Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Physics-Informed Time-Series Large Language Model for Extreme Typhoon Wave Height Forecasting

Zhiguo He

hezhiguo@zju.edu.cn

Zhejiang University

Yiting Zhao

Zhejiang University

Dongrui Han

Zhejiang University

Hui Shen

Marine Monitoring and Forecasting Center of Zhejiang Province

Ye Zhu

Marine Monitoring and Forecasting Center of Zhejiang Province

Xiangbing Kong

Université du Québec à Rimouski

Li Li

Zhejiang University

Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11525261/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Physics-Informed Time-Series Large Language Model for Extreme Typhoon Wave Height Forecasting

Zhiguo He[1,2,*], Yiting Zhao[1,2], Dongrui Han[1], Hui Shen[3], Ye Zhu[3], Xiangbing Kong[4], and Li Li[1]

[1]State Key Laboratory of Ocean Sensing & Ocean College, Zhejiang University, Zhoushan, 316021, PR China.

[2]Hainan Institute of Zhejiang University, Sanya, 572000, PR China.

[3]Marine Monitoring and Forecasting Center of Zhejiang Province, Hangzhou, 310007, PR China.

[4]Department of Engineering, University of Quebec at Rimouski, Rimouski, G5L 3A1, Canada.

***Corresponding author: Zhiguo He (hezhiguo@zju.edu.cn).**

Abstract

Typhoon-induced extreme waves pose severe threats to coastal communities, infrastructure, and maritime operations worldwide. Accurate and rapid forecasting of significant wave height is therefore critical for disaster response. Physics-based numerical wave models provide physically grounded predictions but can be computationally demanding for high-resolution forecasting, while purely data-driven models often struggle with rare, highly non-stationary extreme events. Recent time-series approaches based on large language models (LLMs) offer new opportunities for long-range forecasting but remain limited in explicitly representing coupled environmental variables and enforcing physically plausible predictions. Here, we develop the Typhoon Wave Large Language Model (TW-LLM), a physics-informed, domain-adapted framework for forecasting typhoon-induced coastal wave height. Building on Time-LLM, TW-LLM introduces cross-variable covariate interaction to explicitly represent dependencies between wave states and environmental forcing, structured domain-informed prompting for semantic conditioning, and a physics-informed training objective that combines magnitude-weighted Huber loss with wind-forcing consistency and temporal-rate constraints. Trained and evaluated using nine years (2017–2025) of hourly observations from eleven coastal buoys in China, TW-LLM achieves root-mean-square errors (RMSEs) of ~0.15 m for short-term (6-h) forecasts, 0.3 m for long-term (96-h) forecasts, and 0.23 m under extreme typhoon conditions. With an inference time as low as 5 s per forecast, TW-LLM provides a computationally efficient framework for near-real-time coastal wave forecasting and early warning, with potential to enhance coastal resilience to increasingly severe climate-related hazards.

1 Introduction

Tropical cyclones generate some of the most destructive coastal hazards worldwide, with storm surges and extreme waves causing substantial economic losses, infrastructure damage, and disruptions to maritime activities [1]. Observations and large-scale hindcasts reveal pronounced spatial and temporal variability in extreme significant wave height (H_{sig}) during tropical cyclones, particularly in nearshore and estuarine environments [2, 3]. As climate change, sea-level rise, and growing coastal exposure amplify tropical-cyclone-related risks [4, 5], accurate and timely forecasting of H_{sig} is increasingly critical for disaster preparedness and emergency response. Yet extreme-wave forecasting remains challenging because typhoon-induced waves are highly non-stationary and governed by complex interactions among atmospheric forcing, ocean conditions, and local coastal environments.

