated on the same test set. In the right panel of the Fig. 1, the circle and square markers indicate the MAE achieved by OWZP-Transformer when tested on the DeepTC and SAF-Net datasets, with values of 2.79 and 2.69, respectively. Both the DeepTC and SAF-Net models already outperform operational forecasts (e.g., CMA, KMA, JMA, JTWC), however, MAE values of dU_RI are still greater than 4 m s^{-1}. The OWZP-Transformer outperforms the existing DL methods in predicting dU_RI with values below 3 m s^{-1}, which is 37.4% to 54.8% lower than existing DL models. In summary, the OWZP-Transformer model performs comparably in TC tracks predictions and outperforms in TC intensity prediction, relative to existing DL models.

Prediction of RI cases

Next, we present the performance of the OWZP-Transformer in predicting TCs that undergo RI. During the test period of 2020–2023, a total 56 TCs were identified over the WNP, of which 28 experienced RI. RI is defined here as a 24 h increase in the CMA's maximum 2-min average sustained wind speed exceeding 29 kt (14.9176 m s^{-1}), a threshold that represents the 95th percentile of over-water 24 h intensity changes in the complete CMA record from 1980 to 2023 in the WNP region (Figure S4), consistent with the approach and value reported by previous study[36]. The OWZP-Transformer successfully detect 61% (17 out of the 28) TCs that experienced RI. Figure 2 displays the tracks and intensities of 17 detected TCs, Figure S5 presents the remaining 11 undetected TCs, and Figure S6 shows the rest 28 TCs that did not undergo RI.

Regarding track prediction, our model captures the tendency of track changes of TCs reasonably well. The model typically predicts TCs with smooth paths accurately, such as Haishen (Fig. 2b), Mindulle (Fig. 2g) and Bolaven (Fig. 2q). Although the model exhibits appreciable position offsets for some TCs with highly convoluted tracks—such as Goni (Fig. 2c), Surigae (Fig. 2e), and Hinnamnor (Fig. 2i)—it nonetheless reproduces both the general trajectory and rapid directional changes quite well for other complex cases, including Chanthu (Fig. 2f), Mawar (Fig. 2l), and Khanun (Fig. 2n). A similar pattern emerges when evaluating TCs with undetected RI events as well as for TCs that did not experience RI: the model maintains high skill for TCs that follow relatively smooth tracks and still delivers competitive accuracy for many of the more intricate trajectories, as shown in Figure S5 and S6. Furthermore, we have identified that Transformer model with the inclusion of basic factors only results in even larger offsets for Khanun (Fig. 2n), implying that inclusion of these factors, i.e., gradient, environmental, and structural factors, is important for improving TC track prediction (figure not shown).

In addition to the tracks, our model effectively captures changes in intensity, including sudden strengthening, weakening, and re-intensification events. For instance, Chanthu (Fig. 2f) rapidly intensified to Category 4 as it moved westward along the 15°N latitude; it then weakened near 126°E before curving northwestward, after which it re-intensified to Category 4 strength (Fig. 2f). Mawar (Fig. 2l) rapidly intensified as it moved northwest, weakened slightly near 14°N, then regained strength while curving northeastward, demonstrating the model's ability to capture multiple phases of intensity change. Similarly, Surigae (Fig. 2e) underwent rapid intensification while moving northwestward, eventually reaching Category 5 intensity, and our model successfully captured this peak intensity, a level of performance that current numerical models have yet to achieve (Fig. 2e). Furthermore, Rai (Fig. 2h) undergoes two RI events, and our OWZP-Transformer successfully captures this double-intensification.

As previously discussed, predicting RI poses a significant challenge for numerical models due to the unclear underlying physical mechanisms. Even state-of-the-art deep learning models struggle to accurately capture all RI cases. It is worth noting that while the typical RI criterion (i.e., >29 knots increase within 24 h) was applied to the IBTrACS data, a more stringent criterion was used for the OWZP-Transformer. For the OWZP-Transformer, we cast intensity forecasting as a regression task, predicting both the continuous 6 h intensity series and the future 24 h intensity change. Thus, we imposed a more stringent criterion due to the availability of two prediction outputs: i) the TC's intensity increase of over 29 knots within 24 h or less using predicted 6 hourly intensity data and ii) future 24 h intensity change (dU_RI) exceeding 29 knots. To quantify the model's detection skill, we employed the Heidke Skill Score (HSS), probability of Detection (POD) and False Alarm Ratio (FAR). Based on the contingency table (TP = 17, FN = 11, FP = 0, TN = 28), the OWZP-Transformer achieves an HSS of 0.61, POD of 0.61 and FAR of 0.00, outperforming DeepTC (HSS = 0.42) and far exceeding the values (HSS below 0.20) reported by other operational agencies[30].This demonstrates the advantages of our model over other numerical model models and existing DL models.

Impact of input variables: ablation study

Based on the 15 input variables categorized into four groups in Table 1, we conducted an ablation study to evaluate the impact of each category on forecasting performance. This was achieved by removing one or several categories and examining the subsequent changes in predictive skill of RMSE. In Fig. 3a-d, the black bar represents the baseline scenario, which uses the full set of 15 predictors evaluated on the test dataset. The other colored bars illustrate the RMSE obtained after removing one or more predictor categories, namely structural (–S), environmental (–E), and gradient (–G) factors. Additionally, the three markers plotted on the –E bars highlight the individual effect of excluding single environmental variables (circle = RH, triangle = SH, square = VWS). Figure 3e presents a normalized

