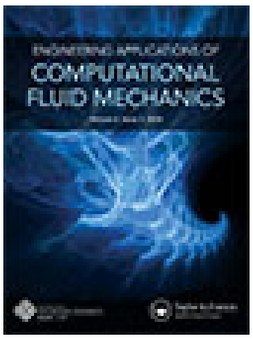

Engineering Applications of Computational Fluid Mechanics

ISSN: 1994-2060 (Print) 1997-003X (Online) Journal homepage: www.tandfonline.com/journals/tcfm20

Investigation of the impact of token embeddings in Transformer-based models on short-term tropical cyclone track and intensity predictions

Yuan-Jiang Zeng, Yi-Qing Ni, Zheng-Wei Chen, Guang-Zhi Zeng, Jia-Yao Wang & Pak-Wai Chan

To cite this article: Yuan-Jiang Zeng, Yi-Qing Ni, Zheng-Wei Chen, Guang-Zhi Zeng, Jia-Yao Wang & Pak-Wai Chan (2025) Investigation of the impact of token embeddings in Transformer-based models on short-term tropical cyclone track and intensity predictions, Engineering Applications of Computational Fluid Mechanics, 19:1, 2538180, DOI: 10.1080/19942060.2025.2538180

To link to this article: https://doi.org/10.1080/19942060.2025.2538180

Published online: 01 Aug 2025.

ENGINEERING APPLICATIONS OF COMPUTATIONAL FLUID MECHANICS
2025, VOL. 19, NO. 1, 2538180
https://doi.org/10.1080/19942060.2025.2538180

Investigation of the impact of token embeddings in Transformer-based models on short-term tropical cyclone track and intensity predictions

Yuan-Jiang Zeng[a,b], Yi-Qing Ni[a,b], Zheng-Wei Chen[a,b], Guang-Zhi Zeng[a,b], Jia-Yao Wang[a] and Pak-Wai Chan[c]

[a]Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong, People's Republic of China; [b]National Rail Transit Electrification and Automation Engineering Technology Research Center (Hong Kong Branch), Hong Kong, People's Republic of China; [c]Hong Kong Observatory, Hong Kong, People's Republic of China

ABSTRACT

Tropical cyclones (TCs) are destructive meteorological phenomena, necessitating accurate predictions of TC track and intensity to reduce risks to human life. This study evaluates three Transformer-based models – vanilla Transformer (Transformer), inverted Transformer (iTransformer), and temporal-variate Transformer (TVFormer) – which are trained, validated, and tested on best track data from 1980 to 2021 from the China Meteorological Administration for TC prediction, integrating temporal, variate, and hybrid token embeddings to analyze temporal and variate correlations. Comparative analysis with four recurrent neural network (RNN) models demonstrates the superiority of the refined Transformer models over RNNs: iTransformer reduces mean absolute error (MAE) and root mean square error (RMSE) by 29.55% and 25.80% (latitude), 50.31% and 46.18% (longitude), 8.71% and 9.98% (pressure), and 8.68% and 9.45% (wind speed), while TVFormer achieves MAE and RMSE reductions of 13.98% and 13.84% (latitude), 39.11% and 38.02% (longitude), 13.69% and 14.02% (pressure), and 12.84% and 12.94% (wind speed) on average. Among Transformer variants, iTransformer excels in track prediction, outperforming Transformer with 21.74% lower MAE and 18.26% lower RMSE for latitude, and 32.73% lower MAE and 24.01% lower RMSE for longitude. TVFormer dominates intensity prediction, reducing pressure errors by 4.42% (MAE) and 3.92% (RMSE) and wind speed errors by 19.21% (MAE) and 14.79% (RMSE) compared to Transformer, while outperforming iTransformer with 4.59% lower MAE and 3.68% lower RMSE for pressure and 3.83% lower MAE and 3.18% lower RMSE for wind speed. Notably, TVFormer also enhances track prediction, with 7.10% reduction in MAE and 7.02% reduction in RMSE for latitude, and 22.84% reduction in MAE and 17.09% reduction in RMSE for longitude compared to Transformer. These results highlight the superiority of iTransformer in track prediction and the efficacy of TVFormer in intensity prediction, thanks to their ability to exploit temporal and variate dependencies, offering potential for TC disaster preparedness systems.

ARTICLE HISTORY
Received 15 February 2025
Accepted 2 July 2025

KEYWORDS
Tropical cyclones; track; intensity; Transformer; token embeddings

1. Introduction

Tropical cyclones (TCs) are among the most intense natural hazards, capable of causing significant devastation to human life, property, and infrastructure. As a result, understanding the track and intensity of TCs is essential as it provides valuable information for various applications. For instance, the design and safety assessment of offshore equipment and infrastructure (Fang et al., 2022; Ju et al., 2021; Zeng et al., 2021), as well as the readiness and survivability assessment of road networks (Hu et al., 2021; Yang et al., 2016) and buildings (Sampson et al., 2012) in coastal cities, rely on accurate TC track and intensity data. On the other hand, the key component of TC early warning systems is the TC prediction results, whose prediction accuracy and inference speed directly affect the effectiveness of early warning systems (Kuleshov et al., 2020; Mandal et al., 2020; Wang et al., 2020). Consequently, rapid and accurate prediction of the track and intensity of TCs is essential to the operational meteorological framework.

There are three main methods for TC prediction: statistical, dynamical, and statistical-dynamical methods (Roy & Kovordányi, 2012). Statistical methods use regression models to establish correlations between

CONTACT Yi-Qing Ni ceyqni@polyu.edu.hk Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, People's Republic of China; National Rail Transit Electrification and Automation Engineering Technology Research Center (Hong Kong Branch), Hung Hom, Kowloon, Hong Kong SAR, People's Republic of China; Zheng-Wei Chen zhengwei.chen@polyu.edu.hk Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, People's Republic of China; National Rail Transit Electrification and Automation Engineering Technology Research Center (Hong Kong Branch), Hung Hom, Kowloon, Hong Kong SAR, People's Republic of China

various indicators from different datasets and the track and intensity of TCs. In contrast, the dynamical methods, also known as numerical predictions and akin to computational fluid dynamics (Chen, Han, et al., 2023; Huo et al., 2023), focus on solving partial differential equations (PDEs) that govern atmospheric circulation. In addition, the statistical-dynamical methods combine statistical and dynamical methods to predict the track and intensity of TCs. Each of these methods has its own merits and demerits. Fast inference speed is the major advantage of the statistical methods, but regression models may be unreliable when predicting TC-related variables with high randomness and variability. The dynamical methods have the potential to be more accurate but are computationally expensive and require extensive expertise in setting their parameters. The advantages and disadvantages of both statistical and dynamical methods are inherent in the statistical-dynamical methods (Tallapragada et al., 2016). Therefore, it is crucial to find a balance between prediction accuracy and computational cost to provide information on TC tracks and intensities for the above assessments. Recent advances in deep learning (DL) have improved many industries, including transportation, biotechnology, finance, etc. (LeCun et al., 2015). DL-based models have demonstrated inference speeds several orders of magnitude faster than dynamical models for weather forecasting and TC predictions (Bi et al., 2023; Chen, Guo, et al., 2023; Chen, Zhong, et al., 2023). In addition, extensive evaluations and experiments have shown that DL-based models outperform traditional statistical and dynamical models in terms of the performance of track predictions (Charlton-Perez et al., 2024; Liu, Hsu, et al., 2024) and intensity predictions (Ma et al., 2023; Meng et al., 2023). Therefore, DL holds promise for fast and accurate TC prediction, which is challenging to achieve with the aforementioned methods (Wang & Li, 2023).

Applying DL to the TC prediction task allows the use of a wide variety of datasets, including best track data, reanalysis data, and satellite data. The best track data in these datasets includes both track and intensity information of each TC, making it a straightforward choice for building DL models. A popular model for TC track and intensity prediction using DL and the best track data is recurrent neural networks (RNNs). The reason is that RNN-based models, such as long short-term memory (LSTM) and gated recurrent unit (GRU), are specifically configured to handle sequence-to-sequence tasks. For example, some studies adopted RNN to forecast TC intensity (Pan et al., 2019), while others used LSTM and GRU to forecast TC track due to their strong ability in representing temporal correlations (Gan et al., 2024; Hao et al., 2024; Lian et al., 2020; Qin et al., 2021). The track and intensity of TCs can also be predicted using improved RNN-based models such as bidirectional gated recurrent unit (BiGRU) and convolutional long short-term memory (ConvLSTM), which have been proposed to enhance the prediction performance of RNN-based models (Song et al., 2022; Tong et al., 2022). In addition to RNN-based models, Transformer is another DL model for processing time series data (Vaswani, 2017). With its powerful parallelisation capability and capacity to represent the whole receptive field of series data, Transformer has found widespread applications beyond its original field of natural language processing, including computer vision, speech recognition, bioinformatics, etc. (Yenduri et al., 2024). Transformer is also suited for TC track and intensity prediction. For instance, the vanilla Transformer was used to predict the track and intensity of TCs occurring in the Northwest Pacific (Gan et al., 2024; Jiang et al., 2023). In contrast to the predictions made using RNN-based models, the use of Transformer in TC track and intensity prediction is still in its early stage, and therefore, more research on the use of Transformer-based models is needed.

The aforementioned works rely on direct applications of the vanilla Transformer (Gan et al., 2024; Jiang et al., 2023); these models employ the self-attention mechanism to depict temporal correlations, which are determined by embedding different variables at the same time step into temporal tokens. Nonetheless, recent studies have also shown the correlation and interaction between the track, intensity and structure, which is important because TCs are the result of chaotic evolutionary dynamics (Chavas et al., 2017; Jiang et al., 2023; Qin et al., 2021; Yenduri et al., 2024). Furthermore, there is usually a significant correlation between the minimum pressure and the mean maximum sustained wind speed, which are two main interchangeable measures of TC intensity (Chavas et al., 2017; Zhao et al., 2024). Thus, these embeddings in vanilla Transformer may downplay the significance of the correlations between track and intensity as well as the correlations between variables indicating TC intensity, both of which are essential in addition to the temporal correlations. In contrast, variate tokens enable Transformer-based models to effectively capture the wind-pressure relationship and track-intensity interaction. Therefore, integrating variate tokens into models has great potential to interpret such interactions and relationships while improving model performance. Consequently, it is critical to investigate the significance of variate correlations in Transformer-based models and their impact on prediction outcomes. Advancements in time series forecasting using Transformer-based models have led to the incorporation of variate tokens or sub-series patches of various

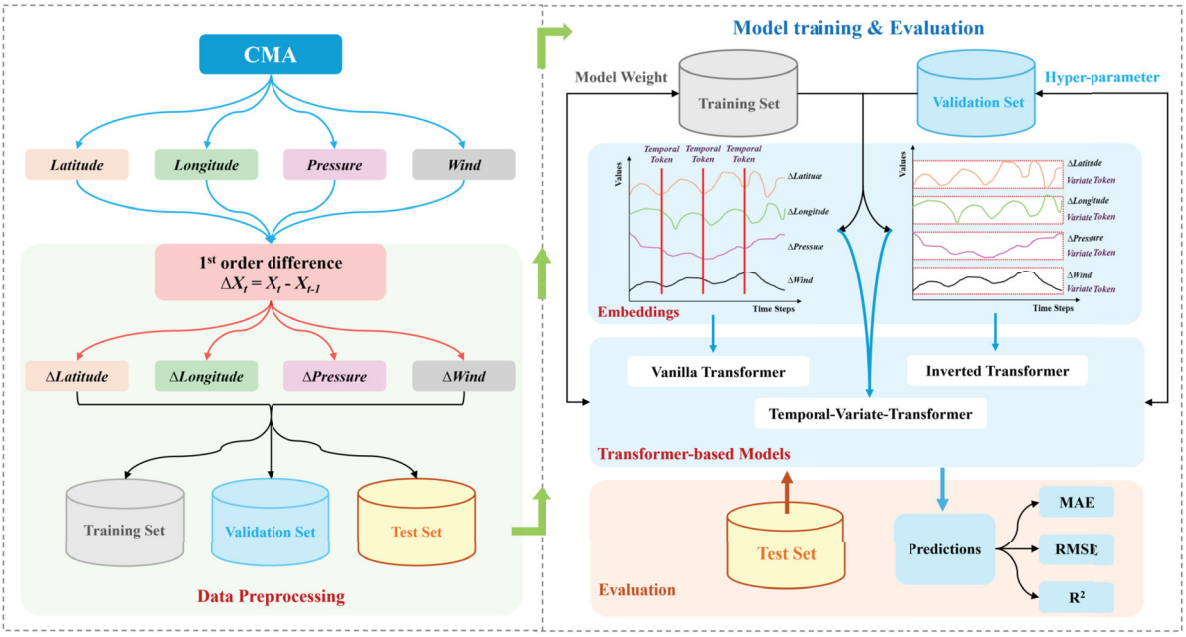

Figure 1. Framework of the proposed study.

variables to account for the significance of variate correlations. Notable examples are the patch time series Transformer (PatchTST) and the inverted Transformer (iTransformer), which both perform well at capturing variate correlations and extending the look-back window of time series data (Liu, Tan, et al., 2024; Nie et al., 2023). However, unlike models that rely solely only on temporal tokens, the utility of Transformer-based models that employ different embeddings to predict TC tracks and intensities remains understudied. In view of this, this paper constructs three Transformer-based models by using the best track data provided by the China Meteorological Administration (CMA) to study the impact of token embeddings on TC track and intensity prediction, given that one-third of TCs occur in the Northwest Pacific (Chan, 2005). This study focuses on exploring the impact of two types of token embeddings, namely temporal tokens and variate tokens, on the prediction of TC tracks and intensities. In addition, a new model that integrates temporal and variate correlations is proposed to evaluate prediction performance.

This study will be presented following the framework shown in Figure 1. Section 2 introduces three Transformer-based models, namely, the vanilla Transformer, the iTransformer, and the temporal-variate Transformer (TVFormer) proposed in this study, as well as the data used. **Section 3** presents the results obtained by the three models, which are then examined and compared in terms of prediction accuracy of TC tracks and intensities. Finally, **Section 4** provides a summary of the study.

2. Methodology

2.1. Data source and processing

2.1.1. Data source

The data used in this study is the best track dataset provided by CMA (tcdata.typhoon.org.cn) (Lu et al., 2021; Ying et al., 2014). With an emphasis on TCs in the Northwest Pacific region, the entire best track dataset covers the years 1949–2023. The dataset includes the track (characterised by the latitude and longitude of the TC centre as shown in Figure 2) and intensity information (defined by the minimum pressure and 2-minute mean maximum sustained wind speed near the TC centre). The international ID, name, and intensity category of TCs are also provided. The dataset primarily uses a 6-hour temporal resolution and encrypts TC records every three hours starting from 2017. This study selects the dataset from 1980 to 2021 because 1980 is widely recognised as the inception of current satellite technology for TC information collection, which helps to improve the accuracy and reliability of the dataset.