Traditional numerical wave models face a persistent trade-off between physical accuracy and computational efficiency. Spectral wave models such as Simulating Waves Nearshore (SWAN) [6] explicitly resolve key wave-generation, propagation, and dissipation processes, but high-resolution coastal applications remain computationally demanding. High-resolution SWAN simulations, for example, can require substantial wall-clock time for historical tropical-cyclone events [7], while operational nested

forecasting systems may require several hours for individual high-resolution estuarine simulations [8]. Large-scale coastal circulation and wave-model configurations can achieve substantial computational acceleration through high-performance computing [9], yet coupled wave forecasts with local refinement and ensemble uncertainty assessment still consume considerable computational resources [10, 11, 12]. Moreover, forecast accuracy during tropical cyclones depends strongly on uncertain atmospheric forcing, boundary conditions, and parameterization of highly non-stationary wave processes [13]. These limitations constrain rapid forecasting and repeated scenario evaluation during evolving extreme events.

Data-driven approaches provide a computationally efficient alternative, but their ability to predict rare and rapidly evolving extremes remains limited. Early machine-learning methods, including support vector regression (SVR) and random forests (RF), can approximate nonlinear relationships between environmental forcing and wave conditions [14, 15], but often lose accuracy outside the range well represented in their training data, particularly during extreme-wave events when H_{sig} exceeds 4 m [16]. Deep learning architectures have substantially advanced time-series forecasting by learning complex temporal and spatial patterns directly from observations [17]. Long short-term memory (LSTM), convolutional neural network (CNN), and hybrid CNN–LSTM models have demonstrated improved wave predictions across buoy and regional datasets [18, 19, 20, 21, 22]. Physics-informed neural operators (PINOs) and related hybrid approaches further seek to improve both robustness and physical consistency by incorporating governing relationships or physical constraints into data-driven learning [23, 24, 25, 26]. Nevertheless, extreme typhoon waves pose a particular challenge: they are sparsely represented in historical observations, strongly non-stationary, and driven by coupled interactions among atmospheric forcing, antecedent wave conditions, and local environments. Consequently, conventional deep-learning models can remain sensitive to domain-specific tuning and exhibit limited generalization under data-sparse extreme conditions [27, 28].

Recent advances in large language models (LLMs) and foundation models for time-series forecasting provide a new opportunity to address these limitations. Frameworks such as Time-LLM reprogram pretrained language models for numerical time series through patch-based representations and cross-modal alignment, exploiting their ability to capture long-range dependencies and transferable temporal patterns [29]. Subsequent developments in patching, representation learning, trend decomposition, and LLM adaptation have further improved forecasting performance across diverse time-series tasks [30, 31, 32, 33]. However, directly adapting general-purpose LLM architectures to extreme ocean forecasting presents two fundamental limitations. First, multivariate ocean processes are physically coupled, whereas generic time-series architectures may learn cross-variable dependencies only implicitly or process channels largely independently. Second, statistical sequence learning alone does not guarantee physically consistent predictions, particularly when extrapolating to rare, rapidly evolving extremes. These limitations are especially consequential for typhoon-induced waves, for which significant wave height (H_{sig}) responds strongly to evolving wind forcing while remaining constrained by physically plausible rates of growth and decay.

To address these challenges, we develop the Typhoon Wave Large Language Model (TW-LLM), a physics-informed, domain-adapted framework for rapid forecasting of extreme typhoon-induced significant wave height (H_{sig}). Building on Time-LLM, TW-LLM introduces three complementary mechanisms tailored to extreme wave dynamics: (1) structural cross-variable attention to explicitly represent interactions between wave states and environmental forcing; (2) domain-informed semantic prompting to provide contextual knowledge of wind–wave processes; and (3) a physics-informed training objective combining magnitude-aware learning with wind-forcing consistency and kinematic constraints. Together, these mechanisms target three major challenges in extreme-wave forecasting—multivariate physical coupling, sparse extreme-event observations, and highly non-stationary wave evolution.