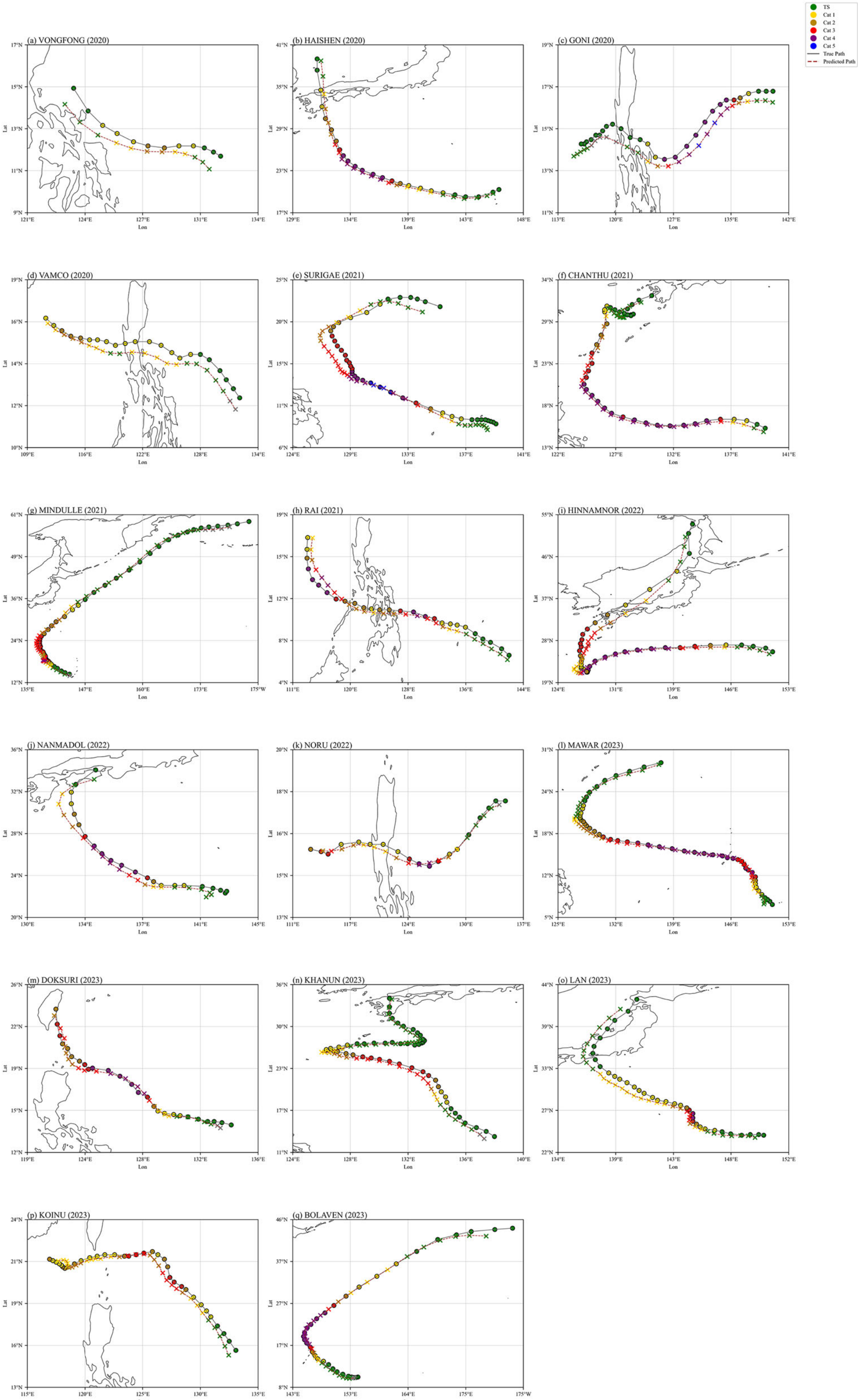

Fig. 2 | OWZP-Transformer prediction results on test dataset for TCs with successfully detected rapid intensification events. The solid lines represent the actual TC path from the IBTrACS, while the dashed lines indicate the predicted path by our OWZP-Transformer model. Colors indicate the intensity of TCs according to the Saffir–Simpson Hurricane Wind Scale classification.

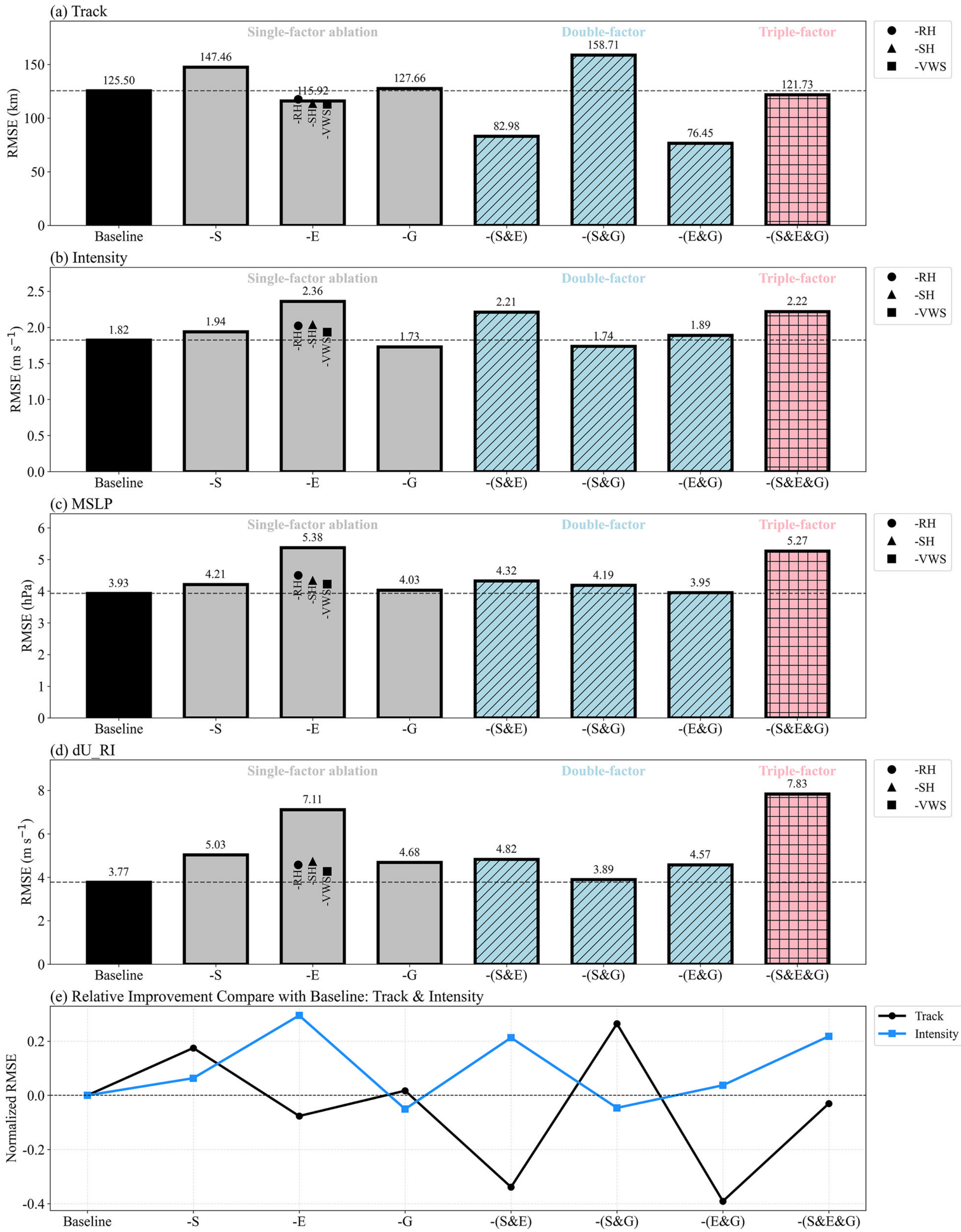

Fig. 3 | Ablation study of the OWZP-Transformer by different predictor categories. **a** RMSE of 6 h track (Haversine distance, km); **b** RMSE of 6 h intensity (m s^{-1}); **c** RMSE of 6 hour MSLP (hPa); **d** RMSE of 24 h future intensity change (m s^{-1}); **e** RMSE of normalized 6 h's Track and Intensity. Colors indicate the removal of different predictor categories, and the three markers denote the effect of eliminating each individual environmental variable.

RMSE line plot, showing the baseline-relative improvement, which is calculated as (value - baseline)/baseline. A positive value indicates that the ablated factor has a positive effect on the model's prediction (i.e., leads to improvement), while a negative value indicates a negative effect (i.e., leads to deterioration in model performance). This representation provides a clearer and more intuitive comparison of the influence of different predictor categories on track and intensity predictions.

The impact of each predictor category on track prediction is clearly illustrated in Fig. 3a. Removing the structural (–S) variables significantly increased the RMSE, indicating a notable deterioration in track forecasting compared to the baseline. This result underscores the positive contribution of the OWZP structural variables in accurately predicting TC tracks. Likewise, omitting only the gradient (–G) category slightly degraded the model's predictive performance. However, simultaneous removal of both structural and gradient categories (–S, –G) resulted in the largest increase in RMSE, highlighting the combined importance and positive influence of these two categories on track forecasting. In contrast, the removal of environmental variables (–E) resulted in improved model performance, as evidenced by a considerable decrease in RMSE compared to the baseline. This improvement was consistent even when removing individual environmental variables (–RH, –SH, or –VWS), suggesting that environmental variables might introduce unnecessary complexity or noise, thereby negatively impacting track predictions. Moreover, further improvements were observed when environmental variables were jointly omitted with either structural or gradient categories (–S&E or –E&G), reinforcing the conclusion that environmental predictors generally contribute negatively to track prediction accuracy.