2.1.2. Data processing

The dataset is divided into three parts: training set, validation set, and test set, which are used for model training and evaluation. In this study, the longitude, latitude, minimum pressure, and 2-minute mean maximum sustained wind speed near the centre of TCs for the years 1980–2013, 2014–2017, and 2018–2021 have been selected as the training set, validation set, and test set,

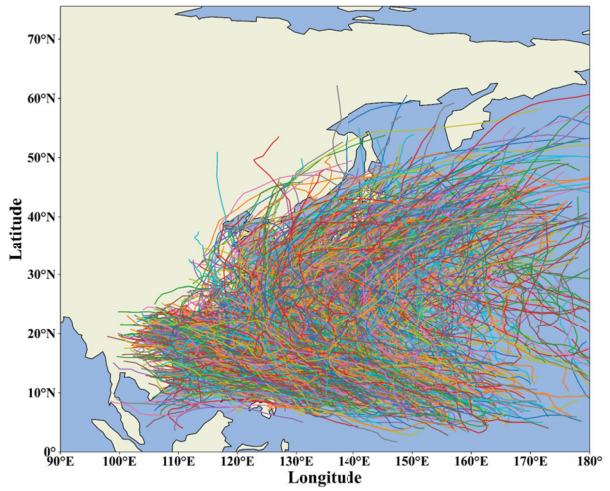

Figure 2. TCs recorded by the best track data from CMA.

Table 1. Summary of the training, validation and test sets.

Dataset	Period	Number of TCs	Number of records	Ratio
Training set	1980–2013	873	26821	79.22%
Validation set	2014–2017	103	3347	9.89%
Test set	2018–2021	103	3689	10.90%
Total	1980–2021	1079	33857	100.00%

respectively. The details are given in Table 1. The ratio of the training set, validation set, and test set is approximately 8:1:1, which is a commonly used data partitioning ratio, and the number of TC records is comparable to that used in TC prediction-related studies (Gan et al., 2024; Jiang et al., 2023; Tong et al., 2022).

TCs labelled 'nameless' are excluded because they were not fully developed or matured, and they are of much shorter duration compared to other TCs. This removal is justified and can contribute positively to maintaining data quality. When multiple 2-minute mean sustained winds are present in the best track data, only the 2-minute mean maximum sustained wind speed near the TC centre is used to ensure data consistency. The 2-minute mean sustained winds associated with coastal severe winds of landfalling TCs and those within a radius of approximately 300–500 km from the TCs are excluded. The temporal resolution is set to 6 hours, which means that only data from 00:00 UTC, 06:00 UTC, 12:00 UTC, and 18:00 UTC are used. Data from 03:00 UTC, 09:00 UTC, 15:00 UTC, and 21:00 UTC since 2017 are not included in the analysis. This study relies on the high-quality data provided by CMA and does not employ any additional data processing techniques to deal with the default values. Our study involves training models for feature evolution, which is expressed as,

$$\Delta x_{t,j} = x_{t,j} - x_{t-1,j} \tag{1}$$

where $x_{t+1,j}$ and $x_{t,j}$ denote the values of feature j at the time steps $t+1$ and t, respectively. In order to reduce the computational cost and improve the convergence of the training process, the following data normalisation is adopted,

$$\Delta x^n_{t,j} = \frac{\Delta x_{t,j} - \Delta x_{j,min}}{\Delta x_{j,max} - \Delta x_{j,min}} \tag{2}$$

where $\Delta x^n_{t,j}$ and $\Delta x_{t,j}$ represent the normalised and original first order difference of feature j at time step t, respectively. By implementing this normalisation, the data is adjusted to fit into the range of [0, 1]. De-normalisation is then performed during the inference stage.

2.2. Problem statement and preliminary

In this study, the short-term prediction of TCs depends on features for both input and output of the model. These features include the track represented by latitude and longitude as well as intensity, which encompasses wind speed and pressure. Consider the time series features $\Delta X = \{\Delta X_1, \Delta X_2, \ldots, \Delta X_t\} \in \mathbb{R}^{T\times C}$, where C denotes the number of features. The features from the preceding h time steps $\Delta X_{t-h+1:t}$ serve as input, while the features from the subsequent p time steps $\Delta X_{t+1:t+p}$ function as output. The task of this study can be generalised as,

$$\Delta X_{t+1:t+p} = f(\Delta X_{t-h+1:t}) \tag{3}$$

where $f(\cdot)$ denotes the model to be trained. TC track and intensity predictions can be framed as a sequence-to-sequence prediction task. Consequently, RNN-based models such as LSTM, GRU, and BiGRU are commonly employed due to their ability to capture temporal dependencies (Gan et al., 2024; Pan et al., 2019; Qin et al., 2021; Song et al., 2022). In addition, some studies explored predictions using convolutional neural network (CNN)-based models (Tong et al., 2022) or CNN-RNN hybrid approaches (Tong et al., 2022) from the perspective of receptive field. These investigations primarily focused on the recurrent processing of TC information. In contrast, Transformer-based models excel at capturing global dependencies and are increasingly becoming the backbone of modern weather forecasting models (Bi et al., 2023; Chen, Guo, et al., 2023; Chen, Zhong, et al., 2023). Therefore, exploring the application of Transformer-based models for predicting TC track and intensity is a promising area. When developing a Transformer-based model for short-term forecasting, the first step involves embedding multivariate data into tokens. This study investigates two different kinds of tokens, namely, temporal tokens and variate tokens, as illustrated in Figure 3, which represent different embeddings. In the vanilla Transformer model, the variables

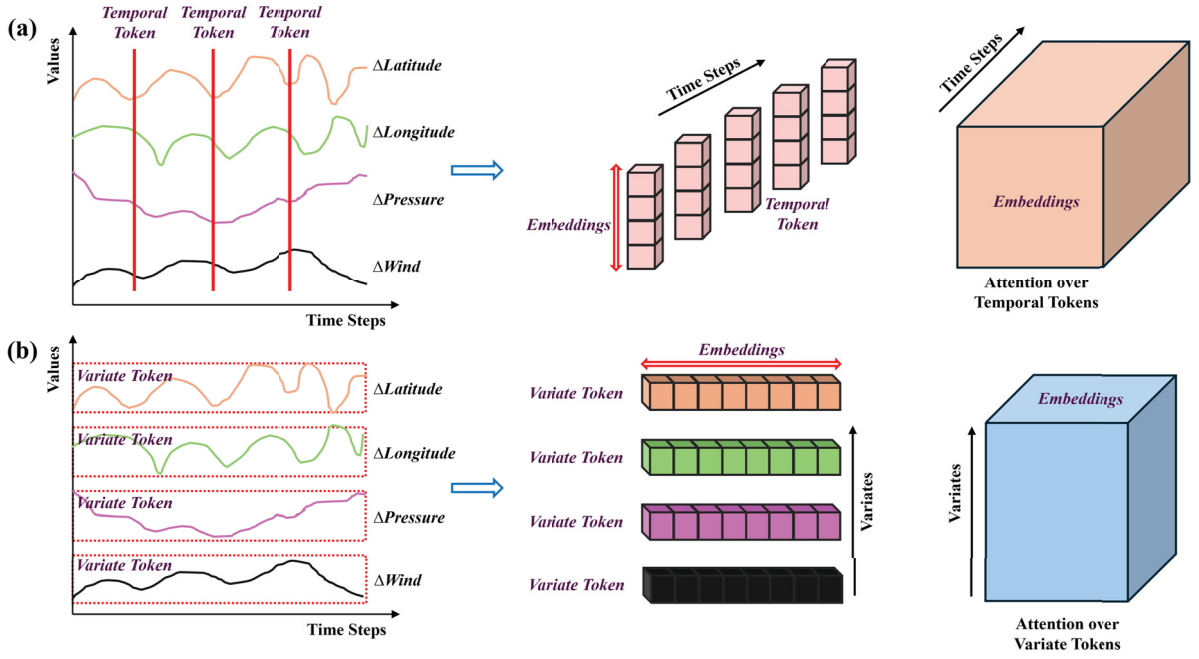

Figure 3. Illustration of token embeddings: (a) temporal tokens; (b) variate tokens.

at each time step are fused to create temporal tokens, resulting in an attention map that illustrates the correlation among various temporal tokens (Figure 3(a)). Conversely, an inverted embedding way, which incorporates each variable at whole time steps as variate tokens, has been introduced to tackle the correlation among different variables (Figure 3(b)) (Liu, Hu, et al., 2024; Nie et al., 2023). These two embeddings can be expressed as,

$$H = Embedding(\Delta X_{t-h+1:t}) \tag{4}$$

In the above, H represents the latent space following the token embedding process, with temporal tokens denoted as $H_T \in \mathbb{R}^{h\times d}$ and variate tokens denoted as $H_V \in \mathbb{R}^{C\times d}$; d refers to the channels of latent space. These two kinds of tokens are investigated in this study due to three main reasons: First, the track and intensity of TCs are distinct variables obtained from various measurements or analysis/reanalysis methods, indicating that they convey different physical interpretations. Embedding these variables which carry distinct meanings into fused but indistinguishable channels may lead to unreasonable results and may hinder model performance. Second, as illustrated in Figure 2, TCs recorded by the CMA show two predominant track directions: westward-moving and eastward-moving (Luo et al., 2022; Qin et al., 2023). This directional tend is affected by multiple factors such as atmospheric circulation and geographical characteristics (Luo et al., 2022; Qin et al., 2023). Figure 4 gives examples of westward – and eastward-moving TC centres, indicating that the first-order differences in latitude and longitude can be summarised as an increase in latitude and a decrease in longitude, or an increase in both latitude and longitude. The correlation between variate tokens may effectively capture this variation pattern. Furthermore, the two variables representing TC intensity – wind speed and pressure – show strong correlation, as illustrated in Figure 4. The two variables exhibit an inverse relationship: a decrease in pressure corresponds to an increase in wind speed, highlighting the importance of employing variate tokens to analyze these correlations. Finally, TC track, intensity and structure can be correlated and interact with each other (Tan et al., 2022); therefore, relying solely on temporal correlations may face challenges in capturing track and intensity changes.

This study will employ three models: the vanilla Transformer, iTransformer, and TVFormer to facilitate the prediction of short-term track and intensity of TCs. Temporal tokens, variate tokens, and a combination of them will be utilised in the three models, which will be described in the next subsection, to explore the impact of different token embeddings on prediction performance.

2.3. Architecture of Transformer-based models

2.3.1. Architecture of vanilla Transformer

Vanilla Transformer is commonly used in time series forecasting due to its superior performance in parallel computation compared to RNN-based models. As shown in Figure 5, the architecture of a vanilla Transformer includes token embeddings, position encoding,

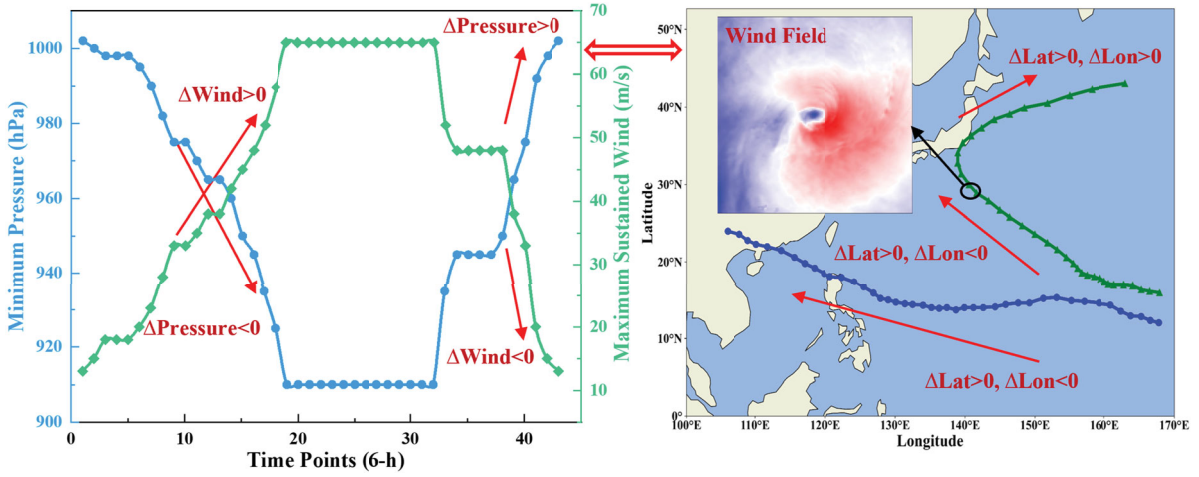

Figure 4. Illustration of interaction between TC track and intensity.