TW-LLM is evaluated using nine years (2017–2025) of hourly observations from eleven coastal buoys exposed to typhoon events. Its performance is assessed across short- and extended-lead forecasts and specifically under extreme typhoon conditions, with comparisons against numerical and data-driven baselines. Model interpretability is further examined using Shapley additive explanations (SHAP) to quantify the contributions of environmental covariates and physically relevant predictors. By combining the long-range representation capability of LLMs with explicit multivariate interactions and physics-informed learning, this study seeks to establish a computationally efficient and physically consistent framework for near-real-time forecasting of extreme typhoon waves.

2 Methodology

2.1 Framework Overview and Input Representation

To bridge the representation gap between continuous multivariate ocean observations and pretrained language models, we develop the Typhoon Wave Large Language Model (TW-LLM), a parameter-efficient forecasting framework built upon Time-LLM [29]. TW-LLM combines temporal patching, explicit cross-variable interaction, cross-modal reprogramming, domain-informed prompting, and physics-informed optimization within a unified architecture. In particular, the framework is designed to address three challenges in extreme typhoon-wave forecasting: coupled dependencies among met-ocean variables, the limited representation of extreme wave conditions in historical observations, and the highly non-stationary evolution of significant wave height.

The framework processes multivariate met-ocean time series consisting of significant wave height H_{sig} and associated environmental covariates, including wind speed, wave direction, and wind direction. At each time step t, the multivariate input is represented as:

$$\mathbf{x}_t \in \mathbb{R}^d, \tag{1}$$

where d denotes the number of variables, including the target variable and environmental covariates. The complete historical input sequence is:

$$\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T} \in \mathbb{R}^{T \times d}, \tag{2}$$

where T is the sequence length.

To mitigate scale heterogeneity across variables, normalization is first applied:

$$\widehat{\mathbf{X}} = \mathrm{Norm}(\mathbf{X}), \tag{3}$$

where the mean and standard deviation used for normalization are estimated from the training set for each variable. The normalized sequence is segmented into overlapping temporal patches of length P (with a predefined stride S):

$$\widehat{\mathbf{X}} \rightarrow \{\mathbf{X}^{(l)}\}_{l=1}^{L},\ \mathbf{X}^{(l)} \in \mathbb{R}^{P \times d}, \tag{4}$$

where $L = \lfloor (T-P)/S \rfloor + 1$ denotes the total number of temporal patches, determined by the patch length and stride.

Under the channel-independence (CI) design, each variable is embedded independently into a latent space of dimension D (the model dimension, a hyperparameter of the framework) via patch embedding layers:

$$\mathbf{Z} \in \mathbb{R}^{B \times N \times L \times D}, \tag{5}$$

where B is the batch size, $N = d$ is the number of variables (including H_{sig} and covariates), and L is the number of temporal patches. This patch-based channel-independent formulation follows the patching design of PatchTST [34] and its implementation in Time-LLM [29].

As illustrated in Figure 1, TW-LLM comprises four functional stages connected by two complementary information pathways: a Data Flow pathway and a Physical-Constraint pathway. First, multivariate met-ocean observations are normalized, segmented into temporal patches, and projected into latent representations. A cross-variable covariate interaction module then explicitly captures dependencies among wave and atmospheric variables. Second, the resulting numerical representations are reprogrammed into the embedding space of a frozen pretrained language model and combined with structured domain-informed prompts. Third, the pretrained Transformer backbone extracts higher-level representations from the combined numerical and semantic information. Finally, a forecasting head maps these representations to the future wave-height sequence. During training, an extreme-aware prediction loss emphasizes high-wave conditions, while physics-informed regularization promotes physically plausible temporal evolution. Thus, the domain-informed prompt and numerical physical regularization serve complementary but distinct functions. The former provides semantic context concerning wind–wave processes during representation learning, whereas the latter directly constrains model optimization through the predicted wave-height sequence.