From Fig. 3b, the removal of either all environmental variables (–E) or even individual environmental variables (–RH, –SH, or –VWS) leads to an evident increase in the RMSE compared to the baseline, clearly highlighting the importance of environmental predictors in intensity forecasting. Similarly, omitting structural variables (–S) also slightly increases the RMSE, demonstrating the positive contribution of the OWZP structural variables toward predicting TC intensity. However, interestingly, the removal of the gradient variables alone (–G) or in combination with structural variables (–S&G) slightly improves the predictive performance, indicating that the gradient variables may negatively influence intensity predictions by introducing unnecessary complexity or noise. Moreover, the removal of any combination involving environmental variables—whether structural and environmental (–S&E), environmental and gradient (–E&G), or all three categories (–S&E&G)—consistently results in poorer model performance. This further emphasizes the crucial role of environmental variables in accurately predicting TC intensity.

From Fig. 3c, d, it is apparent that the removal of any predictor category results in increased RMSE values for predictions of both MSLP and 24 h future intensity change (dU_RI) relative to the baseline. This consistent deterioration in model performance underscores the importance of retaining all four predictor categories (structural, environmental, gradient, and basic variables) for accurate intensity-related forecasts.

Figure 3e provides a clearer comparative visualization of the relative impacts on track and intensity predictions after removing various predictor categories, using normalized RMSE values to eliminate scale differences between these two metrics. Specifically, structural (OWZP) and gradient variables have a beneficial influence on track prediction, while structural and environmental variables are crucial for accurate intensity prediction. Therefore, the final selection of these 15 predictors represents a careful balance designed to simultaneously optimize the forecasts of both track and intensity. Focusing exclusively on enhancing the accuracy of one aspect (e.g., track prediction) would likely compromise performance in intensity forecasting, and vice versa. Consequently, our choice reflects a balanced strategy aiming to achieve overall robust and reliable predictions across multiple forecasting targets.

Interpretation of variables' contribution

We further assess the contribution of 15 parameters by using DeepLIFT and DeepLiftShap to avoid opacity and black-box nature and to enhance the interpretability of DL models. Among various XAI algorithms, the DeepLIFT[32] and DeepLiftShap[33] methods align most closely with domain expertise, offering higher qualitative interpretability and more meaningful insights in line with expert knowledge[37].

Figures S7 and 4 display the results of the feature importance analysis using DeepLIFT and DeepLiftShap, respectively, to explain the contribution of each input variable X to each predictand y. DeepLIFT employs a zero baseline, outputting the mean absolute contribution score, while DeepLiftShap uses three different baselines to approximate the mean absolute Shapley value. A comparison of the two methods in Figs. S7 and 4 reveals that DeepLiftShap's output rankings in feature importance are nearly identical or even same to those from DeepLIFT. This similarity arises from the fact that DeepLiftShap is fundamentally an extension of DeepLIFT, with the addition of multiple baselines to ensure more accurate and consistent estimation of feature importance.

When evaluating the importance of various input variables on the prediction targets y, it is evident that a predict and y's past values have the largest contribution to the prediction of its future values. This is expected, as our deep learning model is performing short-term forecasts, where the predictive power relies heavily on the autocorrelation of the variables themselves. Beyond the self-contributions, the basic factors—representing fundamental TC characteristics—also exhibit significant contributions to the prediction results, further validating their role in forecasting TC behavior.

Apart from the basic variables and self-contributions, environmental and structural factors, such as the specific humidity at 925 hPa, relative humidity at 925 hPa and OWZ parameter at 500 hPa, play a substantial role in forecasting intensity and intensity change over the next 24 h (Fig. 4). This observation reinforces the effectiveness of the OWZ parameter in improving prediction accuracy. Furthermore, gradient factors, including steering flow, moving angle and translation speed, also contribute meaningfully to track predictions.

It is important to note that the improvement in predictive performance from each input variable should not be interpreted solely through their contribution score or Shapley value. The model's final output is the result of complex interactions among multiple variables. Figures S7 and 4 provide insights into the relative importance of the variables, and they align well with our domain expertise and intuition. Moreover, they directly demonstrate the significant role of the OWZ parameter in enhancing the prediction of TC intensity and intensity changes.

Discussion

In this study, we developed a model named OWZP-Transformer, capable of simultaneously predicting track, intensity, and 24 h future intensity changes of TCs. Additionally, we categorized the input variables into four distinct groups: basic factors, gradient factors, environmental factors, and structural factors. Importantly, for the first time, we incorporated a structural factor—the Okubo-Weiss-Zeta (OWZ) parameter—into the model, enabling a more accurate representation of the processes involved in TC formation and development. Comparisons with existing DL models revealed that the inclusion of the OWZ parameter significantly enhanced the model's predictive performance. Specifically, our model demonstrated superior performance compared to other DL models by achieving 25.1–37.8% improvement in predicting intensity over the next 6 h and 37.4–54.8% improvement in predicting future 24 h intensity change.

Moreover, our study provided insights into the model's predictive mechanisms through an ablation study and two explainable AI techniques, DeepLIFT and DeepLiftShap, quantifying the importance of individual input variables. These methods highlighted that structural and environmental variables significantly contribute to TC intensity predictions, whereas structural and gradient variables are critical for accurate track predictions. These findings not only underscore the robustness of the OWZP-Transformer model but also demonstrate the practical utility of employing explainable AI techniques in meteorological research to elucidate complex predictive mechanisms.