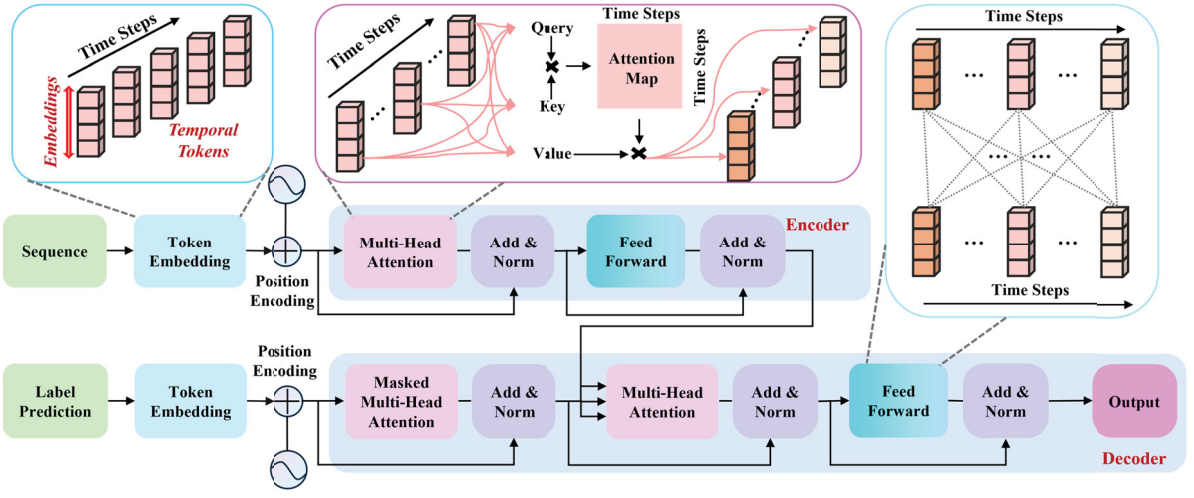

Figure 5. Overview of architecture of vanilla Transformer.

encoder blocks, and decoder blocks. This study specifically focuses on token embeddings; as mentioned earlier, the generation of temporal tokens occurs after token embedding in the vanilla Transformer. The temporal tokens are then further processed by adding position encoding. The resulting embedding, E_T, is as follows,

$$E_T = H_T + PE(\Delta X_{t-h+1:t}) \tag{5}$$

where PE(·) stands for position encoding. This study adopts the typical sinusoidal position encoding approach (Vaswani, 2017) where embeddings are used to compute the attention map, and employs the salient innovation of self-attention in the following manner,

$$Attention(Q_T, K_T, V_T) = softmax\left(\frac{Q_T {K_T}^T}{\sqrt{d_k}}\right) V_T \tag{6}$$

$$\begin{cases} Q_T = E_T W_T^Q \\ K_T = E_T W_T^K \\ V_T = E_T W_T^V \end{cases} \tag{7}$$

where Q_T, K_T, and V_T represent the query, key, and value, respectively, which can be derived from the embedded latent space through linear weighting parameters $W_T^{(\cdot)}$. The attention map is obtained by calculating the dot product of Q_T and the transposed K_T, followed by scaling it by a factor of $1/\sqrt{d_k}$. Moreover, the computation of multi-head attention significantly improves computational efficiency. The self-attention is computed for each head, representing a subspace of feature representations, and is ultimately concatenated and linearly projected to produce the final attention map. The attention mechanism allows to represent similarities between the embedded latent spaces across various time steps. Consequently, the Transformer employs an attention mechanism to

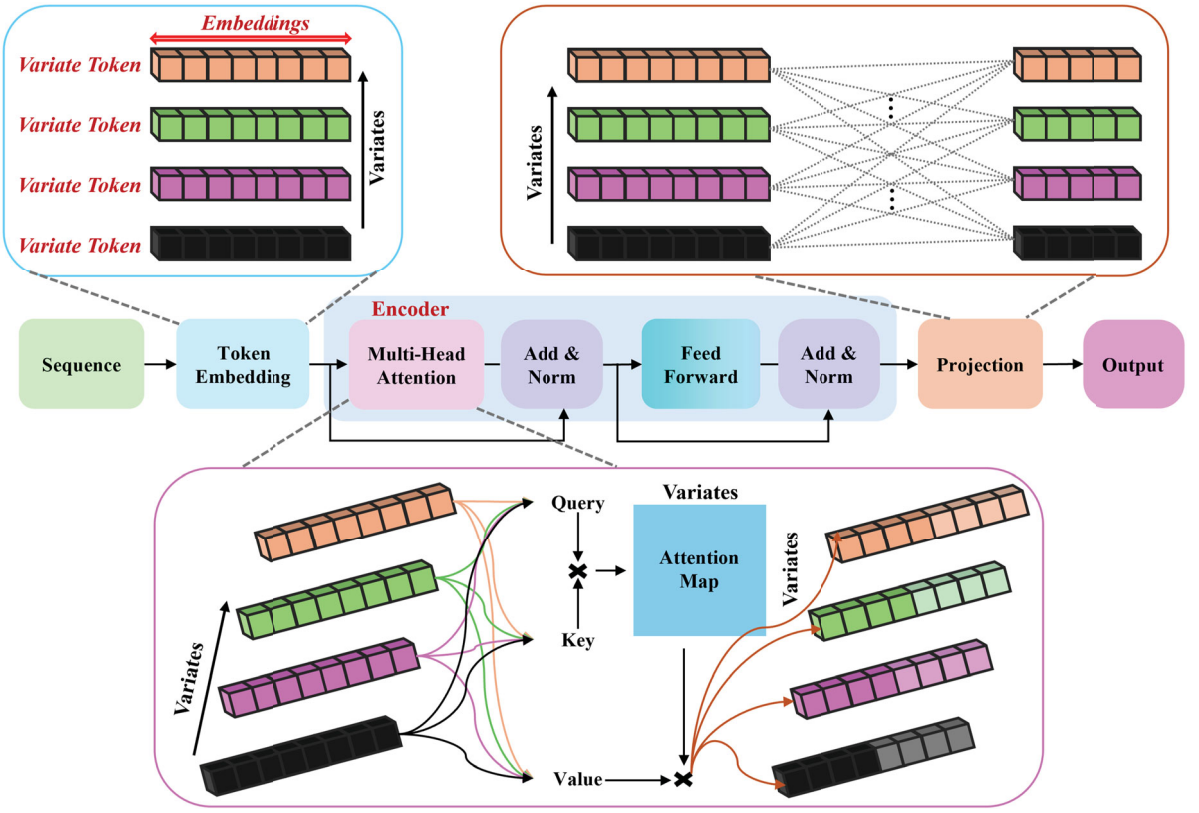

Figure 6. Overview of the architecture of iTransformer.

assess the correlations of embeddings $E_T \in \mathbb{R}^{h\times d}$. The resulting attention map $A_T \in \mathbb{R}^{h\times h\times d}$ is thus primarily concentrated on the temporal dependency.

The encoder block and decoder block share similar layers; however, there are two significant distinctions between them. The first is that each decoder layer includes two self-attention computations, while each encoder layer only performs it once. The second is that the sequence length after token embedding and position encoding provides the input to the encoder for self-attention calculation, while length of the prediction and label serve as the input to the decoder for masked self-attention computation.

2.3.2. Architecture of iTransformer

iTransformer is an innovative model that aims to address the limitations of the vanilla Transformer in time series forecasting tasks. It demonstrates superior performance compared to other Transformer-based models on public benchmark datasets (Liu, Hu, et al., 2024). Figure 6 illustrates the architecture of iTransformer, indicating that its components are the same as those in the vanilla Transformer. In iTransformer, since the token embeddings are processed along the temporal dimension while maintaining the distinguishability of the variate dimension, the position encoding is not necessarily required. The attention map can be computed using the self-attention mechanism with the generated variate tokens as follows.

$$Attention(Q_V, K_V, V_V) = softmax\left(\frac{Q_V {K_V}^T}{\sqrt{d_k}}\right) V_V \quad (8)$$

$$\begin{cases} Q_V = H_V W_V^Q \\ K_V = H_V W_V^K \\ V_V = H_V W_V^V \end{cases} \quad (9)$$

where Q_V, K_V, and V_V denote the query, key, and value, respectively, which are obtained from the variate tokens via the linear weight parameters $W_T^{(\cdot)}$. The self-attention mechanism produces an attention map $A_V \in \mathbb{R}^{C\times C\times d}$, which can subsequently be used to interact with V_V, allowing greater weights to be assigned to paired variables that exhibit higher correlation. The focus of the attention mechanism in iTransformer is to capture multivariate correlations rather than the temporal correlations typically found in the vanilla Transformer.

It is worth noting that after the computation of self-attention, applying add-and-norm, and passing through feedforward networks, the dimension of the output encoder block, $O_V \in \mathbb{R}^{C\times d}$, remains the same as the dimension of the variate tokens. A linear projection is then used to directly map this output to the desired prediction $P \in \mathbb{R}^{C\times p}$, which eliminates the need for a

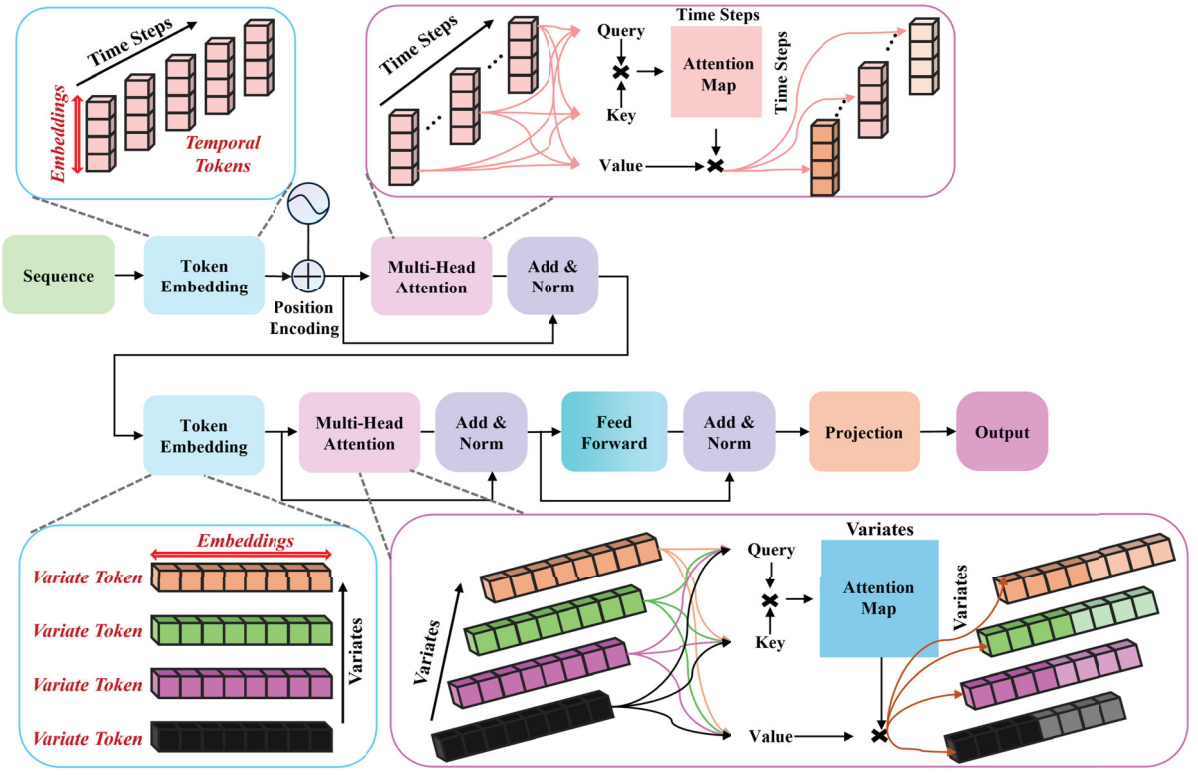

Figure 7. Overview of architecture of TVFormer.

decoder block in the vanilla Transformer within the iTransformer. The rationale is that iTransformer maintains the variate dimension unchanged to underscore the multivariate correlation, which is already captured by the attention map derived from historical data. If the same decoder architecture of the vanilla Transformer is used, no additional multivariate information will be acquired. Additionally, the temporal dimension in the vanilla Transformer is preserved since it is essential for computing the attention map required to generate the target prediction. In contrast, the temporal dimension in iTransformer is integrated and assembled, allowing for a direct mapping of the final prediction length. If the same decoder architecture as the vanilla Transformer is used, the temporal dimension of the decoder input remains embedded, resulting in no distinction compared to directly mapping the encoder output. Consequently, direct projection is sufficient and can simplify the model, thus reducing the computational cost.

2.3.3. Architecture of TVFormer

The aforementioned vanilla Transformer and iTransformer focus on self-attention computations utilising only temporal tokens and variate tokens, respectively. Inspired by hybrid convolutional attention modules such as the Convolutional Block Attention Module (CBAM) (Woo et al., 2018), in this study, a Transformer-based model that integrates dual attention calculations using both temporal and variate tokens and draws, named TVFormer, is proposed to predict the track and intensity of TCs. As shown in Figure 7, the core concept of the proposed TVFormer involves generating temporal tokens to compute attention, and then deriving variate tokens from the above to perform a second attention computation. This two-step process aims to effectively capture both temporal correlations and variate correlations. Specifically, the latent space is obtained using the procedure described in Equation (5), followed by attention computation using Equations (6) and (7). The resulting output, $O_T \in \mathbb{R}^{h \times d}$, is first embedded as follows,

$$H_R = Embedding(O_T) \tag{10}$$

The embedded features $H_R \in \mathbb{R}^{h \times C}$ can then be used to generate variate tokens. Attention is subsequently computed using Equations (8) and (9), and prediction is achieved via projection, which is the same as used in iTransformer. This process illustrates that the interim features H_R encapsulate information regarding temporal correlations, while the attention computation with variate tokens ensures the effective representation of multivariate correlations.

2.4. Evaluation metrics

The prediction results are evaluated using three metrics: mean absolute error (MAE), root mean square error (RMSE), and R^2 score, using the following formulas:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}| \tag{11}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} \tag{12}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \tag{13}$$

where y_i, $\widehat{y_i}$, and $\overline{y_i}$ denote the target values, predicted values, and mean value of the variable i, respectively; n represents the total number of values associated with the variable i. Lower MAE and RMSE and higher R^2 scores indicate better model performance.

2.5. Experimental setup

In this study, two RNN-based models – LSTM and GRU – along with their improved versions, ConvLSTM and BiGRU, are employed to compare their prediction performance with the Transformer-based models. Hyperparameters are set according to the commonly used settings in previous studies (Gan et al., 2024; Hao et al., 2024; Meng & Song, 2024; Song et al., 2022; Tong et al., 2022) and refined by the authors through grid search. The sequence length is configured to 32 (Jiang et al., 2023), and the label and prediction lengths are set in a 1:1 ratio. An unfixed time window is adopted, the minimum length of the input sequence is set to 7, and necessary padding is used to ensure the required length. The hidden layer size of the RNN-based models is set to 128. The Transformer-based models each consist of two encoder blocks and one decoder block, featuring 8 attention heads and a dropout rate of 0.1. The dimensions of the latent space and feed-forward networks are set to 512 and 2048, respectively. The batch size for all models is 32, and the learning rate is set to 0.001. The number of training epochs is set high, three trials are conducted for each model, and an early stopping criterion is implemented. The best models, determined by the lowest validation loss, are selected for performance evaluation. The activation function used is *GELU*, and the *Adam* optimiser is adopted. Training is performed on a platform equipped with 64 GB of RAM and a single RTX 6000 Ada 48 GB GPU, using Python 3.10 and PyTorch 2.1.0 for programming and training.

3. Results and discussion

3.1. Comparison with LSTM and GRU

Table 2 compares the prediction performance of the RNN-based models and the Transformer-based models, where the predictions with the smallest errors are underlined (for the sake of brevity, the term 'vanilla Transformer' is shortened to 'Transformer' hereafter). It is evident that among the RNN-based models studied, GRU outperforms LSTM in predicting track and intensity; this conclusion is consistent with previous studies (Jiang et al., 2023; Song et al., 2022). Meanwhile, when it comes to the Transformer-based models studied, iTransformer and TVFormer emerge as the frontrunners in terms of track prediction and intensity prediction, respectively. The performance of these two models surpasses that of Transformer, GRU, and LSTM, further demonstrating the potential of Transformer-based models in predicting the short-term track and intensity of TCs. However, it is worth noting that although Transformer provides more accurate prediction than LSTM, it does not consistently outperform GRU, especially in terms of wind speed and 24-hour track forecasts.