(figure 1)

2.2 Cross-Variable Covariate Interaction

The original channel-independent formulation is effective for preserving variable-specific temporal characteristics but does not explicitly represent interactions among different physical variables. This limitation is particularly relevant to typhoon-wave forecasting because wave evolution reflects coupled influences from atmospheric forcing and antecedent sea states. We therefore introduce a cross-variable covariate interaction module before the numerical reprogramming stage. For each temporal patch $l \in \{1,\ldots,L\}$, the latent representations of all variables are collected as:

$$\mathbf{Z}_l \in \mathbb{R}^{N\times D}. \tag{6}$$

Following the scaled dot-product attention formulation [35], a multi-head self-attention layer is applied across the variable dimension:

$$\widetilde{\mathbf{Z}}_l = \mathrm{MultiHeadAttention}(\mathbf{Z}_l,\mathbf{Z}_l,\mathbf{Z}_l), \tag{7}$$

where each variable attends to all others within the same temporal patch. For the i-th variable, its updated representation is:

$$\tilde{\mathbf{z}}_{l,i} = \sum_{j=1}^{N} \alpha_{ij}^{(l)}\, \mathbf{z}_{l,j}, \tag{8}$$

where $\alpha_{ij}^{(l)}$ denotes the attention weight from variable j to variable i at patch l, and $\mathbf{z}_{l,j}$ is the latent representation of variable j. This formulation enables the model to dynamically integrate information from physically related covariates, such as wind speed contributing to wave growth. Here, the learnable query, key, and value projection matrices map each variable representation into the corresponding query, key, and value spaces, respectively, and the softmax operation normalizes the attention weights across all variables within each temporal patch. The updated representations have the following form:

$$\widetilde{\mathbf{Z}} \in \mathbb{R}^{B\times N\times L\times D}. \tag{9}$$

Before reprogramming, the batch and variable dimensions are combined to obtain:

$$\mathbf{Z}_{\mathrm{patch}} = \mathrm{Reshape}(\widetilde{\mathbf{Z}}) \in \mathbb{R}^{BN\times L\times D}, \tag{10}$$

where BN represents the total number of variable-specific sequences processed independently by the subsequent reprogramming layer. This operation preserves the variable-specific sequence structure required by the subsequent Time-LLM reprogramming procedure [29] while allowing information exchange among variables beforehand. TW-LLM therefore retains the computational advantages of

channel-independent temporal patching but relaxes the assumption that environmental variables evolve independently.

2.3 Cross-Modal Reprogramming and Domain-Informed Prompt Conditioning

Prompts are constructed to provide high-level contextual guidance and domain knowledge. Rather than encoding explicit differential equations or rigid numerical constraints, $\mathrm{T}_{\mathrm{phys}}$ represents the textual encoding of qualitative physical knowledge, including wind-driven wave energy input, coastal shoaling, and wave dissipation. These physical descriptions serve as semantic priors that provide contextual guidance to the frozen language model, while numerical physical consistency is further encouraged during training through the physics-informed regularization described in Section 2.5.

Following Time-LLM [29], the structurally refined numerical representations are reprogrammed into the embedding space of the pretrained language model. The input embedding matrix of the frozen language model is defined as:

$$\mathbf{E}_{\mathrm{vocab}} \in \mathbb{R}^{V \times D_{\mathrm{llm}}}, \tag{11}$$

where V is the vocabulary size and D_{llm} is the dimension of the language-model embedding space. A learnable mapping layer is used to obtain a compact set of K source embeddings:

$$\mathbf{E}_{\mathrm{src}} = \mathbf{M}\mathbf{E}_{\mathrm{vocab}} \in \mathbb{R}^{K \times D_{\mathrm{llm}}}, \tag{12}$$

where $\mathbf{M} \in \mathbb{R}^{K \times V}$ is the learnable mapping matrix and K denotes the number of source embeddings used for reprogramming. The pretrained vocabulary embedding matrix remains frozen, while the mapping layer provides a compact trainable interface between the numerical time-series representation and the language-model embedding space.