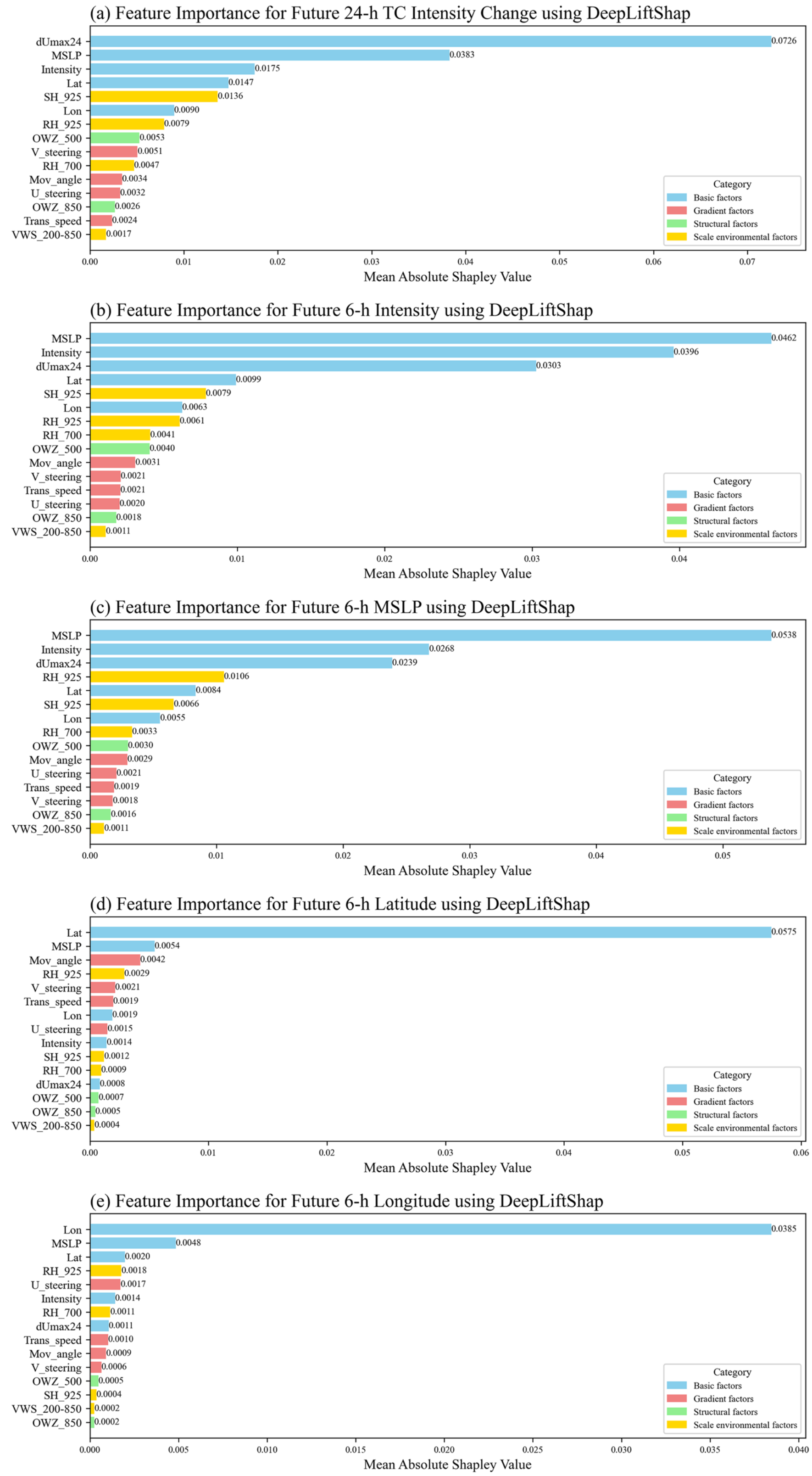

Fig. 4 | Feature importance test using DeepLiftShap. Contribution of each variable in predicting **a** 24 h intensity change, **b** 6 h intensity, **c** 6 h MSLP, **d** 6 h latitude, and **e** 6 h longitude, listed in order of highest contribution. Colors represent four categories.

Focusing on 6 h track and intensity prediction and 24 h rapid intensification, this study establishes the short-step accuracy needed for stable multi-step extensions and for capturing fast-evolving conditions. High-quality 6 h forecasts are critical for autoregressive systems as they underpin longer-term predictions by ensuring stability and sensitivity to rapid change. Our work establishes this foundational accuracy, which is essential for scalable extensions. We introduce the OWZ parameter as a physics-aware predictor and demonstrate its added value through ablation study and feature importance analysis. The OWZP-Transformer improves both track and intensity skill, addressing a key gap in structural representation for short-term TC forecasting. Together, these contributions justify the present short-lead scope and provide a sound basis for sequential extensions to multi-day forecasts.

In conclusion, the OWZP-Transformer model presents a significant advancement in the field of TC forecasting. By effectively combining multiple prediction tasks and incorporating interpretability into the modeling process, it provides not only superior accuracy but also enhanced understanding of critical factors influencing TC behavior. The continued refinement and expansion of this approach hold great potential for substantially improving both the scientific understanding and practical management of tropical cyclone-related hazards.

Methods

IBTrACS data

The International Best Track Archive for Climate Stewardship (IBTrACS)[38] is an extensive compilation of rigorously verified global TC best-track data gathered from multiple agencies worldwide. We use TC latitude and longitude (°), maximum 2-min average sustained wind speed (intensity, kt) and minimum sea level pressure (MSLP, hPa) from the 6 hourly (00, 06, 12, 18 UTC) CMA records in the WNP region from 1980 to 2023. A TC event is defined as starting when its intensity reaches tropical storm level (intensity ≥ 17 m s^{-1}). TCs with fewer than 13 data points were omitted from the analysis, a threshold chosen to ensure that each retained TC contains enough sequential information to serve as valid input to the model. The intensity change is determined by calculating the difference between the current intensity and the intensity 24 h in the past. The Haversine distance is adopted to estimate the translation speed. Specifically, the great-circle distance between two consecutive best-track positions with coordinates (φ_1, λ_1) and (φ_2, λ_2) is computed using the haversine formula[39] in Eq. (1), where λ denotes longitude and φ denotes latitude, $R = 6371\text{km}$, $\Delta\varphi = \varphi_2 - \varphi_1$ and $\Delta\lambda = \lambda_2 - \lambda_1$.

$$d = 2R\arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\varphi}{2}\right) + \cos\varphi_1\cos\varphi_2\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right) \tag{1}$$

Equation (1) provides the along-surface distance d in kilometers and is free from planar-projection errors. The translation speed is subsequently derived by dividing the great-circle distance by the 6 h temporal resolution:

$$speed = \frac{d}{6} \tag{2}$$

We also compute TCs' moving angle θ, as illustrated in Eq. (3), which is defined as the great-circle bearing between successive 6 h fixes and is obtained with the two-argument atan2 function and normalized to the interval $[0, 2\pi)$ rad.

$$\theta = \left(atan2\left(\sin\Delta\lambda\cos\varphi_2, \cos\varphi_1\sin\varphi_2 - \sin\varphi_1\cos\varphi_2\cos\Delta\lambda\right) + 2\pi\right)\text{mod}2\pi \tag{3}$$

where mod denotes the modulo operation ensuring $\theta \in [0, 2\pi)$.

The predictors of latitude, longitude, current intensity, MSLP, the intensity change over the previous 24 h (dUmax24) are classified into basic factors that belong to TC basic features. The predictors of moving angle and translation speed are classified into gradient factors that describe the TC change over time.

ERA5 reanalysis data

ERA5 represents the fifth iteration of global climate reanalysis data, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF)[40]. Launched in 2018, ERA5 has significantly enhanced its temporal and spatial resolution, moving from 6-hourly to hourly intervals, and improving its horizontal resolution from 79 km to 31 km $(0.25^\circ \times 0.25^\circ)$[40]. In this study, the 6 hourly relative humidity (RH, %) at 925 hPa and 700 hPa, specific humidity (SH, g kg^{-1}) at 925 hPa, zonal wind (u, m s^{-1}) and meridional wind (v, m s^{-1}) from 1980 to 2023 are used. The VWS is calculated using the zonal and meridional wind components at 200 hPa and 850 hPa, as defined by following Eq. (4):

$$\text{VWS} = \sqrt{\left(u_{200} - u_{850}\right)^2 + \left(v_{200} - v_{850}\right)^2} \tag{4}$$

The Okubo–Weiss–Zeta (OWZ) parameter[31] at 500 hPa and 850 hPa represent low-deformation vorticity[41] and is also calculated using the zonal and meridional wind components as shown in Eq. (5):

$$\text{OWZ} = \text{sgn}(f) \times (\zeta + f) \times \max\left[\frac{\zeta^2 - \left(E^2 + F^2\right)}{\zeta^2}, 0\right] \tag{5}$$

where $f = 2\Omega\sin\varphi$ is the Coriolis parameter, Ω is the Earth's angular velocity and φ is latitude; $\zeta = \partial v/\partial x - \partial u/\partial y$ is the vertical component of relative vorticity, $E = \partial u/\partial x + \partial v/\partial y$ is the stretching deformation, $F = \partial v/\partial x + \partial u/\partial y$ is the shearing deformation.