In addition, two more models, ConvLSTM and BiGRU, are selected as baselines for the improved LSTM and GRU. ConvLSTM demonstrates an improvement over LSTM, while BiGRU shows higher performance than GRU, which is consistent with previous research results (Song et al., 2022; Tong et al., 2022). It is also worth noting that the Transformer models generally have lower prediction accuracy for track and wind speed compared to ConvLSTM and BiGRU. However, overall, iTransformer and TVFormer tend to provide more accurate TC track and intensity forecasts compared to ConvLSTM and BiGRU, except for the 6-hour intensity forecast of iTransformer. This suggests that while the attention mechanism in Transformer has improved model performance in track and intensity prediction, the results indicate that temporal tokens may not be the most effective choice for applying the attention mechanism in these predictions. This finding highlights the need to study the impact of different token embeddings on the performance of Transformer-based models.

3.2. Prediction performance of Transformer-based models

3.2.1. Track prediction

The track predictions obtained from the three Transformer-based models for 6-hour, 12-hour, 18-hour, and 24-hour lead times are illustrated in Figures 8–11, respectively. Overall, the predictions are effective with the predicted track points fluctuating closely around the true

Table 2. Comparison of prediction accuracy using RNN-based models and Transformer-based models. The smallest errors for each lead time are underlined.

		Latitude (°N)		Longitude (°E)		Pressure (hPa)		Wind Speed (m/s)	
Lead Times	Models	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
6-h	LSTM	0.51	0.66	0.90	1.25	3.31	5.39	1.93	2.99
	ConvLSTM	0.39	0.51	0.42	0.59	2.92	4.31	1.70	2.40
	GRU	0.41	0.56	0.55	0.80	3.23	4.77	1.90	2.65
	BiGRU	0.37	0.49	0.39	0.56	2.94	4.40	1.73	2.43
	Transformer	0.38	0.50	0.50	0.69	2.91	4.34	1.90	2.60
	iTransformer	0.29	0.38	0.33	0.49	2.93	4.33	1.73	2.40
	TVFormer	0.37	0.48	0.39	0.57	2.81	4.23	1.68	2.35
12-h	LSTM	1.00	1.30	1.81	2.50	6.28	9.46	3.69	5.28
	ConvLSTM	0.79	1.04	0.87	1.24	5.14	7.45	2.95	4.17
	GRU	0.80	1.07	1.06	1.53	5.61	8.08	3.26	4.51
	BiGRU	0.75	1.00	0.82	1.16	5.28	7.70	3.02	4.24
	Transformer	0.76	1.01	0.99	1.35	5.13	7.46	3.39	4.58
	iTransformer	0.60	0.81	0.70	1.03	5.13	7.44	2.96	4.16
	TVFormer	0.69	0.92	0.80	1.14	4.90	7.19	2.86	4.04
18-h	LSTM	1.49	1.92	2.74	3.77	8.89	13.03	5.26	7.31
	ConvLSTM	1.21	1.59	1.37	1.96	7.33	10.40	4.19	5.82
	GRU	1.16	1.56	1.53	2.20	7.68	10.87	4.46	6.10
	BiGRU	1.15	1.53	1.29	1.84	7.71	10.85	4.34	5.95
	Transformer	1.17	1.55	1.51	2.05	7.23	10.28	4.81	6.42
	iTransformer	0.96	1.31	1.15	1.66	7.24	10.23	4.19	5.75
	TVFormer	1.10	1.45	1.24	1.75	6.94	9.87	4.02	5.55
24-h	LSTM	1.96	2.53	3.69	5.07	11.14	16.09	6.61	9.08
	ConvLSTM	1.64	2.16	1.92	2.74	9.24	12.92	5.29	7.28
	GRU	1.49	2.01	1.97	2.81	9.39	13.18	5.47	7.42
	BiGRU	1.57	2.09	1.82	2.60	9.96	13.66	5.54	7.48
	Transformer	1.61	2.12	2.11	2.83	9.08	12.68	6.10	8.06
	iTransformer	1.37	1.88	1.67	2.40	9.09	12.68	5.23	7.13
	TVFormer	1.50	1.99	1.73	2.45	8.67	12.16	5.03	6.90

values. Notably, for the 24-hour lead time track prediction, the R^2 values of the predicted tracks from all three Transformer-based models exceed 0.96, indicating strong prediction capability. In addition to MAE and RMSE, the R^2 values also demonstrate that iTransformer provides the highest prediction accuracy for all examined lead times, followed by TVFormer, while Transformer ranks last. Regarding the prediction accuracy of latitude and longitude, it is observed that the predicted longitude consistently exhibits larger errors than latitude in all four lead times. This observation could be explained by the spatial distribution of tracks illustrated in Figure 2, which shows that the variations in longitude, specifically from 100°E to 180°, are more pronounced than those in latitude, ranging from 0° to 60°N. Moreover, as illustrated in Figure 4, the variations in latitude are mainly positive, indicating that the TCs tend to move northward in the latitude dimension. The change in longitude can be either positive or negative, indicating that the TCs can move westward or eastward along the longitude dimension. Consequently, changes in longitude are more pronounced than in latitude, regardless of magnitude or direction. As a result, the models may struggle to capture these larger changes, resulting in larger errors in longitude prediction.

In the scenario of 6-hour lead time track prediction, the latitude predictions generated by iTransformer are significantly better than those generated by Transformer and TVFormer. The MAE of iTransformer is 24% and 22% smaller than the MAEs of Transformer and TVFormer, respectively. Moreover, the predicted longitude of iTransformer and TVFormer significantly outperforms the predicted latitude of Transformer, with the MAEs of iTransformer and TVFormer being 34% and 15% smaller than that of Transformer, respectively. When the lead time for prediction increases from 6 hours to 12, 18, and 24 hours, similar trends can be observed. The predicted latitude of TVFormer is comparable to that of Transformer with TVFormer performing slightly better, while iTransformer provides much better performance. In terms of longitude prediction, TVFormer is close to iTransformer with iTransformer slightly outperforming TVFormer, while Transformer has the lowest performance among the three models. It is worth noting that prediction error increases with the increase of lead time due to error accumulation, which is a common challenge in time series prediction tasks. Specifically, when the lead time changes from 6 hours to 24 hours, the latitude MAE increases by 1.23°, 1.08°, and 1.13°, respectively, while the latitude RMSE increases by 1.62°, 1.50°, and 1.51° for Transformer, iTransformer, and TVFormer. Meanwhile, the longitude MAE increases by 1.61°, 1.34°, and 1.34°, respectively, while the longitude RMSE increases by 2.14°, 1.91°, and 1.88° for the same models over the same lead time interval. The statistical confidence intervals of the three metrics are also shown in Figures 8–11.

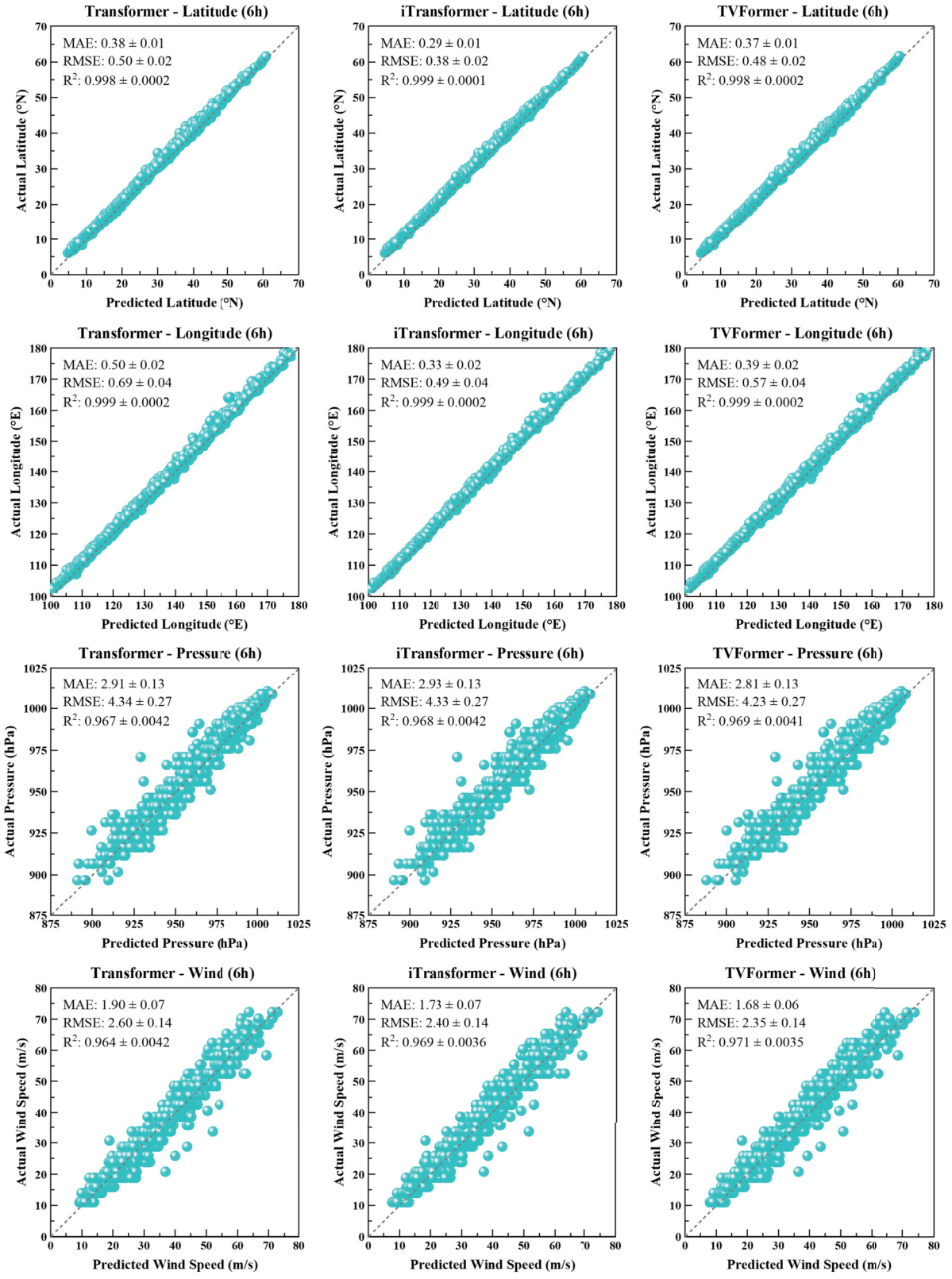

Figure 8. Scatter plots of 6-h track and intensity predictions using three Transformer-based models.

It is noteworthy that the track prediction intervals of iTransformer are similar at a lead time of 6 hours, and as the lead time increases, its intervals are smaller than those of the other two Transformer-based models, indicating that iTransformer has strong prediction capabilities with low uncertainty. However, similar to the error accumulation observed in MAE, RMSE, and R^2, the intervals for all three Transformer-based models also expand with

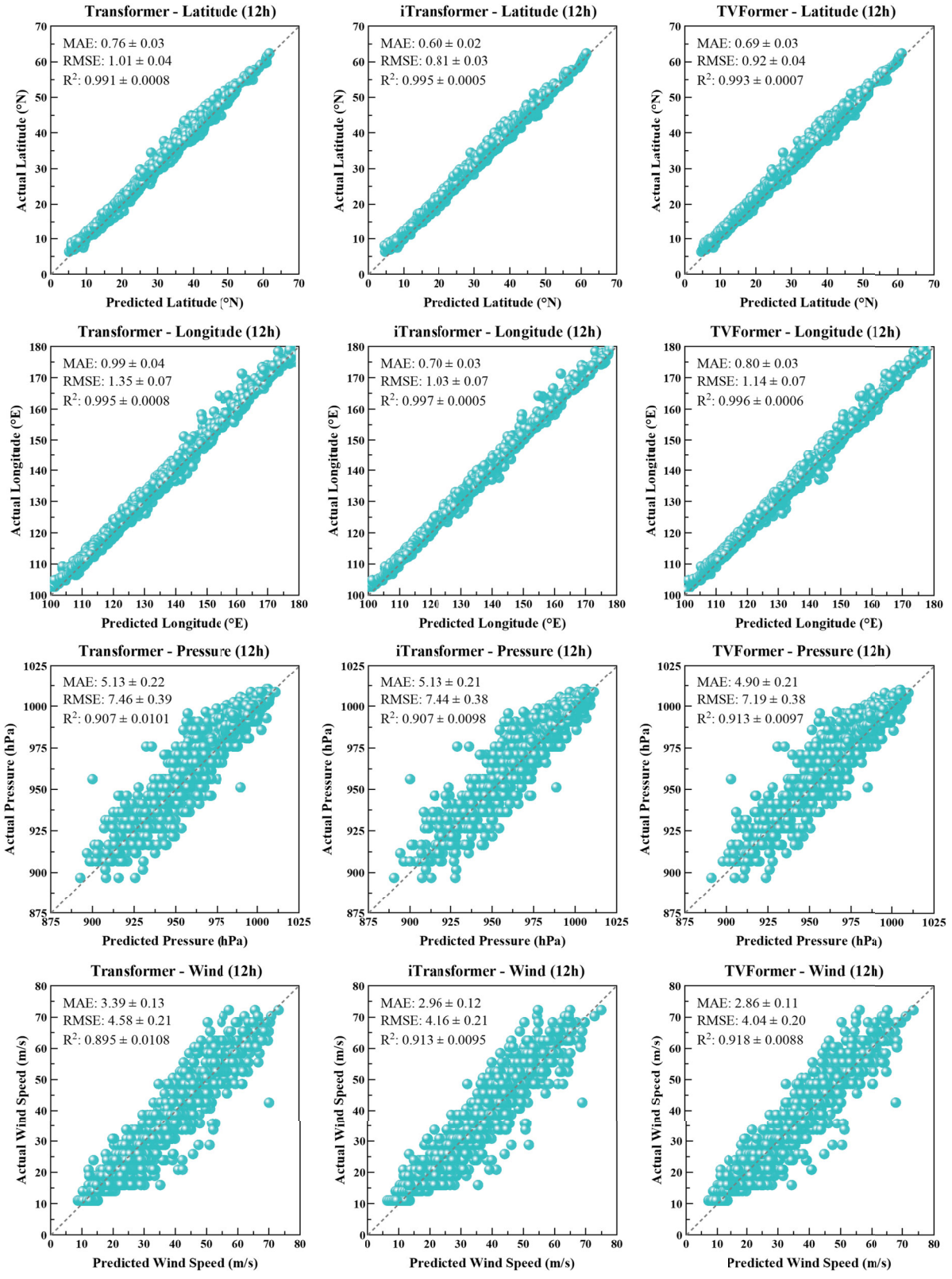

Figure 9. Scatter plots of 12-h track and intensity predictions using three Transformer-based models.

increasing lead time. This limitation has been pointed out in related evaluations (Gan et al., 2024; Jiang et al., 2023; Song et al., 2022).