The reshaped temporal patch representations $\mathbf{Z}_{\mathrm{patch}} \in \mathbb{R}^{BN \times L \times D}$ are then aligned with the source embeddings through a cross-attention-based reprogramming layer:

$$\mathbf{Z}_{\mathrm{repro}} = \mathrm{Reprogramming}\left(\mathbf{Z}_{\mathrm{patch}}, \mathbf{E}_{\mathrm{src}}, \mathbf{E}_{\mathrm{src}}\right) \in \mathbb{R}^{BN \times L \times D_{\mathrm{llm}}}, \tag{13}$$

where the temporal patch representations serve as queries, while the source embeddings serve as both keys and values. This operation transforms the continuous numerical features into representations compatible with the embedding space of the pretrained language model without discretizing the numerical observations into vocabulary tokens.

In parallel, the contextual and physical prompt components are concatenated into a textual prefix:

$$\mathrm{T}_{\mathrm{prompt}} = [\mathrm{T}_{\mathrm{cont}}; \mathrm{T}_{\mathrm{phys}}], \tag{14}$$

where $\mathrm{T}_{\mathrm{cont}}$ contains dataset, environmental, statistical, and event-related information, while $\mathrm{T}_{\mathrm{phys}}$ provides qualitative descriptions of relevant wave-generation, propagation, shoaling, and dissipation processes. The resulting text sequence is tokenized and processed by the frozen language-model embedding layer to obtain:

$$\mathbf{E}_{\mathrm{prompt}} = \mathrm{Embed}\left(\mathrm{T}_{\mathrm{prompt}}\right) \in \mathbb{R}^{BN \times L_{\mathrm{prompt}} \times D_{\mathrm{llm}}}, \tag{15}$$

where L_{prompt} denotes the number of prompt tokens. Importantly, the prompt embeddings are not used as queries, keys, or values in the numerical reprogramming operation. Instead, they are introduced separately as semantic prefix information after the numerical features have been transformed into the language-model embedding space. This separation allows domain knowledge to provide high-level contextual guidance while leaving numerical physical consistency to the physics-informed regularization during model optimization.

A systematic prompt design is employed as a prefix layer alongside the reprogrammed time-series patches. We use a structured prompt with six complementary components:

Setting: Role: Expert AI model for coastal typhoon wave time-series forecasting.

Context: Marine Environment: coastal met-ocean observations; Typhoon Scenario: Non-stationary wave evolution under typhoon forcing

Statistical Descriptors: The input sequence has a minimum of <min_val>, a maximum of <max_val>, and a median of <median_val>. The overall trend is <upward/downward>. The dominant temporal lags are <lag_val>.

Domain Guidance: Wind forcing contributes to wave growth; energy dissipation contributes to wave decay; local coastal conditions can modify the wave response.

Directive: Task: Predict the next <Horizon> hours of the H_sig sequence.

Requirement: Maintain temporal continuity and physically plausible wave evolution.

2.4 LLM Input Formation and Forecast Generation

Because the reprogrammed numerical features already reside in the embedding space of the pretrained language model, no additional tokenization or vocabulary-level discretization is required. The numerical representations retain the variable-specific sequence form:

$$\mathbf{E}_{\mathrm{ts}}=\mathbf{Z}_{\mathrm{repro}}\in\mathbb{R}^{BN\times L\times D_{\mathrm{llm}}}. \tag{16}$$

The structured prompt embeddings are prepended to the reprogrammed temporal representations along the sequence dimension:

$$\mathbf{E}_{\mathrm{input}}=[\mathbf{E}_{\mathrm{prompt}};\mathbf{E}_{\mathrm{ts}}]\in\mathbb{R}^{BN\times(L_{\mathrm{prompt}}+L)\times D_{\mathrm{llm}}}. \tag{17}$$

The resulting sequence is supplied directly to the frozen pretrained Transformer through its input-embedding interface. The pretrained LLM therefore receives two complementary forms of information: numerical temporal representations obtained through cross-modal reprogramming and domain-specific contextual information provided through structured prompts. The resulting hidden representations are passed to a trainable forecasting head to generate the predicted significant wave-height sequence.