The steering flow[42,43] is evaluated as the depth-mean environmental wind within the 200–850 hPa layer, spatially averaged over a 270 km circle centered on the cyclone. For each horizontal grid point x, the zonal (u) and meridional (v) wind components are first combined into pressure-thickness-weighted layer means:

$$\bar{u}(x,t) = \frac{\sum_{i=1}^{4}\Delta p_i\, u\left(p_i, x, t\right)}{\sum_{i=1}^{4}\Delta p_i}, \bar{v}(x,t) = \frac{\sum_{i=1}^{4}\Delta p_i\, v\left(p_i, x, t\right)}{\sum_{i=1}^{4}\Delta p_i} \tag{6}$$

where $p_i = \{200, 500, 700, 850\}$ hPa and Δp_i denotes the midpoint pressure thickness of each layer. These layer-mean components are then horizontally averaged over all grid points whose great-circle distance from the TC center (φ_c, λ_c) is ≤270 km[44], giving:

$$u_s(t) = \frac{1}{M}\sum_{d(x,x_c)\le R}\bar{u}(x,t),\ v_s(t) = \frac{1}{M}\sum_{d(x,x_c)\le R}\bar{v}(x,t) \tag{7}$$

with $R = 270km$ and M is the number of grid points inside the circle. The resulting pair (u_s, v_s) furnishes a single steering-flow vector for each6 h TC position.

The predictors of VWS, RH and SH are classified into environmental factors, the OWZ parameter is classified into structural factor, and the zonal and meridional steering flow components are classified into gradient factors.

Input data preparation

We prepare 15 variables as a predictor of the models (Table 1). The predictors, including latitude, longitude, intensity, MSLP, translation speed, moving angle, dUmax24 obtained from IBTrACS data, as well as OWZP at 850 hPa and 500 hPa, VWS between 200 hPa and 850 hPa, RH at 925 hPa and 700 hPa, SH at 925 hPa, zonal and meridional steering flow obtained from ERA5 data. For each TC point, the latitude, longitude, moving angle, translation speed, intensity, MSLP, dUmax24 values obtained from IBTrACS and steering flow calculated from ERA5 data are inherently 0-dimensional data (scalars). However, variables such as OWZP, VWS, RH, and SH at specified levels are originally 2-dimensional data. To efficiently integrate these 2-dimensional data into our model, we convert them into 0-dimensional data by extracting either the maximum or the average values within a defined region. After conducting sensitivity tests on the validation dataset, we ultimately selected a 10°×10° region as the final averaging area. Specifically, for each center point of a TC recorded in the IBTrACS dataset, we identify the corresponding time and the nearest latitude and longitude coordinates in the ERA5 dataset. Then, for each TC point, we delineate a 10°×10° region. Within this region, we extract the maximum value for the OWZ parameter and compute the average values for VWS, RH, and SH to represent the corresponding factor for that TC point. Since positive OWZP values indicate regions where vorticity dominates over deformation, the maximum OWZP value within the inner-core reflects the most dynamically coherent and intense cyclonic structure, making it an appropriate proxy for assessing TC core strength. By doing so, we transform each TC factors within a certain region into a corresponding set of 0-dimensional factors.

Those scalar predictors are employed as input variables (X) for the DL model. The output variables (y), i.e., predictands, for the DL model consist of latitude, longitude, intensity, and MSLP in the next 6 h, as well as 24 h future intensity change (dU_RI). For all X and y variables, we apply a min-max scaling process to standardize their values within a range of 0 to 1 or −1 to 1. This normalization helps reduce the impact of varying measurement scales across the data.

We designate the last four years of data (2020 to 2023) as the test set to evaluate the model's prediction performance on unseen data. The data from 1980 to 2017 serves as the training set, while the data from 2018 and 2019 is used as the validation set. The validation set is instrumental in selecting the optimal hyperparameter combination and preventing model overfitting. The final data dimensions for the training, validation, and test sets are summarized in Table S1. After dividing the dataset into training, validation, and test sets, we employ a sliding window approach—where the inputs for each window are always based on the true observed data rather than the model's own forecasts— to generate the TC track and intensity for the next 6 h, as well as the intensity change over the next 24 h (dU_RI), using the past 24 h of data. Given that our data has a 6 h interval, the past 24 h window contains 5 TC points. For the X variables, the first dimension corresponds to the sample size, the second dimension corresponds to the past 24 h of data, and the third dimension includes all X variable features. Similarly, for the y variables, the first dimension corresponds to the sample size, the second dimension corresponds to the next time step (6 h forecast for track and intensity, and 24 h forecast for intensity change), and the third dimension includes all y variable targets.

OWZP-Transformer model

The Transformer model was initially proposed to address natural language processing (NLP) tasks[45]. In recent years, the Transformer architecture has gained significant attention beyond its initial success in NLP field, leading to its application in time series forecasting by numerous researchers[46]. In this study, we explore the effectiveness of the Transformer model specifically in predicting TC development and its characteristics. Given the inclusion of the OWZ parameters, we refer to our model as the OWZP-Transformer. The model's general design is depicted in Figure S1. The Transformer architecture consists of an encoder and a decoder component. The encoder is designed to derive rich features from the time series data using the multi-head self-attention (MHA) mechanism[45], while the decoder is tasked with reconstructing these extracted features.

The encoder is tasked with representing the past dataset and is made up of multiple elements. The input embedding for the TC data is implemented as a simple linear layer that transforms the input with dimensions [Batch size, Number of samples = 5, Predictors = 15] into the model's embedding space. To account for the sequential nature of the data, positional encoding is applied along the time dimension. This process assigns a unique positional value to each time step t in the input sequence. The specific formula for positional encoding is outlined as follows:

$$\begin{cases} P_t^{(2i)} = \sin\left(w_i t\right), 2i \le d \\ P_t^{(2i+1)} = \cos\left(w_i t\right), 2i + 1 \le d \end{cases} \tag{8}$$

where $w_i = \frac{1}{10000^{2i/d}}$, t is the index of the time step, and d is the dimension of the embedding.

The configuration of MHA mechanism is depicted on the right side of Figure S2. MHA is widely regarded as the leading method for extracting global context in tasks that involve long-term time series forecasting[47]. Within MHA, the input vector sequence X is transformed into h distinct query (Q_h), key (K_h), and value (V_h) matrices, as illustrated in Eq. (9).

$$\begin{cases} Q_h = XW_h^Q \\ K_h = XW_h^K \\ V_h = XW_h^V \end{cases} \tag{9}$$

Here W_h^Q, W_h^K and W_h^V represent the learnable parameters within the linear layers, and h refers to the count of heads in the multi-head configuration. This allows the model to establish multiple sub-spaces, enabling it to focus on various facets of the data to capture complementary features[48]. Following the linear transformations, a sequence of output vectors is calculated using scaled dot-product attention, as shown on the left side of Figure S2 and in Eq. (10).

$$(O_h = \text{Attention}\left(Q_h, K_h, V_h\right) = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_K}}\right) V_h \tag{10}$$

In this calculation, O_h represents the scaled dot-product attention output, and $\sqrt{d_K}$ is the scaling factor applied to the weight matrix. The final step involves concatenating these O_h outputs and applying a linear projection:

$$\mathrm{MultiHeadSelfAttention(Q, K, V) = Concatenate\left(O_0, O_1, \ldots, O_h\right) W^O} \tag{11}$$

where W^O is the learnable parameter of the MHA mechanism.

The add & norm operation integrates a residual connection, which is then followed by layer normalization to stabilize the learning process. The feedforward network consists of a two-layer linear architecture that employs the GELU[49] activation function to introduce non-linearity.

In the decoder, the architecture is largely similar but includes an additional feature known as masked MHA. This mechanism is crucial for ensuring that future data points are not visible to the model during the prediction process, thereby enforcing a sequential approach to generating outputs.