The superior results achieved by iTransformer show that using variate tokens for attention computation can effectively capture the multivariate correlations implicit

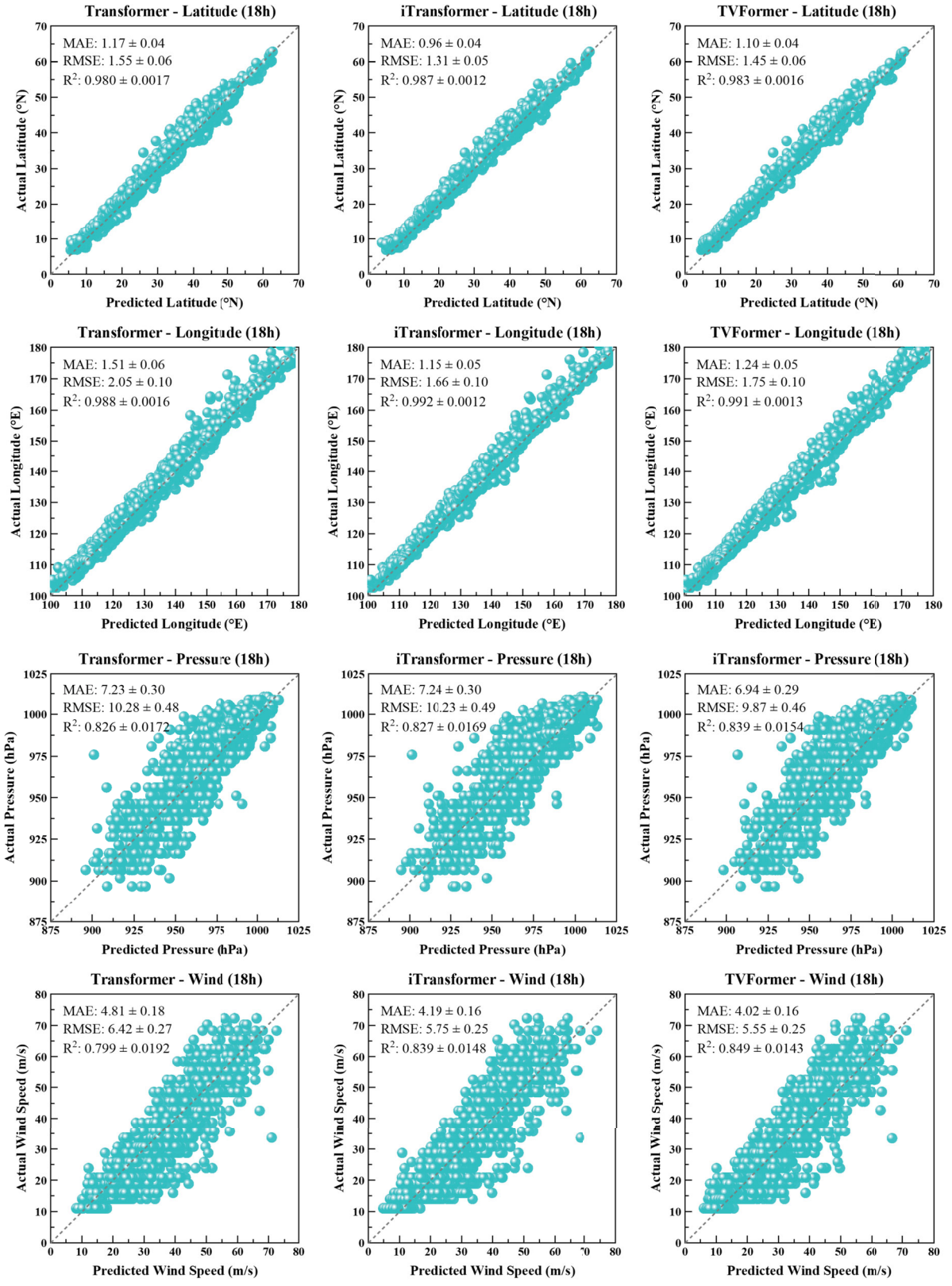

Figure 10. Scatter plots of 18-h track and intensity predictions using three Transformer-based models.

in the variations in track data, which are more significant than temporal dependencies. This capability is beneficial for producing accurate predictions and suppressing error accumulation. In contrast, using temporal tokens for attention computation is less effective and may even prevent the Transformer from generating

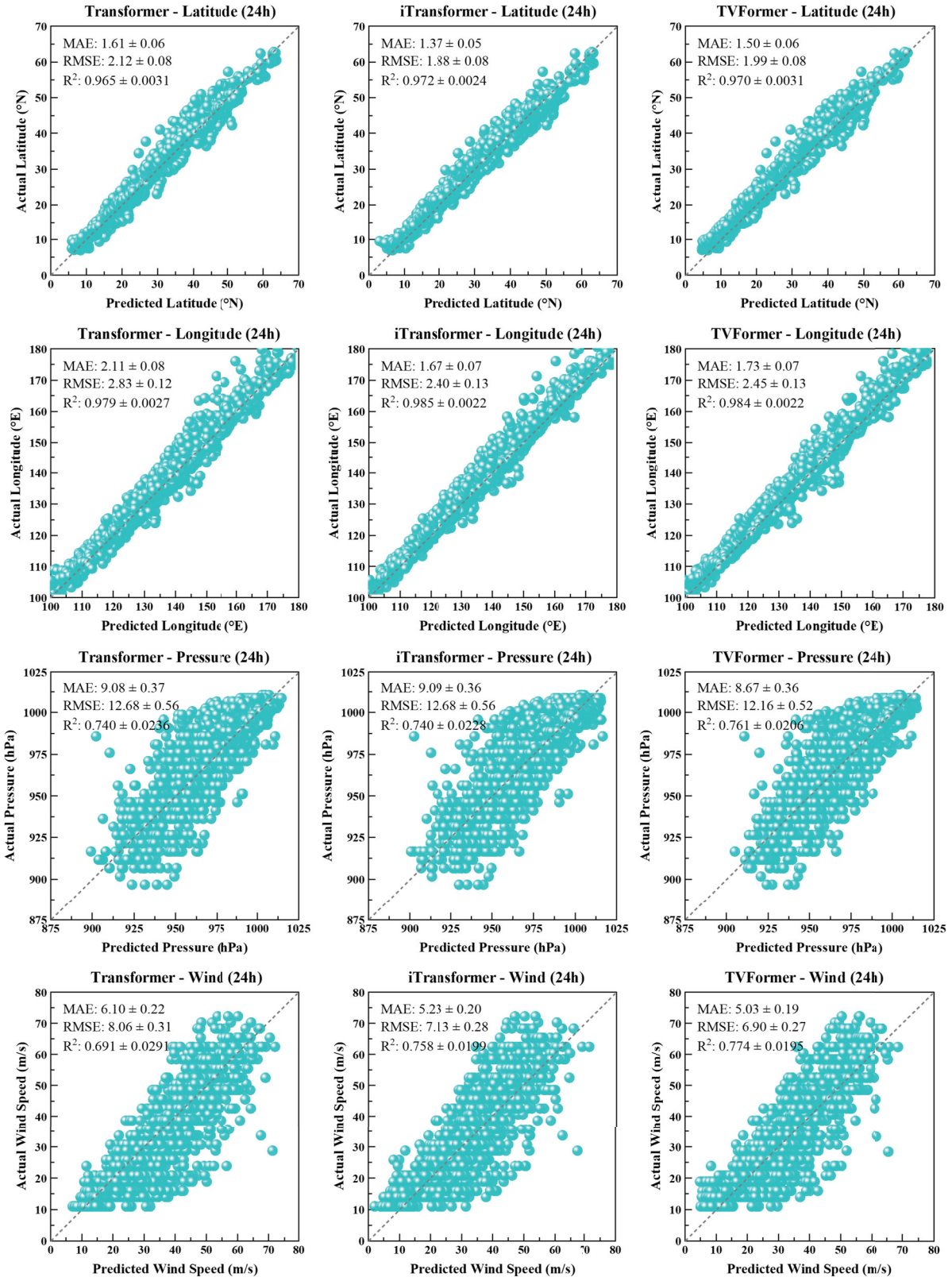

Figure 11. Scatter plots of 24-h track and intensity predictions using three Transformer-based models.

efficient predictions. Consequently, the performance of TVFormer lies between Transformer and iTransformer, which may be attributed to the balance between the adverse impact of attention computation using temporal tokens and the positive impact of attention computation using variate tokens.

3.2.2. Intensity prediction

Figures 8–11 provide the intensity predictions generated by the three Transformer-based models for lead times of 6 hours, 12, 18, and 24 hours, respectively. Overall, TVFormer shows superior prediction ability, followed by iTransformer and then Transformer. Furthermore, the R^2 values for intensity predictions are notably smaller than those for track predictions. This observation can also be seen from the large fluctuations in intensity predictions in the scatter plot. This is reasonable because the variations in wind speed and pressure as shown in Figure 4 are significantly larger than the variations in latitude and longitude. In addition, the R^2 margins for wind speed prediction by TVFormer are the smallest overall. At lead times of 18 and 24 hours, the R^2 margins of TVFormer are also noticeably lower than those of other models.

Compared to iTransformer and Transformer, the MAE of TVFormer achieves 4% smaller in pressure prediction with 6-hour lead time. On the other hand, the MAE of iTransformer is slightly larger than that of Transformer, while slightly smaller than that of Transformer. Similar trends are observed as the lead time increases. Furthermore, when increasing the lead time from 6 hours to 24 hours, the MAE of pressure prediction increases by 6.17, 6.16, and 5.86 hPa for Transformer, iTransformer, and TVFormer, respectively, and the RMSE of pressure prediction increases by 8.34, 8.35, and 7.93 hPa, respectively.

For wind speed prediction with 6-hour lead time, TVFormer and iTransformer outperform Transformer with 13% and 9% smaller MAEs, respectively. As the lead time increases, the MAEs of TVFormer and iTransformer are consistently smaller than that of Transformer. Moreover, both TVFormer and iTransformer are more effective in mitigating error accumulation. Specifically, when the lead time extends from 6 hours to 24 hours, the MAE of Transformer increases by 4.20 m/s, while the MAEs of iTransformer and TVFormer increase less, by 3.50 and 3.35 m/s, respectively. Meanwhile, the RMSE increases by 5.46 m/s for Transformer, compared to 4.73 m/s for iTransformer and 4.55 m/s for TVFormer. Similar to track forecasting, the statistical confidence intervals for the metrics also increase as the lead time extends, while the error increases and the R^2 score decreases. Nevertheless, TVFormer still shows smaller uncertainties in wind speed and pressure predictions. As the lead time increases, the error accumulation becomes more pronounced, which represents a limitation of time series forecasting.

Given that iTransformer and Transformer show similar performance in pressure prediction, both models may struggle to achieve higher accuracy when relying solely on attention computations using variate or temporal tokens. In contrast, TVFormer provides more accurate predictions by leveraging the advantage of attention computation with both temporal tokens and variate tokens. Moreover, the superior performance of iTransformer in wind speed predictions suggests that the multivariate correlations it captures play an important role in the accuracy. One of the influencing factors may be the ability of iTransformer to identify the correlations illustrated in Figure 4. In addition, iTransformer can effectively and implicitly capture the interaction between track and wind structure, which affects wind speed (Tan et al., 2022). It is worth noting that among the three Transformer-based models, TVFormer's superior performance in predicting wind speed further reinforces the benefits of using attention computation that combines both temporal and variate tokens. In summary, TC intensity prediction can be significantly enhanced by integrating temporal and variate tokens into Transformer-based models.

3.3. Model scalability and efficiency

The Transformer-based models have shown potential in predicting TC track and intensity in the Northwest Pacific. To further evaluate the scalability and generalisation ability of the models trained on the CMA best track data, we obtain TCs over the South Indian and North Atlantic regions from 2018 to 2021 from the International Best Track Archive for Climate Stewardship (IBTrACS) (Knapp et al., 2010). In addition, ConvLSTM and BiGRU models that showed strong performance are included for comparison. For each model, the MAEs for different lead times are averaged to calculate the overall MAEs for different variables, and the corresponding results are presented in Figure 12.

As shown in Figure 12(a), the Transformer-based models show better prediction performance than ConvLSTM and BiGRU in the South Indian region, except for Transformer in predicting wind speed. Specifically, iTransformer outperforms the other four models in predicting TC track, followed by TVFormer. In addition, TVFormer demonstrates the best performance in predicting intensity, while iTransformer and Transformer also perform well in wind speed and pressure prediction, outperforming ConvLSTM and BiGRU. Figure 12(b) shows the performance of TC track and intensity predictions in the North Atlantic. iTransformer continues to demonstrate strong capabilities in track prediction; however, TVFormer fails to maintain its strong performance in pressure and wind speed prediction. Specifically, while TVFormer predicts wind speed most accurately among the five models studied, its pressure predictions have errors of approximately 2% larger than those of ConvLSTM and iTransformer, and about 10% larger than those of Transformer.

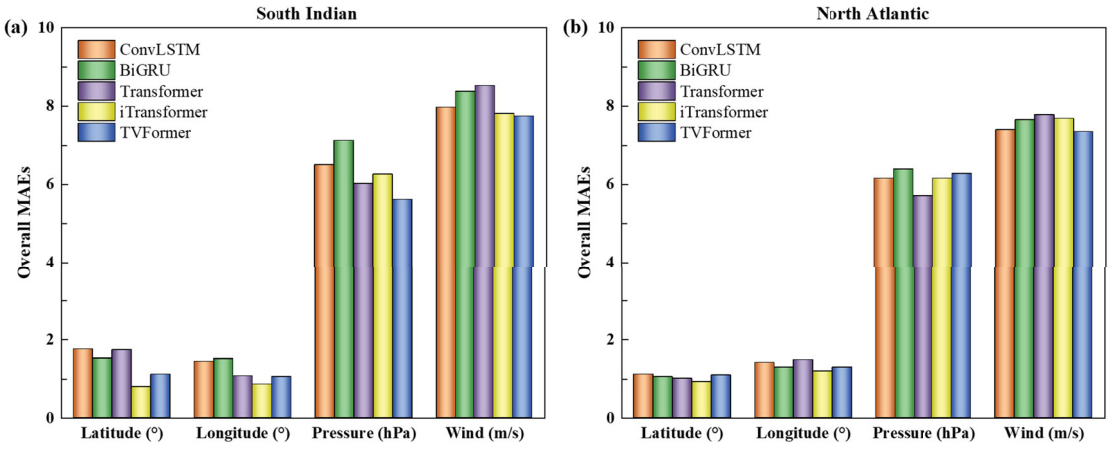

Figure 12. Comparison of overall MAEs for track and intensity predictions in different regions: (a) South Indian; (b) North Atlantic.