2.5 Physics-Informed Training and Optimization

TW-LLM is trained using a composite objective designed to address both statistical accuracy and physically plausible extreme-wave evolution. The structural covariate interaction and reprogramming modules transform the multivariate met-ocean observations into representations suitable for the frozen language model, while the training objective directly constrains the predicted significant wave height sequence. Specifically, the optimization objective consists of a magnitude-weighted Huber loss and a physics-informed regularization term. The former improves robustness and emphasizes high-wave-height events, whereas the latter constrains the temporal evolution of the predicted wave height according to wind forcing and limits excessively rapid variations. Thus, the physical information introduced through prompting acts as a semantic prior, while the physics-informed loss provides an independent numerical constraint during parameter optimization.

$\widehat{\mathbf{Y}}\in\mathbb{R}^{B\times H}$ denotes the predicted significant wave-height sequences and $\mathbf{Y}\in\mathbb{R}^{B\times H}$ the corresponding observations, where B is the batch size and H is the forecasting horizon. The training objective is defined as

$$L=L_{\mathrm{pred}}+\lambda_{\mathrm{phys}}L_{\mathrm{phys}}, \tag{18}$$

where L_{pred} is the magnitude-weighted Huber prediction loss, L_{phys} is the physics-informed regularization term, and λ_{phys} controls the relative contribution of the physical regularization. In the implementation, the predicted and observed wave heights are transformed back to their physical units before the physics-informed loss is evaluated, ensuring that the thresholds and wind-forcing coefficients are interpreted in physically meaningful units. This inverse-standardization procedure is explicitly implemented before the loss calculation in the training loop.

2.5.1 Extreme-Aware Magnitude-Weighted Huber Prediction Loss

Extreme typhoon-wave observations are substantially less frequent than observations of normal sea states but are disproportionately important for coastal hazard forecasting. A conventional mean loss can therefore be dominated by the much more numerous moderate-wave observations. To obtain robust predictions across both normal and extreme sea states, we employ the Huber loss as the basic prediction objective. For an individual prediction element, the Huber loss is defined as [36]

$$L_{\text{Huber}}(y_i,\hat{y}_i)=\begin{cases}\frac{1}{2}(y_i-\hat{y}_i)^2, & |y_i-\hat{y}_i|\le\delta\\ \delta|y_i-\hat{y}_i|-\frac{1}{2}\delta^2, & |y_i-\hat{y}_i|>\delta\end{cases}, \tag{19}$$

where y and $\hat{y}$ denote the observed and predicted significant wave heights, respectively, and δ is the transition threshold between the quadratic and linear regimes. In this study, δ=1.0 m is used, following common practice in robust regression [36]. The implementation follows this piecewise formulation directly.

Because high-wave-height observations are less frequent but are particularly important for typhoon-related wave forecasting, a magnitude-dependent weighting mechanism is further introduced. The prediction loss is given by

$$L_{\text{pred}}=\frac{1}{BH}\sum_{b=1}^{B}\sum_{i=1}^{H}W_{b,i}\,\text{Huber}(y_{b,i},\hat{y}_{b,i}), \tag{20}$$

where $W_{b,i}$ is a sample-dependent weight determined by the observed wave height:

$$W_{b,i}=1+\kappa_w\phi(y_{b,i}), \tag{21}$$

with

$$\phi(y_{b,i})=\min\left[\max(y_{b,i}-\eta,0),c\right]. \tag{22}$$

Here, η denotes the wave-height threshold above which additional weighting is activated, c limits the maximum value of the additional weighting component $\phi(\cdot)$, and κ_w controls the strength of the emphasis on high-wave-height observations. The current implementation uses η=1.5 m, c=2.0 m, and κ_w=1.0 as the default values. The bounded form of $\phi(\cdot)$ prevents a small number of extreme observations from receiving excessively large weights. These parameters are explicitly defined in the training configuration and passed to the loss function.