Building on this, our implementation employs a dual-encoder, single-decoder configuration tailored for multivariate, sequence-to-sequence regression of TC variables. Each input sample consists of a sliding window of five consecutive snapshots, represented as a tensor of dimension [batch size, 5-time steps, 15 predictors]. A linear input embedding layer first projects the 15 continuous predictors into the latent space and adds fixed sinusoidal positional encodings. The embedded sequence is processed by two encoder blocks arranged in series and each block contains one Transformer encoder layer. Their stacked memories are then supplied to one decoder block composed of two Transformer decoder layers. Within the decoder, the first self-attention sub-layer is masked to prevent access to positions beyond the current prediction step, while the subsequent cross-attention sub-layer conditions on the encoder memories to fuse historical information. Finally, the decoder's last time step is passed through a linear projection to produce a five-element output vector and returned as a tensor of shape [batch size, 1, 5]. No softmax activation is applied because all targets are real-valued. This dual-encoder, single-decoder design differs from previous Transformer-based TC prediction models[5,34], which typically utilize a single encoder structure, thereby enabling our model to capture more comprehensive temporal dependencies and enhance multi-task learning performance.

The network is trained end-to-end with a composite loss function designed to simultaneously optimize track and intensity predictions. Specifically, the composite loss function L combines the Haversine distance for track prediction errors (latitude and longitude) and the root-mean-square error (RMSE) for intensity-related terms, defined mathematically as below:

$$L = \alpha \cdot \text{Haversine}\left(lat_{\text{pred}}, lon_{\text{pred}}, lat_{\text{true}}, lon_{\text{true}}\right) + \beta \cdot \sqrt{\frac{1}{3N}\sum_{i=1}^{N}\sum_{j=1}^{3}\left(y_{ij}^{\text{pred}} - y_{ij}^{\text{true}}\right)^2 + \epsilon} \tag{12}$$

Here, α and β are weighting coefficients, both set to 1 in our experiments; N represents the batch size, the intensity-related terms (y_{ij}) include the predicted and true values of intensity, MSLP and 24 h future intensity change, and ϵ is a small constant (e.g., 10^{-6}) added for numerical stability to prevent undefined gradients during backpropagation when the intensity error is zero. The Haversine loss component is computed using the standard spherical distance formula in Eq. (1). Such a composite loss formulation represents a methodological advancement over previous studies[5,34], where single-task or standard loss functions were commonly used, allowing our model to jointly optimize physically distinct prediction targets in a unified framework.

Forecast evaluation

In this research, the performance of the models is assessed using four key metrics: root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2) and mean signed error (bias) as illustrated in Eqs. (13) to (16).

$$RMSE = \sqrt{\frac{1}{N}\sum\left(Y - \hat{Y}\right)^2} \tag{13}$$

$$MAE = \frac{1}{N}\sum|Y - \hat{Y}| \tag{14}$$

$$R^2 = 1 - \frac{\sum\left(Y - \hat{Y}\right)^2}{\sum\left(Y - \bar{Y}\right)^2} \tag{15}$$

$$\text{Bias} = \frac{1}{N}\sum\left(\hat{Y} - Y\right) \tag{16}$$

Here N represents the overall count of samples; Y denotes to the real values; $\hat{Y}$ refers to the forecasting values; $\bar{Y}$ indicates the mean of real values. MAE provides a straightforward measure of the average magnitude of errors in the predictions, with lower values indicating better model performance. RMSE gives more weight to larger errors and is thus sensitive to outliers, where smaller RMSE values indicate higher predictive accuracy. R^2 value closer to 1 signifies that the model explains a large portion of the variance, indicating strong predictive power. Bias quantifies systematic over- or under-estimation: a positive value indicates a consistent over-prediction, a negative value indicates under-prediction, and an ideal model yields a bias close to zero.

In addition, the Heidke Skill Score (HSS), Probability of Detection (POD) and False Alarm Ratio (FAR) were employed to evaluate the performance of predictions for RI events. The HSS is a widely used metric for assessing the performance of models dealing with class imbalance. It is calculated using accuracy (ACC) and the standard forecast (SF), as defined by Heidke[50]:

$$\text{HSS} = \frac{\text{ACC} - \text{SF}}{1 - \text{SF}} \tag{17}$$

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{N} \tag{18}$$

$$\text{SF} = \frac{(\text{TP} + \text{FP})}{N} \times \frac{(\text{TP} + \text{FN})}{N} \times \frac{(\text{TN} + \text{FN})}{N} \times \frac{(\text{TN} + \text{FP})}{N} \tag{19}$$

In these formulas, TP (true positives), TN (true negatives), FP (false positives), and FN (false negatives) represent the respective counts for each outcome, and N denotes the total number of instances. An HSS of 1 indicates perfect predictions, while a negative value signifies no predictive skill. A higher POD and lower FAR indicate better discrimination between RI and non-RI cases[51], as shown in Eqs. (20) to (21).

$$\text{POD} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{20}$$

$$\text{FAR} = \frac{\text{FP}}{\text{TP} + \text{FP}} \tag{21}$$

DL model setting

The OWZP-Transformer model was implemented using PyTorch 2.3.1 and trained on the High-Performance Computing (HPC) cluster with a single NVIDIA Tesla-V100 GPU. GELU[49] was selected as the activation function, and the AdamW[52] optimization algorithm was applied for gradient descent. To identify the optimal model architecture, Bayesian optimization[53] was employed through the Optuna Python package[54], exploring the best combination of embedding dimension, number of attention heads, feedforward network dimension, number of encoder layers and decoder layers. Bayesian optimization was also utilized to fine-tune the learning rate, dropout rate, and batch size.

Given the model's rapid convergence on the training data, the number of training epochs was set to 200. Early stopping[55] was implemented, terminating the training process if the validation loss fails to improve over 20 consecutive epochs. Additionally, a learning rate adjustment strategy[56] was incorporated during training. Bayesian optimization was performed over 10 trials, which took approximately 1 h to complete. The final Transformer model configuration was as follows: batch size of 128, learning rate of 0.001, dropout rate of 0.1, embedding dimension of 128, 16 attention heads, a feedforward network dimension of 128, 1 encoder layer, and 2 decoder layers. Training the OWZP-Transformer on the Tesla-V100 GPU took only a few minutes per run. The model was subsequently employed to predict and get the outputs which were then inverse transformed to their actual units, and the model's performance was evaluated using different evaluation metrics.

Feature importance explanation

To address the opacity and black-box nature of DL models, and to assess the contribution of 15 parameters that we defined before to improving short-term forecast accuracy, XAI[57] techniques are employed to enhance the interpretability of DL models, thereby improving trust and model validity. These methods play a crucial role in identifying and highlighting key contributors to the output predictions of complex DL models[58]. Among various XAI algorithms, the Deep Learning Important Features (DeepLIFT)[32] and DeepLIFT + Shapley values (DeepLiftShap)[33] methods align most closely with domain expertise, offering higher qualitative interpretability and more meaningful insights in line with expert knowledge[37]. These two XAI algorithms were selected to assess the different contributions of variables in the prediction outcomes.

DeepLIFT determines the contribution of each input feature by evaluating the difference in the model's output between a reference (baseline) input and the actual input[32]. It utilizes backpropagation to compute these contribution scores. Here, we choose zero baseline for DeepLIFT. For contribution scores, a positive value indicates that the input variable X is positively correlated with the predictand y, while a negative value signifies a negative correlation. To facilitate a more straightforward comparison of variable importance, we use the absolute contribution score, where larger values indicate a greater contribution of the variable to the prediction outcome.