The findings indicate that iTransformer exhibits strong scalability in predicting TC track, while TVFormer shows strong performance but lacks scalability when directly applying the model trained on the CMA best track data to predict intensity in the North Atlantic region. Notably, Transformer achieves the most accurate pressure forecasts in the North Atlantic, highlighting the important role of temporal tokens in this region. Furthermore, given the varying characteristics of TCs across different regions and the discrepancies in best track data from various agencies (Knapp & Kruk, 2010; Schreck et al., 2014), fine-tuning or retraining is recommended to ensure that the models can generate accurate predictions.

Using a single GPU card, the runtimes of Transformer, iTransformer, and TVFormer in predicting the Northwest Pacific test set are 14.42, 11.44, and 14.50 s, respectively; the runtimes for predicting the South Indian test set (59 TCs) are 8.38, 8.23, and 8.67 s, respectively; and the runtimes for predicting the North Atlantic test set (84 TCs) are 9.52, 7.37, and 9.64 s, respectively. Furthermore, these test sets cover weather categories of different intensities such as tropical depressions, typhoons, and hurricanes. The fast inference speed of these Transformer-based models offers a significant advantage over traditional numerical forecasting (Bi et al., 2023; Kim et al., 2023; Liu, Tan, et al., 2024), making them a valuable supplement for real-time forecasting systems or operational centres such as CMA or National Oceanic and Atmospheric Administration (NOAA) (Boussioux et al., 2022).

3.4. Model interpretability analysis

To further investigate feature importance, permutation importance analysis is conducted, and the results are presented in Figure 13. The overall loss change of Transformer is different from that of iTransformer, TVFormer, and ConvLSTM; specifically, permuting any of the track, pressure, or wind speed features leads to a decrease in overall loss, and the magnitudes of these changes are relatively similar. This feature importance ranking of Transformer shows that the features are effectively fused through the token embedding. In contrast, for the other three models, intensity plays the largest role in feature importance compared to track, pressure is crucial for ConvLSTM and iTransformer, and wind speed is crucial for TVFormer. This observation aligns with previous analyses indicating that the magnitudes of intensity and track variations differ significantly.

The overall loss changes of iTransformer and TVFormer are larger than those of Transformer and ConvLSTM. This discrepancy may be attributed to the design of Transformer and ConvLSTM, which focuses on capturing temporal dependencies, resulting in smaller overall loss changes due to feature fusion. As illustrated in Figure 13(e), the feature importance indicates that longitude and wind speed are important for iTransformer, highlighting its ability to effectively capture variate correlations. Consequently, the permutation of either track or intensity can notably affect model accuracy. Similarly, Figure 13(f) shows that pressure and longitude are important for TVFormer, reflecting its ability to capture variate correlations. It is noted that the loss changes related to wind speed and pressure are more obvious in TVFormer than in iTransformer, while the loss changes associated with track are more significant in iTransformer. This pattern provides insights into the feature importance of iTransformer and TVFormer and their evaluation results.

Figure 14 shows the attention maps of Transformer, iTransformer, and TVFormer, which are calculated based on the last encoder layer as described in Mylonas et al., 2024. While there is debate about whether attention maps accurately reflect model decisions (Hao et al., 2021; Kovaleva et al., 2019; Tsai et al., 2019), they remain useful for understanding model performance. It is seen that

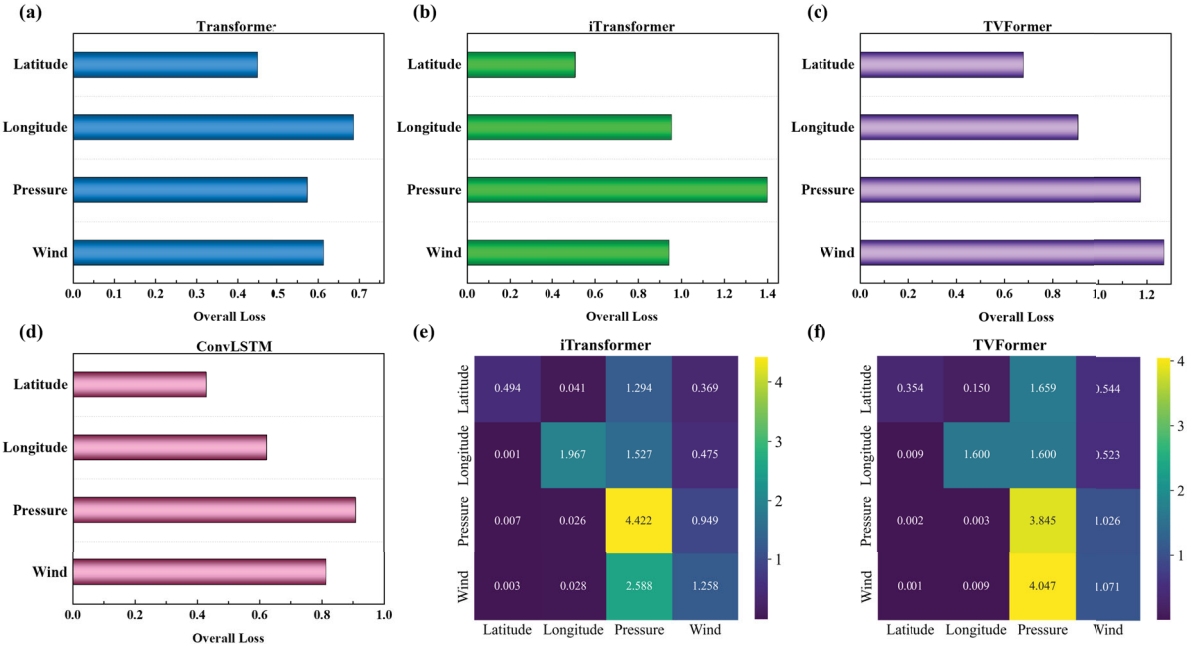

Figure 13. Results of permutation importance: (a)-(d) overall loss changes of Transformer, iTransformer, TVFormer, and ConvLSTM; (e)-(f) loss changes of variables in iTransformer and TVFormer.

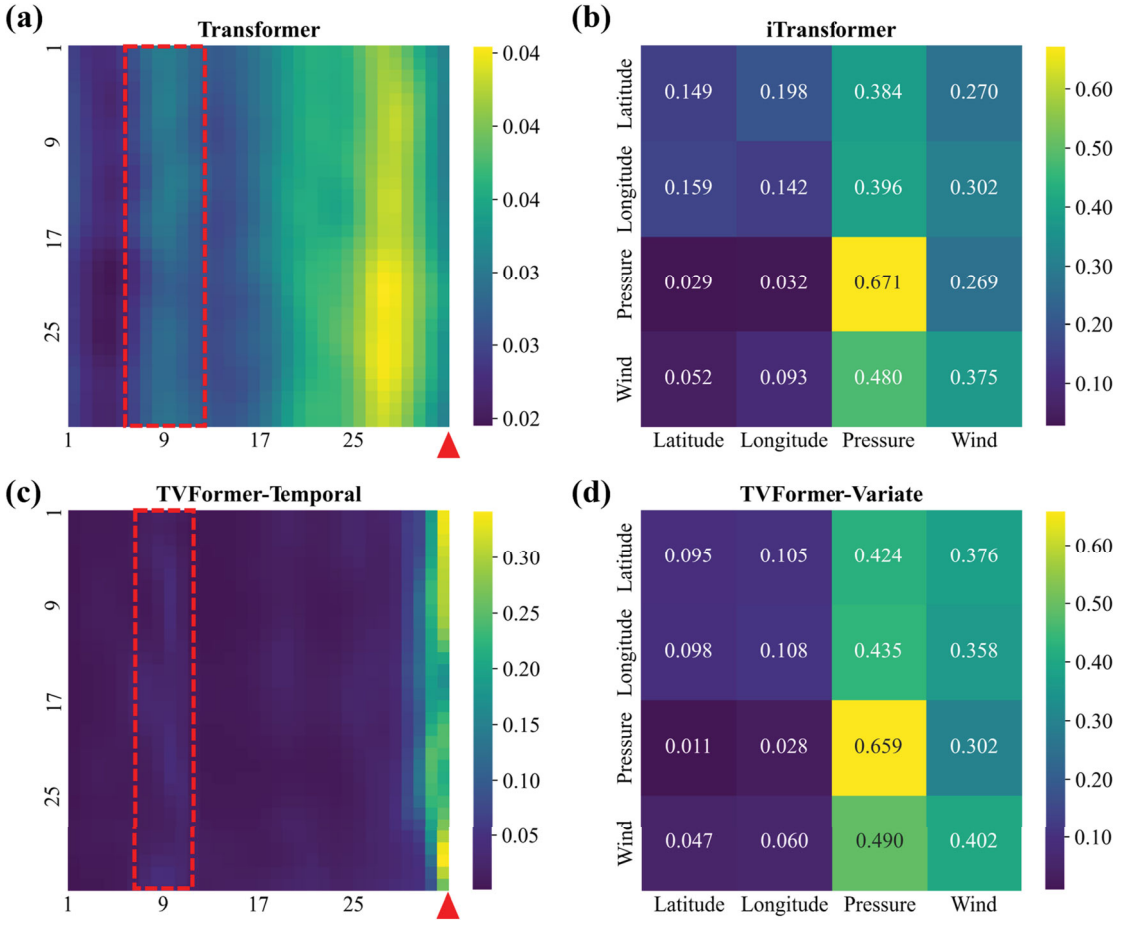

Figure 14. Visualisation of attention maps calculated from the last encoder layer: (a) Transformer; (b) iTransformer; (c)-(d) TVFormer.

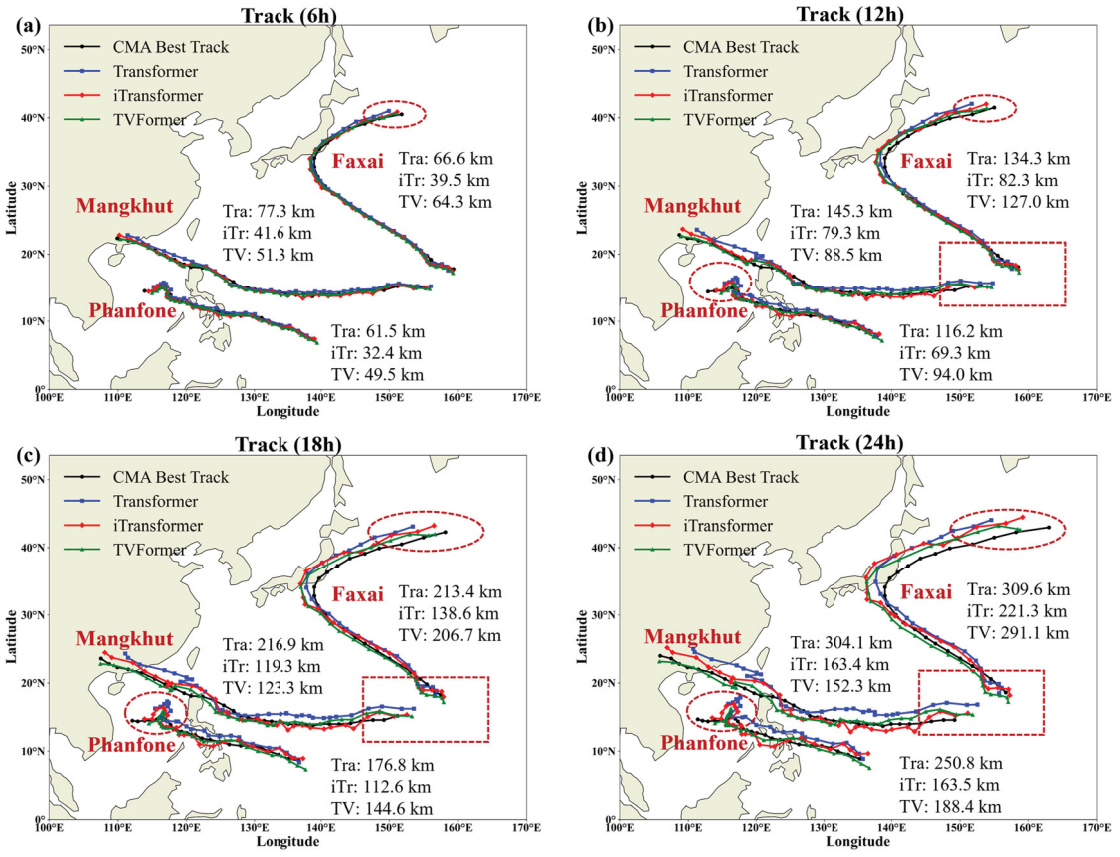

Figure 15. Track predictions for three cases: (a) 6-h; (b) 12-h; (c) 18-h; (d) 24-h. Here 'TV', 'Tra', and 'iTr' represent TVFormer, Transformer, and iTransformer, respectively.

the dimensions of the attention maps of Transformer and iTransformer are different, which aligns with their respective architectures. The attention map of Transformer reveals that larger attention weights are assigned to points near the reference point of the initial prediction (marked with a red triangle in the figure). It is worth noting that the attention weights are also significant near the beginning of the sequence and in the marked area in Figure 14(a). This suggests that long-term sequential dependencies are crucial for predicting TC track and intensity, which is consistent with previous studies (Jiang et al., 2023). The attention map of iTransformer demonstrates strong variate correlations, including correlations between latitude and longitude, between wind speed and pressure, and between track and intensity. This observation aligns with the data shown in Figure 4. Notably, iTransformer exhibits greater attention weights between pressure and track, and between wind speed and track, proving once again that pressure and wind speed are key indicators for tracking TCs (Roy & Kovordányi, 2012; Tan et al., 2022). Looking at the attention maps of TVFormer, it is evident that its temporal component is different from that of Transformer. Larger attention weights are observed at five points near the reference prediction point (marked with a red triangle in the figure), and the marked area in Figure 14(c) also shows higher attention weights compared to the neighbouring points. Moreover, the attention weights for intensity and track in the attention map of TVFormer are more significant compared to iTransformer, which is consistent with previous studies on track-intensity interactions (Knaff & Zehr, 2007). In view of the above, the strong ability of TVFormer in predicting TC intensity suggests that both temporal and variate dependencies are crucial factors.