This formulation retains the robustness of the Huber objective while increasing the contribution of relatively high wave-height observations to parameter optimization. Therefore, the model does not simply minimize the average error across all sea states but gives additional attention to the upper range of wave-height observations that are more consequential for extreme-event forecasting.

2.5.2 Physics-Informed Regularization

To further constrain the temporal behavior of the predicted significant wave height, we introduce a soft physics-informed regularization term based on two physically motivated terms: wind-forcing consistency and a kinematic rate constraint. Rather than imposing the complete wave-action balance equation, the

proposed regularization uses simplified physical proxies that can be evaluated directly from the available wind and wave observations.

The physics-informed regularization is defined as

$$L_{\mathrm{phys}}=L_{\mathrm{wind}}+L_{\mathrm{grad}}, \tag{23}$$

where L_{wind} constrains the predicted wave-growth tendency according to wind speed and directional alignment, while L_{grad} penalizes excessively large temporal changes in the predicted wave height.

Because the observations contain significant wave height, wave direction, wind speed, and wind direction, the wind-forcing term incorporates both wind magnitude and the directional relationship between wind and wave propagation. The temporal change in the predicted wave height is approximated using a first-order finite difference:

$$\Delta_t\hat{y}_{b,i}=\hat{y}_{b,i+1}-\hat{y}_{b,i},\ i=1,\ldots,H-1. \tag{24}$$

For hourly observations, $\Delta_t\hat{y}$ has units of m/h and represents the predicted wave-height growth or decay rate between two consecutive forecasting steps.

The directional alignment factor is defined as

$$A_{b,i}=\left(\max\left[0,\cos\left(\theta^{w}_{b,i+1}-\theta^{\mathrm{wave}}_{b,i+1}\right)\right]\right)^2, \tag{25}$$

where θ^{w} and θ^{wave} denote wind direction and wave direction, respectively. The angular difference is converted from degrees to radians before evaluating the cosine function. This implementation is directly reflected in the loss function, where the directional alignment is calculated from the wind and wave directions at step i+1.

A simplified wind-supported growth-rate upper-bound proxy is then formulated as

$$G_{b,i}=\kappa_{\mathrm{energy}}U^{4}_{10,b,i+1}A_{b,i}, \tag{26}$$

where U_{10} denotes the wind speed at a height of 10 m at step i+1 and κ_{energy} is a scaling coefficient. The coefficient κ_{energy} is introduced here as an empirical scaling coefficient that converts the wind-speed-based energy-input proxy into a wave-height growth-rate scale. Its functional form is motivated by established wind-wave input formulations, including the wind-energy input parameterization of Cavaleri and Malanotte-Rizzoli [37], which relates wave generation to wind forcing and wind-wave directional alignment [37]. Because the present term is used as a simplified regularization proxy rather than as a direct implementation of the original wave-energy source term, κ_{energy} is treated as a tunable coefficient. In the current implementation, $\kappa_{\mathrm{energy}}=2.0\times10^{-5}$ is adopted.

The wind-forcing consistency loss is calculated as the mean squared violation when the predicted wave-height growth rate exceeds this wind-energy upper bound:

$$L_{\mathrm{wind}}=\frac{1}{B(H-1)}\sum_{b=1}^{B}\sum_{i=1}^{H-1}\left[\mathrm{ReLU}\left(\Delta_t\hat{y}_{b,i}-G_{b,i}\right)\right]^2. \tag{27}$$

This term should be interpreted as a first-order physical proxy rather than an exact wave-energy balance equation. It encourages the predicted wave-height evolution to remain broadly compatible with stronger wind forcing and favorable wind-wave directional alignment by penalizing wave growth rates that exceed the wind-supported capacity, without claiming that wave growth is completely determined by wind speed alone. The corresponding implementation calculates the squared ReLU violation between the finite-difference wave-height change and the wind-energy upper-bound proxy.