Suppose we have a neural network prediction model with an input layer consisting of neurons $\{x_1, x_2, x_3, \ldots, x_n\}$, several hidden layers with neurons $\{h_1, h_2, h_3, \ldots, h_n\}$, and a target output o. Let $f(x)$ represent the activation of a given neuron, while $f(x')$ represents the reference activation. DeepLIFT computes the contribution scores as shown in Eq. (22):

$$\Delta o = f(x) - f(x') = \sum_{i=1}^{n} C_{\Delta x_i \Delta o} \tag{22}$$

where $C_{\Delta x_i \Delta o}$ represents the contribution score for each input variable x_i, $\Delta x_i = x_i - x_i'$ is the difference between the actual input and the reference input and $\Delta o = f(x) - f(x')$ is the difference in output. Equation (22) is known as the 'summation to delta' property[32], indicating that the sum of all input changes equals the overall output difference. To calculate the contribution score $C_{\Delta x_i \Delta o}$ for each input variable, a multiplier $m_{\Delta x_i \Delta o}$ is defined by averaging the differences, as shown in Eq. (23).

$$m_{\Delta x_i \Delta o} = \frac{C_{\Delta x_i \Delta o}}{\Delta x_i} \tag{23}$$

Since deep neural networks contain input, hidden, and output layers, we can define the contribution multiplier in Eq. (24) by applying iterative chain rules across each layer.

$$m_{\Delta x_i \Delta o} = \sum\nolimits_j m_{\Delta x_i \Delta h_j} \cdot m_{\Delta h_j \Delta o} \tag{24}$$

Using this chain rule, the contribution in terms of multipliers can be computed for a given target through backpropagation, as shown in Figure S3. Specifically, by combining Eqs. (23) and (24), we can use Eq. (25) to calculate the contribution score of each variable.

$$C_{\Delta x_i \Delta o} = \left(\sum\nolimits_j m_{\Delta x_i \Delta h_j} \cdot m_{\Delta h_j \Delta o}\right) * \Delta x_i \tag{25}$$

DeepLiftShap extends the DeepLIFT algorithm to approximate Shapley values[33], a concept from game theory that represents the average marginal contribution of each feature across all possible feature combinations[59], and is widely used to explain variable contributions in predictive models. For each input variable, the method calculates the DeepLIFT attributions relative to different baselines and then computes the average of the resulting attributions[60]. This averaging process allows for a more accurate and consistent estimation of feature importance across varying contexts, making it a valuable tool for interpreting complex DL models. Here, we choose 3 different baselines: zero input, mean of input, randomly sampled input and use mean absolute Shapley values to focus on the magnitude of each variable's contribution, where higher values indicate a stronger influence on the prediction outcome.

DeepLiftShap recursively computes feature contributions using the chain rule, backpropagating from the output layer to the input layer. This process provides an efficient approximation of Shapley values. DeepLiftShap breaks down the contribution of each feature into multipliers through each layer using Eq. (24). The final Shapley values $\phi_i(f(x))$ are computed by recursively passing these multipliers and combining them with the difference between the input x and the baseline:

$$\phi_i(f(x)) = m_{\Delta x_i \Delta o} \cdot (x_i - E(x_i)) \tag{26}$$

When using multiple baselines, DeepLiftShap averages the results from all baselines to ensure that the computed attribution values adhere to the properties of Shapley values (Eq. 27).

$$\phi_i(f(x)) = \frac{1}{k} \sum_{j=1}^{k} m_{\Delta x_i \Delta o_j} \cdot \left(x_i - r_j\right) \tag{27}$$

Here, r_j represents the j-th baseline, which is a reference input. $m_{\Delta x_i \Delta o_j}$ refers to the multiplier corresponding to the j-th baseline, representing the contribution propagated from the input layer to the output layer using the chain rule. $x_i - r_j$ is the difference between the input x_i and the j-th baseline r_j.

Data availability

The best track data of the IBTrACS are available at https://www.ncei.noaa.gov/products/international-best-track-archive. The ERA5 reanalysis dataset can be downloaded at https://cds.climate.copernicus.eu/datasets/reanalysis-era5-pressure-levels?tab=overview.

Code availability

Our OWZP-Transformer for TC track and intensity prediction is available on GitHub at https://github.com/William-Linn/OWZP-Transformer.

Received: 11 March 2025; Accepted: 21 August 2025;
Published online: 03 November 2025

References

1. Dare, R. A. & McBride, J. L. Sea surface temperature response to tropical cyclones. *Mon. Weather Rev.* **139**, 3798–3808 (2011).
2. Fang, G. et al. Toward a refined estimation of typhoon wind hazards: Parametric modeling and upstream terrain effects. *J. Wind Eng. Ind. Aerodyn.* **209**, 104460 (2021).
3. Sampson, C. R., Jeffries, R. A. & Neumann, C. J. *Tropical cyclone forecasters reference guide 4. Tropical Cyclone Motion*. https://doi.org/10.21236/ada302329 (1995).
4. Keenan, T. Statistical forecasting of tropical cyclone movement in the Australian region. *Q. J. R. Meteorol. Soc.* **111**, 603–615 (1985).
5. Jiang, W. et al. Transformer-based tropical cyclone track and intensity forecasting. *J. Wind Eng. Ind. Aerodyn.* **238**, 105440 (2023).
6. Elsberry, R. L. Recent advancements in dynamical tropical cyclone track predictions. *Meteorol. Atmos. Phys.* **56**, 81–99 (1995).
7. Peng, X., Fei, J., Huang, X. & Cheng, X. Evaluation and error analysis of official forecasts of tropical cyclones during 2005–14 over the Western North Pacific. Part I: storm tracks. *Weather Forecast* **32**, 689–712 (2017).
8. Fudeyasu, H., Ito, K. & Miyamoto, Y. Characteristics of tropical cyclone rapid intensification Over The Western North Pacific. *J. Clim.* **31**, 8917–8930 (2018).
9. Lee, C.-Y., Tippett, M. K., Sobel, A. H. & Camargo, S. J. Rapid intensification and the bimodal distribution of tropical cyclone intensity. *Nat. Commun.* **7**, 10625–10625 (2016).
10. Roy, C. & Kovordányi, R. Tropical cyclone track forecasting techniques — A review. *Atmos. Res* **104–105**, 40–69 (2012).
11. Xu, J., Wang, Y. & Yang, C. Quantifying the environmental effects on tropical cyclone intensity change using a simple dynamically based dynamical system model. https://doi.org/10.1175/JAS-D-23-0058.1 (2023).
12. Chen, X., Wang, Y., Zhao, K. & Wu, D. A numerical study on rapid intensification of Typhoon Vicente (2012) in the South China Sea. Part I: verification of simulation, storm-scale evolution, and environmental contribution. https://doi.org/10.1175/MWR-D-16-0147.1 (2017).
13. Su, H. et al. Applying satellite observations of tropical cyclone internal structures to rapid intensification forecast with machine learning. *Geophys. Res. Lett.* **47**, e2020GL089102 (2020).
14. Khallaf, R. & Khallaf, M. Classification and analysis of deep learning applications in construction: a systematic literature review. *Autom. Constr.* **129**, 103760 (2021).
15. Bi, K. et al. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538 (2023).
16. Chen, G., Yang, M., Zhang, X. & Wan, R. Verification of tropical cyclone operational forecast in (2019).