3.5. Case study

To further demonstrate the performance of the three Transformer-based models, three TCs, Phanfone, Mangkhut, and Faxai, are selected as specific examples. These cases are often used to verify related studies on short-term TC track and intensity predictions using the best track data (Gan et al., 2024; Jiang et al., 2023). TCs Phanfone and Mangkhut moved from southeast to northwest, while TC Faxai initially moved from southeast to northwest and then changed direction to move northeast.

The track predictions of the three TCs at four lead times are illustrated in Figure 15. The mean position error shows that iTransformer has the best performance, followed by TVFormer and then Transformer. This ranking

aligns with the overall evaluation results on the test set. While all three Transformer-based models show potential for predicting short-term TC tracks, the most significant differences occur in the predicted centres during the initial gradual formation and final decay stages of the TCs. The models exhibit a tendency to generate less accurate predictions during these stages, as marked by the dashed ellipses and dashed rectangles in Figure 15. This is understandable, because the translation speeds of the TCs during these stages differ from those in other stages (Sun et al., 2017). Nevertheless, both iTransformer and TVFormer show higher accuracy in predicting TC centres during these periods. For example, at the 6-hour lead time, although all three models provide similar predictions overall, iTransformer and TVFormer yield more accurate estimates of the true centres during the decay

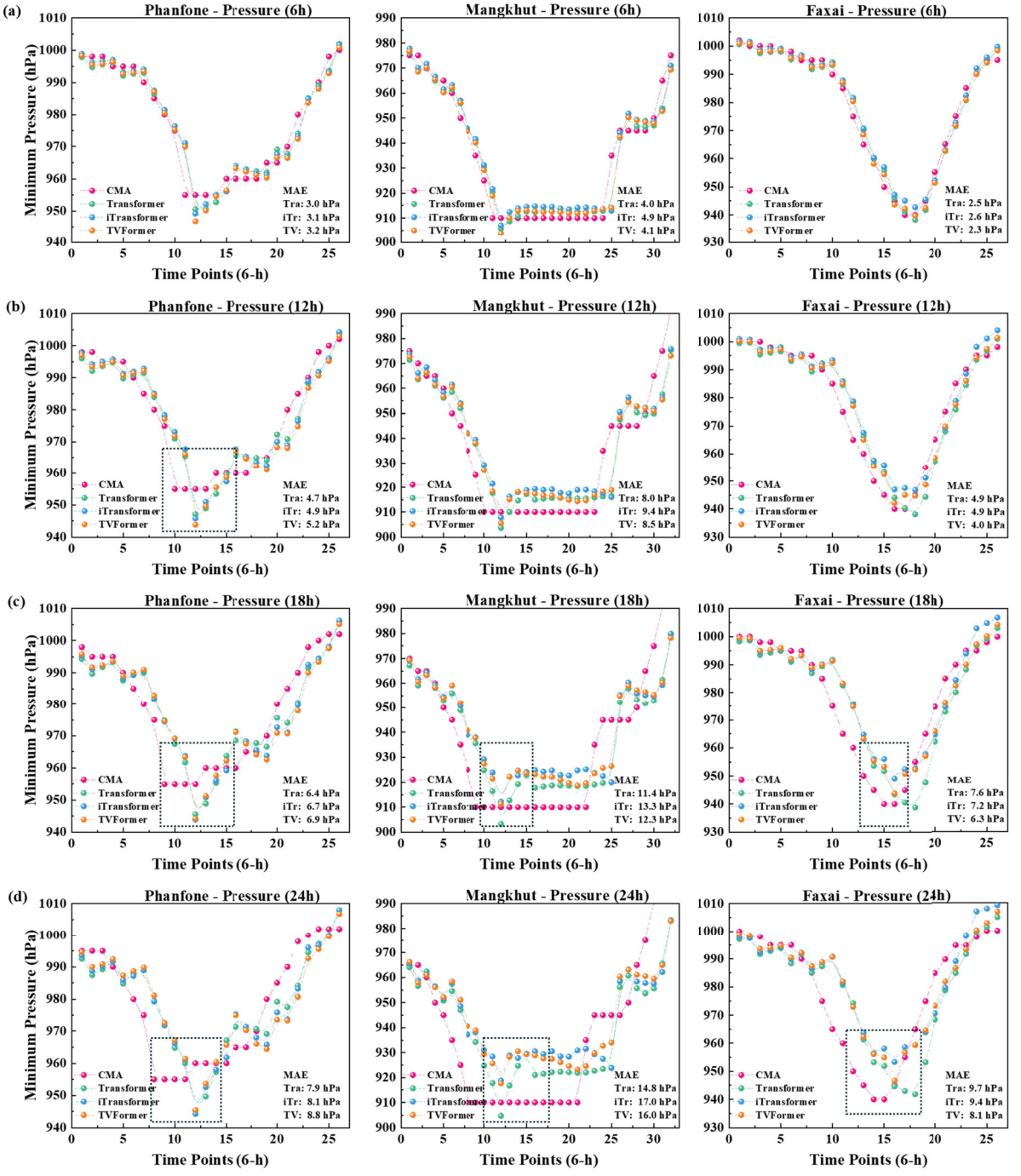

Figure 16. Intensity predictions for three cases: (a) 6-h pressure prediction; (b) 12-h pressure prediction; (c) 18-h pressure prediction; (d) 24-h pressure prediction; (e) 6-h wind speed prediction; (f) 12-h wind speed prediction; (g) 18-h wind speed prediction; (h) 24-h wind speed prediction. Here 'TV', 'Tra', and 'iTr' represent TVFormer, Transformer, and iTransformer, respectively.

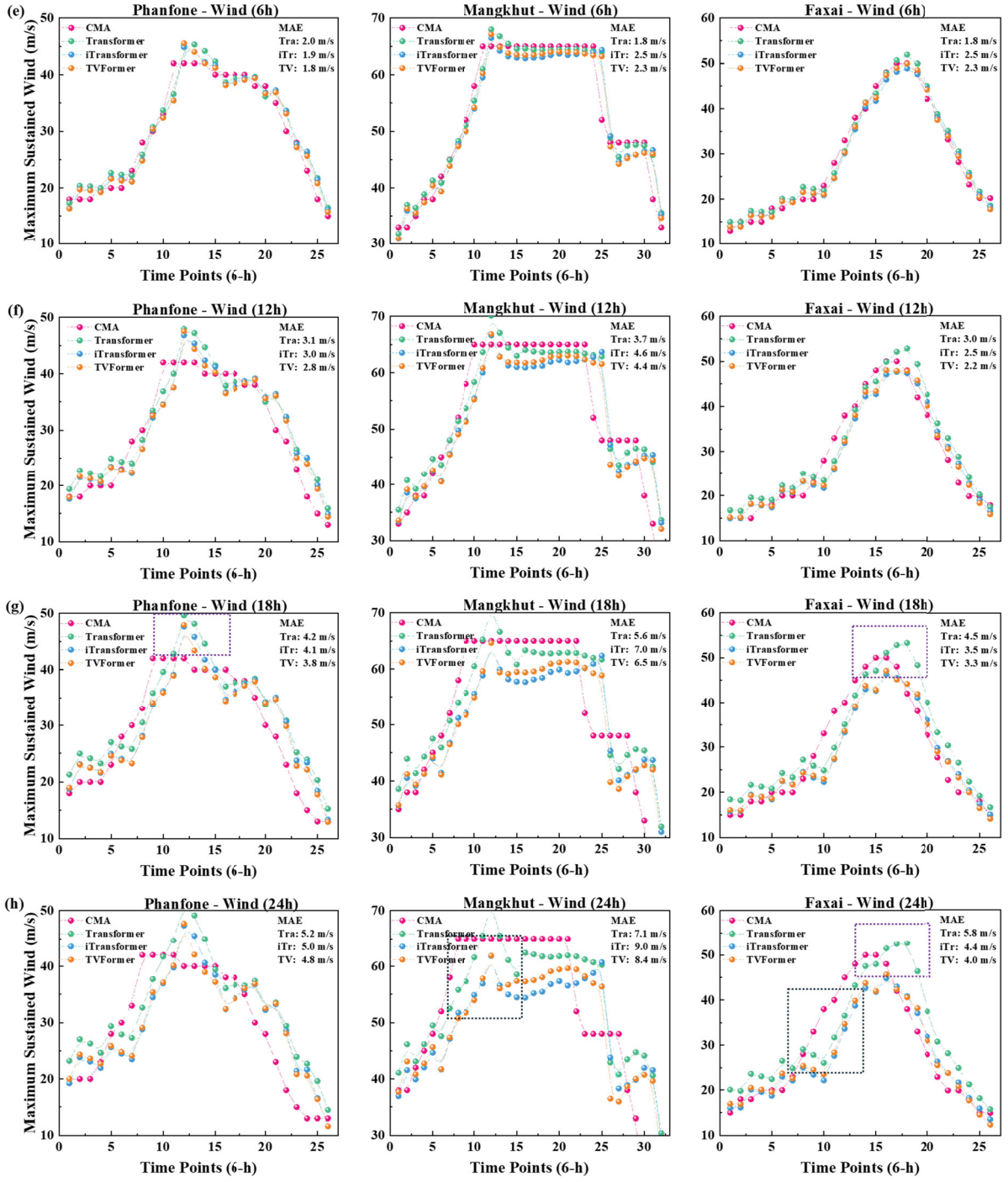

Figure 16. Continued.

stage of TC Faxai. Similar findings are observed at lead times of 12, 18, and 24 hours. Moreover, during the decay stage of TC Phanfone, Transformer shows a larger position error, while iTransformer and TVFormer perform more effectively. In addition, at the initial time points of TCs Faxai and Mangkhut, the predicted TC centres of iTransformer and TVFormer are closer to the true centres.

The intensity predictions of the three TCs at four lead times are illustrated in Figure 16. Overall, the three Transformer-based models effectively capture the trends of TC pressure and wind speed. In the three selected cases, Transformer performs well in predicting the pressure of TCs Phanfone and Mangkhut, while TVFormer excels in predicting the pressure of TC Faxai and in this case iTransformer performs poorly. This finding aligns

with previous studies and indicates that the contribution of temporal tokens in attention computation is critical for accurate pressure prediction. It is noteworthy that all three models tend to overestimate the minimum pressure values compared to the true minimum pressures, as shown by the black dashed rectangles in the figure. Furthermore, the time lag effect becomes evident as the lead time increases, which may be due to the models' tendency to adhere to historical patterns. For wind speed prediction, TVFormer shows strong performance overall. The time lag effect in the wind speed prediction is also observable, as marked by the black dashed rectangles in the figure. However, with the integration of both temporal and variate tokens, peak wind speed values predicted by TVFormer are much closer to the true peak values as shown by the purple dashed rectangles, while Transformer tends to overestimate the wind speed.

4. Conclusions and future research directions

In this study, the effects of different token embeddings in Transformer-based models on the performance of short-term TC track and intensity prediction are investigated. Three models – Transformer, iTransformer, and TVFormer – are trained and evaluated using the best track data from CMA. These models focus on temporal tokens, variate tokens, and integration of both token types generated through different embedding methods. Two Transformer variants, iTransformer and TVFormer, have demonstrated improved prediction accuracy. Their runtime for forecasting 100 TCs is measured in seconds, while traditional numerical forecasting takes hours to calculate one TC. This fast runtime, along with reliance on minimal sequential information for prediction, shows great potential as a complementary tool for early warning systems deployed in the real world. By comparing the performance of the three Transformer-based models and four RNN-based models, the following conclusions can be drawn:

- iTransformer and TVFormer demonstrate superior performance compared to RNN-based models in predicting short-term TC track and intensity, as evidenced by the evaluation metrics. Specifically, iTransformer and TVFormer reduce the MAE by 0.88° and 0.79° compared to LSTM, by 0.24° and 0.14° compared to GRU, by 0.19° and 0.10° compared to ConvLSTM, and by 0.14° and 0.04° compared to BiGRU in track prediction. For pressure prediction, their average MAE values are reduced by 1.31 and 1.58 hPa compared to LSTM, by 0.38 and 0.65 hPa compared to GRU, by 0.06 and 0.33 hPa compared to ConvLSTM, and by 0.38 and 0.64 hPa compared to BiGRU. In wind speed prediction, the average MAE values of iTransformer and TVFormer are reduced by 0.85 and 1.46 m/s compared to LSTM, by 0.38 and 0.65 m/s compared to GRU, by 0.06 and 0.33 m/s compared to ConvLSTM, and by 0.38 and 0.64 m/s compared to BiGRU.
- For track prediction, iTransformer outperforms both TVFormer and Transformer, achieving average MAE reductions of 0.11° and 0.18° in latitude, and 0.08° and 0.32° in longitude, respectively. In addition, TVFormer outperforms Transformer with average MAE reductions of 0.06° in latitude and 0.24° in longitude, due to the integration of variate tokens. The attention map of iTransformer reveals that track and intensity correlations are well captured, highlighting the importance of variate tokens, which are more effective in enhancing track prediction than embedding temporal tokens at the same time step.
- For pressure prediction, TVFormer achieves mean MAE reductions of 0.27 and 0.26 hPa compared to iTransformer and Transformer, respectively. For wind speed prediction, it achieves reductions of 0.13 and 0.65 m/s compared to iTransformer and Transformer, respectively. Furthermore, permutation importance analysis shows that intensity contributes most to the model accuracy of both iTransformer and TVFormer. The integration of temporal and variate tokens in TVFormer enhances accuracy in intensity prediction. Furthermore, the temporal and variate components of TVFormer's attention maps differ from those of Transformer and iTransformer, demonstrating the effectiveness of this integration in capturing correlations.

Future research could be enhanced by expanding the dataset to include data from different basins, enabling fine-tuning or retraining of the model. This approach will improve the scalability and accuracy of model predictions of TC tracks and intensities in various regions. In addition, since TCs are significantly influenced by environmental circulation, combining multi-source data such as reanalysis data can further improve prediction performance by integrating physical constraints. Finally, Transformer-based models present an alternative to statistical-dynamical approaches, with the potential to improve the capabilities of numerical models and ensemble forecasts.

Acknowledgements

The authors would like to express their gratitude for the computing platform provided by the University Research Facility in Big Data Analytics (UBDA) at The Hong Kong Polytechnic University.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China [Grant No. T22-501/23-R] and the Innovation and Technology Commission of the Hong Kong Special Administrative Region, China [Grant No. KBBY1].

Data availability

The best track data used in this study can be downloaded from the China Meteorological Administration (tcdata.typhoon.org.cn).