17. Tong, B., Sun, X., Fu, J., He, Y. & Chan, P. Identification of tropical cyclones via deep convolutional neural network based on satellite cloud images. *Atmos. Meas. Tech.* **15**, 1829–1848 (2022).
18. Gao, S. et al. A nowcasting model for the prediction of typhoon tracks based on a long short-term memory neural network. *Acta Oceanol. Sin.* **37**, 8–12 (2018).
19. Rüttgers, M., Lee, S., Jeon, S. & You, D. Prediction of a typhoon track using a generative adversarial network and satellite images. *Sci. Rep.* **9**, 6057–6057 (2019).
20. Lian, J., Dong, P., Zhang, Y. & Pan, J. A novel deep learning approach for tropical cyclone track prediction based on auto-encoder and gated recurrent unit networks. *Appl. Sci.* **10**, 3965 (2020).
21. Xu, G. et al. AM-ConvGRU: a spatio-temporal model for typhoon path prediction. *Neural Comput. Appl.* **34**, 5905–5921 (2022).
22. Song, T., Li, Y., Meng, F., Xie, P. & Xu, D. A novel deep learning model by BiGRU with attention mechanism for tropical cyclone track prediction in the Northwest Pacific. *J. Appl. Meteorol. Climatol.* **61**, 3–12 (2022).
23. Pan, B., Xu, X. & Shi, Z. Tropical cyclone intensity prediction based on recurrent neural networks. *Electron. Lett.* **55**, 413–415 (2019).
24. Yuan, S., Wang, C., Mu, B., Zhou, F. & Duan, W. Typhoon intensity forecasting based on LSTM using the Rolling Forecast Method. *Algorithms* **14**, 83 (2021).
25. Tong, B., Wang, X., Fu, J. Y., Chan, P. W. & He, Y. C. Short-term prediction of the intensity and track of tropical cyclone via ConvLSTM model. *J. Wind Eng. Ind. Aerodyn.* **226**, 105026 (2022).
26. Na, Y., Na, B. & Son, S. Near real-time predictions of tropical cyclone trajectory and intensity in the northwestern Pacific Ocean using echo state network. *Clim. Dyn.* **58**, 651–667 (2021).
27. Huang, C., Bai, C., Chan, S. & Zhang, J. MMSTN: a multi-modal spatial-temporal network for tropical cyclone short-term prediction. *Geophys. Res. Lett.* **49**, e2021GL096898 (2022).
28. Yang, Q., Lee, C.-Y. & Tippett, M. K. A long short-term memory model for global rapid intensification prediction. *Weather Forecast* **35**, 1203–1220 (2020).
29. Xu, G., Lin, K., Li, X. & Ye, Y. SAF-Net: A spatio-temporal deep learning method for typhoon intensity prediction. *Pattern Recognit. Lett.* **155**, 121–127 (2022).
30. Kim, J.-H., Ham, Y.-G., Kim, D., Li, T. & Ma, C. Improvement in forecasting short-term tropical cyclone intensity change and their rapid intensification using deep learning. *Artif. Intell. Earth Syst*. **3**, (2024).
31. Tory, K. J., Dare, R. A., Davidson, N. E., McBride, J. L. & Chand, S. S. The importance of low-deformation vorticity in tropical cyclone formation. *Atmos. Chem. Phys.* **13**, 2115–2132 (2013).
32. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning* 3145–3153 (PMLR, 2017).
33. Lundberg, S. M. & Lee, S.-I. A Unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).
34. Gan, S. L., Fu, J. Y., Zhao, G. F., Chan, P. W. & He, Y. C. Short-term prediction of tropical cyclone track and intensity via four mainstream deep learning techniques. *J. Wind Eng. Ind. Aerodyn.* **244**, 105633 (2024).
35. Tian, W. et al. Short-term rolling prediction of tropical cyclone intensity based on multi-task learning with fusion of deviation-angle variance and satellite imagery. *Adv. Atmospheric Sci.* **42**, 111–128 (2025).
36. Li, L., Li, Y. & Tang, Y. On the duration of tropical cyclone rapid intensification. https://doi.org/10.1029/2024GL108578.
37. Kakavandi, F., Han, P., de Reus, R., Larsen, P. G. & Zhang, H. Interpretable fault detection approach with deep neural networks to industrial applications. In *Proc. International Conference on Control, Automation and Diagnosis (ICCAD)* Vol. 32 pp. 1–7 (IEEE, 2023).
38. Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J. & Neumann, C. J. The International Best Track Archive for Climate Stewardship (IBTrACS). *Bull. Am. Meteorol. Soc.* **91**, 363–376 (2010).
39. Sinnott, R. W. Virtues of the Haversine. *Sky Telesc* **68**, 158 (1984).
40. Soci, C. et al. The ERA5 global reanalysis from 1940 to 2022. *Q. J. R. Meteorol. Soc*. https://doi.org/10.1002/qj.4803 (2024).
41. Bell, S. S., Chand, S. S., Tory, K. J. & Turville, C. Statistical assessment of the OWZ tropical cyclone tracking scheme in ERA-interim. *J. Clim*. **31**, (2018).
42. Chan, J. C. L. & Williams, R. T. Analytical and numerical studies of the beta-effect in tropical cyclone motion. Part I: zero mean flow (1987).
43. Fiorino, M. & Elsberry, R. L. Some aspects of vortex structure related to tropical cyclone motion (1989).
44. Wu, L. & Chen, X. Revisiting the steering principal of tropical cyclone motion in a numerical experiment. *Atmos. Chem. Phys.* **16**, 14925–14936 (2016).
45. Vaswani, A. et al. Attention is all you need. *Adv Neural Inf Process Syst.* **30** (2017).
46. Chen, Z., Chen, D., Zhang, X., Yuan, Z. & Cheng, X. Learning graph structures with transformer for multivariate time-series anomaly detection in IoT. *IEEE Internet Things J.* **9**, 9179–9189 (2022).
47. Cao, D. & Zhang, S. AD-autoformer: decomposition transformers with attention distilling for long sequence time-series forecasting. *J. Supercomput.* **80**, 21128–21148 (2024).
48. Li, S. et al. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Adv Neural Inf Process Syst.* **32**, (2019).
49. Hendrycks, D. & Gimpel, K. Gaussian Error Linear Units (GELUs). Preprint at https://doi.org/10.48550/arXiv.1606.08415 (2023).
50. Heidke, P. Berechnung Des Erfolges Und Der Güte Der Windstärkevorhersagen Im Sturmwarnungsdienst. *Geogr. Ann*. (1926).
51. Wang, C., Yang, N. & Li, X. Advancing forecasting capabilities: a contrastive learning model for forecasting tropical cyclone rapid intensification. *Proc. Natl. Acad. Sci.* **122**, e2415501122 (2025).
52. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. Preprint at https://doi.org/10.48550/arXiv.1711.05101 (2019).
53. Frazier, P. I. A tutorial on Bayesian optimization. Preprint at https://doi.org/10.48550/arXiv.1807.02811 (2018).
54. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna. https://doi.org/10.1145/3292500.3330701 (ACM, 2019).
55. Rice, L., Wong, E. & Kolter, Z. Overfitting in adversarially robust deep learning. in *International conference on machine learning* 8093–8104 (PMLR, 2020).
56. Yedida, R., Saha, S. & Prashanth, T. LipschitzLR: Using theoretically computed adaptive learning rates for fast convergence. *Appl. Intell.* **51**, 1460–1478 (2020).
57. Barredo Arrieta, A. et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020).
58. Adadi, A. & Berrada, M. Peeking inside the black-box: a survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **6**, 52138–52160 (2018).
59. Lipovetsky, S. & Conklin, M. Analysis of regression in game theory approach. *Appl. Stoch. Models Bus. Ind.* **17**, 319–330 (2001).
60. Kokhlikyan, N. et al. Captum: a unified and generic model interpretability library for PyTorch. Preprint at https://doi.org/10.48550/arXiv.2009.07896 (2020).

Acknowledgements

This research was supported by the Hong Kong Research Grants Council Early Career Scheme (9048309) and the National Natural Science Foundation of China (9240162).

Author contributions
Z. Lin: Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. J.-E. Chu: Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Visualization, Writing—original draft, review and editing. Y.-G. Ham: Conceptualization, Writing—review and editing.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44387-025-00037-3.

Correspondence and requests for materials should be addressed to Jung-Eun Chu.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.