References

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

Boussioux, L., Zeng, C., Guénais, T., & Bertsimas, D. (2022). Hurricane forecasting: A novel multimodal machine learning framework. *Weather and Forecasting*, *37*(6), 817–831. https://doi.org/10.1175/WAF-D-21-0091.1

Chan, J. (2005). Interannual and interdecadal variations of tropical cyclone activity over the western North Pacific. *Meteorology and Atmospheric Physics*, *89*(1-4), 143–152. https://doi.org/10.1007/s00703-005-0126-y

Charlton-Perez, A., Dacre, H., Driscoll, S., Gray, S., Harvey, B., Harvey, N., Hunt, K., Lee, R., Swaminathan, R., & Vandaele, R. (2024). Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciarán. *npj Climate and Atmospheric Science*, *7*(1), 93. https://doi.org/10.1038/s41612-024-00638-w

Chavas, D., Reed, K., & Knaff, J. (2017). Physical understanding of the tropical cyclone wind-pressure relationship. *Nature Communications*, *8*(1), 1360–1311. https://doi.org/10.1038/s41467-017-01546-9

Chen, Z., Guo, Z., Ni, Y., Liu, T., & Zhang, J. (2023). A suction method to mitigate pressure waves induced by high-speed maglev trains passing through tunnels. *Sustainable Cities and Society*, *96*, 104682. https://doi.org/10.1016/j.scs.2023.104682

Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J., Chen, X., Ma, L., Zhang, T., & Su, R. (2023). Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint*, https://doi.org/10.48550/arXiv.2304.02948

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

Fang, G., Pang, W., Zhao, L., Xu, K., Cao, S., & Ge, Y. (2022). Tropical-cyclone-wind-induced flutter failure analysis of long-span bridges. *Engineering Failure Analysis*, *132*, 105933. http://doi.org/10.1016/j.engfailanal.2021.105933

Gan, S., Fu, J., Zhao, G., Chan, P., & He, Y. (2024). Short-term prediction of tropical cyclone track and intensity via four mainstream deep learning techniques. *Journal of Wind Engineering and Industrial Aerodynamics*, *244*, 105633. https://doi.org/10.1016/j.jweia.2023.105633

Hao, Y., Dong, L., Wei, F., & Xu, K. (2021). *Self-attention attribution: Interpreting information interactions inside transformer*. Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, pp. 12963–12971).

Hao, P., Zhao, Y., Li, S., Song, J., & Gao, Y. (2024). Deep learning approaches in predicting tropical cyclone tracks: An analysis focused on the Northwest Pacific region. *Ocean Modelling*, *192*, 102444. https://doi.org/10.1016/j.ocemod.2024.102444

Hu, F., Yang, S., & Thompson, R. (2021). Resilience-driven road network retrofit optimization subject to tropical cyclones induced roadside tree blowdown. *International Journal of Disaster Risk Science*, *12*(1), 72–89. https://doi.org/10.1007/s13753-020-00301-x

Huo, X., Liu, T., Chen, Z., Li, W., Niu, J., & Gao, H. (2023). Aerodynamic characteristics of double-connected train groups composed of different kinds of high-speed trains under crosswinds: A comparison study. *Alexandria Engineering Journal*, *64*, 465–481. https://doi.org/10.1016/j.aej.2022.09.011

Jiang, W., Zhang, D., Hu, G., Wu, T., Liu, L., Xiao, Y., & Duan, Z. (2023). Transformer-based tropical cyclone track and intensity forecasting. *Journal of Wind Engineering and Industrial Aerodynamics*, *238*, 105440. https://doi.org/10.1016/j.jweia.2023.105440

Ju, S.-H., Hsu, H.-H., & Hsiao, T.-Y. (2021). Three-dimensional wind fields of tropical cyclones for wind turbine structures. *Ocean Engineering*, *237*, 109437. http://doi.org/10.1016/j.oceaneng.2021.109437

Kim, K., Yoon, D., Cha, D., & Im, J. (2023). Improved tropical cyclone track simulation over the Western North Pacific using the WRF model and a machine learning method. *Asia-Pacific Journal of Atmospheric Sciences*, *59*(3), 283–296. https://doi.org/10.1007/s13143-022-00313-1

Knaff, J., & Zehr, R. (2007). Reexamination of tropical cyclone wind – pressure relationships. *Weather and Forecasting*, *22*(1), 71–88. https://doi.org/10.1175/WAF965.1

Knapp, K., & Kruk, M. (2010). Quantifying interagency differences in tropical cyclone best-track wind speed estimates. *Monthly Weather Review*, *138*(4), 1459–1473. https://doi.org/10.1175/2009MWR3123.1

Knapp, K., Kruk, M., Levinson, D., Diamond, H., & Neumann, C. (2010). The international best track archive for climate stewardship (IBTrACS) unifying tropical cyclone data. *Bulletin of the American Meteorological Society*, *91*(3), 363–376. https://doi.org/10.1175/2009BAMS2755.1

Kovaleva, O., Romanov, A., Rogers, A., & Rumshisky, A. (2019). Revealing the dark secrets of BERT. *arXiv preprint*.

Kuleshov, Y., Gregory, P., Watkins, A., & Fawcett, R. (2020). Tropical cyclone early warnings for the regions of the Southern Hemisphere: Strengthening resilience to tropical cyclones in small island developing states and least developed countries. *Natural Hazards*, *104*(2), 1295–1313. https://doi.org/10.1007/s11069-020-04214-2

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lian, J., Dong, P., Zhang, Y., & Pan, J. (2020). A novel deep learning approach for tropical cyclone track prediction based on auto-encoder and gated recurrent unit networks. *Applied sciences*, *10*(11), 3965. https://doi.org/10.3390/app10113965

Liu, C., Hsu, K., Peng, M., Chen, D., Chang, P., Hsiao, L., Fong, C., Hong, J., Cheng, C., & Lu, K. (2024). Evaluation of five global AI models for predicting weather in Eastern Asia and Western Pacific. *npj Climate and Atmospheric Science*, *7*(1), 221. https://doi.org/10.1038/s41612-024-00769-0

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., & Long, M. (2024). iTransformer: Inverted Transformers are effective for time series forecasting. *arXiv preprint*, https://doi.org/10.48550/arxiv.2310.06625

Liu, H., Tan, Z., Wang, Y., Tang, J., Satoh, M., Lei, L., Gu, J., Zhang, Y., Nie, G., & Chen, Q. (2024). A hybrid machine learning/physics-based modeling framework for 2-week extended prediction of tropical cyclones. *Journal of Geophysical Research: Machine Learning and Computation*, *1*(3), e2024JH000207. doi:10.1029/2024JH000207

Lu, X., Yu, H., Ying, M., Zhao, B., Zhang, S., Lin, L., Bai, L., & Wan, R. (2021). Western North Pacific tropical cyclone database created by the China Meteorological Administration. *Advances in Atmospheric Sciences*, *38*(4), 690–699. https://doi.org/10.1007/s00376-020-0211-7

Luo, X., Yang, L., Chen, S., Liang, D., Chan, J., & Wang, D. (2022). The decadal variation of eastward-moving tropical cyclones in the South China Sea during 1980–2020. *Geophysical Research Letters*, *49*(5), e2021GL096640. https://doi.org/10.1029/2021GL096640

Ma, D., Wang, L., Fang, S., & Lin, J. (2023). Tropical cyclone intensity prediction by inter – and intra-pattern fusion based on multi-source data. *Environmental Research Letters*, *18*(1), 014020. https://doi.org/10.1088/1748-9326/aca9e2

Mandal, A., Ramakrishnan, R., Pandey, S., Rao, A., & Kumar, P. (2020). An early warning system for inundation forecast due to a tropical cyclone along the east coast of India. *Natural Hazards*, *103*(2), 2277–2293. https://doi.org/10.1007/s11069-020-04082-w

Meng, F., & Song, T. (2024). Uncertainty forecasting system for tropical cyclone tracks based on conformal prediction. *Expert Systems with Applications*, *249*, 123743. https://doi.org/10.1016/j.eswa.2024.123743

Meng, F., Yang, K., Yao, Y., Wang, Z., & Song, T. (2023). Tropical cyclone intensity probabilistic forecasting system based on deep learning. *International Journal of Intelligent Systems*, *2023*(1), 3569538. https://doi.org/10.1155/2023/3569538

Mylonas, N., Mollas, I., & Tsoumakas, G. (2024). An attention matrix for every decision: Faithfulness-based arbitration among multiple attention-based interpretations of transformers in text classification. *Data Mining and Knowledge Discovery*, *38*(1), 128–153. https://doi.org/10.1007/s10618-023-00962-4

Nie, Y., Nguyen, N., Sinthong, P., & Kalagnanam, J. (2023). A time series is worth 64 words: long-term forecasting with Transformers. *arXiv preprint*, https://doi.org/10.48550/arxiv.2211.14730

Pan, B., Xu, X., & Shi, Z. (2019). Tropical cyclone intensity prediction based on recurrent neural networks. *Electronics Letters*, *55*(7), 413–415. https://doi.org/10.1049/el.2018.8178

Qin, W., Tang, J., Lu, C., & Lao, S. (2021). Trajectory prediction based on long short-term memory network and Kalman filter using hurricanes as an example. *Computational Geosciences*, *25*(3), 1005–1023. https://doi.org/10.1007/s10596-021-10037-2

Qin, X., Yao, J., Hu, J., & Li, C. (2023). Characteristics of the tropical cyclones before making landfall in China. *International Journal of Climatology*, *43*(9), 3963–3976. https://doi.org/10.1002/joc.8067

Roy, C., & Kovordányi, R. (2012). Tropical cyclone track forecasting techniques — a review. *Atmospheric Research*, *104*, 40–69. https://doi.org/10.1016/j.atmosres.2011.09.012

Sampson, C. R., Schumacher, A. B., Knaff, J. A., DeMaria, M., Fukada, E. M., Sisko, C. A, Roberts, D. P., Winters, K. A., & Wilson, H. M. (2012). Objective guidance for use in setting tropical cyclone conditions of readiness. *Weather and Forecasting*, *27*(*4*), 1052–1060. http://doi.org/10.1175/WAF-D-12-00008.1

Schreck, C., Knapp, K., & Kossin, J. (2014). The impact of best track discrepancies on global tropical cyclone climatologies using IBTrACS. *Monthly Weather Review*, *142*(10), 3881–3899. https://doi.org/10.1175/MWR-D-14-00021.1

Song, T., Li, Y., Meng, F., Xie, P., & Xu, D. (2022). A novel deep learning model by BiGRU with attention mechanism for tropical cyclone track prediction in the Northwest Pacific. *Journal of Applied Meteorology and Climatology*, *61*(1), 3–12. https://doi.org/10.1175/JAMC-D-20-0291.1

Sun, J., Wang, G., Zuo, J., Ling, Z., & Liu, D. (2017). Role of surface warming in the northward shift of tropical cyclone tracks over the South China Sea in November. *Acta Oceanologica Sinica*, *36*(5), 67–72. https://doi.org/10.1007/s13131-017-1061-8

Tallapragada, V., Kieu, C., Trahan, S., Liu, Q., Wang, W., Zhang, Z., Tong, M., Zhang, B., Zhu, L., & Strahl, B. (2016). Forecasting tropical cyclones in the Western North Pacific basin using the NCEP operational HWRF model: Model upgrades and evaluation of real-time performance in 2013. *Weather and Forecasting*, *31*(3), 877–894. https://doi.org/10.1175/WAF-D-14-00139.1

Tan, Z., Lei, L., Wang, Y., Xu, Y., & Zhang, Y. (2022). Typhoon track, intensity, and structure: From theory to prediction. *Advances in Atmospheric Sciences*, *39*(11), 1789–1799. https://doi.org/10.1007/s00376-022-2212-1

Tong, B., Wang, X., Fu, J., Chan, P., & He, Y. (2022). Short-term prediction of the intensity and track of tropical cyclone via ConvLSTM model. *Journal of Wind Engineering and Industrial Aerodynamics*, *226*, 105026. https://doi.org/10.1016/j.jweia.2022.105026

Tsai, Y., Bai, S., Yamada, M., Morency, L., & Salakhutdinov, R. (2019). Transformer dissection: a unified understanding of transformer's attention via the lens of kernel. *arXiv preprint*, https://doi.org/10.48550/arXiv.1908.11775

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, https://doi.org/10.48550/arXiv.1706.03762

Wang, C., & Li, X. (2023). Deep learning in extracting tropical cyclone intensity and wind radius information from satellite infrared images – A review. *Atmospheric and Oceanic Science Letters*, *16*(4), 100373–100371. https://doi.org/10.1016/j.aosl.2023.100373

Wang, L., Zhou, Y., Lei, X., Zhou, Y., Bi, H., & Mao, X. (2020). Predominant factors of disaster caused by tropical cyclones in South China coast and implications for early warning systems. *Science of the Total Environment*, *726*, 138556. https://doi.org/10.1016/j.scitotenv.2020.138556

Woo, S., Park, J., Lee, J., & Kweon, I. (2018). CBAM: Convolutional block attention module. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, & Yair Weiss (Eds.), *Computer Vision – ECCV 2018, 11211* (pp. 3–19). Springer.

Yang, S., Hu, F., & Jaeger, C. (2016). Impact factors and risk analysis of tropical cyclones on a highway network. *Risk Analysis*, *36*(2), 262–277. https://doi.org/10.1111/risa.12463

Yenduri, G., Ramalingam, M., Selvi, G., Supriya, Y., Srivastava, G., Maddikunta, P., Raj, G., Jhaveri, R., Prabadevi, B., Wang, W., Vasilakos, A., & Gadekallu, T. (2024). GPT (generative pre-trained Transformer) – a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access*, *12*, 54608–54649. https://doi.org/10.1109/ACCESS.2024.3389497

Ying, M., Zhang, W., Yu, H., Lu, X., Feng, J., Fan, Y., Zhu, Y., & Chen, D. (2014). An overview of the China meteorological administration tropical cyclone database. *Journal of Atmospheric and Oceanic Technology*, *31*(2), 287–301. https://doi.org/10.1175/JTECH-D-12-00119.1

Zeng, D., Zhang, H., Li, Q., & Ellingwood, B. R. (2021). Tropical cyclone damage assessment of distributed infrastructure systems under spatially correlated wind speeds. *Structural Safety*, *91*, 102080. http://doi.org/10.1016/j.strusafe.2021.102080

Zhao, B., Wu, L., Wang, G., Zhang, J., Liu, L., Zhao, C., Zhuang, Z., Xia, C., Xue, Y., Li, X., & Qiao, F. (2024). A numerical study of tropical cyclone and ocean responses to air-sea momentum flux at high winds. *Journal of Geophysical Research: Oceans*, *129*(7), e2024JC020956. https://doi.org/10.1029/2024JC020956