In addition to wind-forcing consistency, the model is discouraged from producing unrealistically abrupt temporal variations. The kinematic regularization is defined as:

$$L_{\text{grad}}=\frac{1}{B(H-1)}\sum_{b=1}^{B}\sum_{i=1}^{H-1}\left[\text{ReLU}\left(|\Delta_i\hat{y}_{b,i}|-\delta_k\right)\right]^2, \tag{28}$$

where δ_k represents the maximum tolerated wave-height change per hour. In accordance with ocean wave dynamics under storm and typhoon conditions [38], the kinematic threshold δ_k typically ranges from 0.06 to 0.10 m/h; the current implementation adopts δ_k=0.08 m/h as the default setting. Therefore, temporal changes below this threshold are not penalized by this term, whereas changes exceeding the threshold incur an increasing quadratic penalty, thereby suppressing non-physical numerical spikes. This formulation is implemented through a ReLU operation applied to the absolute first difference of the predicted wave-height sequence.

2.5.3 Composite Objective, Parameter Updating, and Inference

Combining the magnitude-weighted prediction loss and physics-informed regularization yields the final optimization objective:

$$L=L_{\text{pred}}+\lambda_{\text{phys}}\left(L_{\text{wind}}+L_{\text{grad}}\right), \tag{29}$$

where λ_{phys} controls the overall strength of the physics-informed regularization. With the default setting λ_{phys}=0.01, the physical terms act as a soft auxiliary objective rather than replacing the primary forecasting objective.

During training, the predicted and observed wave heights, together with the wind speed, wind direction, and wave direction, are converted from their standardized values back to physical units before the composite loss is evaluated. The resulting loss is then backpropagated through the trainable components of TW-LLM. The pretrained language-model parameters remain frozen, whereas the trainable numerical representation, cross-variable interaction, reprogramming, and forecasting components are updated according to the combined prediction and physical objectives. The implementation explicitly constructs the Physics-Informed-Wave-Loss, computes the prediction and physical loss components, and performs parameter updates using the resulting total loss.

During inference, the physics-informed loss is not applied because ground-truth future wave heights are unavailable. The trained model directly generates the future significant wave-height sequence, which is subsequently transformed back to physical units for output.

3 Data and Experimental Setup

3.1 Observational Dataset

TW-LLM is evaluated using a multi-year observational dataset collected from eleven coastal wave buoys along the Zhejiang coast, China. Although the model is developed as a general framework for extreme typhoon-wave forecasting, the Zhejiang buoy network provides a challenging regional testbed because the area is frequently affected by western North Pacific tropical cyclones and encompasses diverse nearshore and offshore environments.

The dataset spans from January 2017 to December 2025 and contains approximately 260,000 hourly records. The observations cover both routine sea states and episodic typhoon-induced extremes, thereby providing a strongly non-stationary dataset for evaluating forecasting performance across a broad range of wave conditions [39]. The primary prediction target is significant wave height, with observed values ranging approximately from 0.1 m to 7.2 m. Available environmental covariates include wind speed (0–25 m/s), wind direction, barometric pressure, and wave period. The exact subset of covariates used by each model is kept consistent when comparisons require identical input information.

The eleven stations span different coastal and bathymetric settings, ranging from relatively shallow nearshore waters (depth < 20 m) to more exposed offshore locations (depth > 50 m). Such spatial

heterogeneity provides an opportunity to evaluate model performance under different local wave climates rather than at a single station. Table 1 summarizes the location, sample size, and basic wave height statistics for each buoy.

Missing observations account for less than 5% of the dataset. Short isolated gaps are reconstructed by linear interpolation, whereas longer gaps are excluded from sequence construction to avoid introducing artificial temporal evolution across extended periods without observations. Importantly, interpolation is performed independently within each temporal subset after data splitting, thereby preventing information from the validation or test periods from propagating into the training data.

Normalization parameters are estimated exclusively from the training subset. Reversible Instance Normalization (ReVIN) is additionally employed to reduce distributional shifts among temporal windows while retaining the ability to recover predictions in physical units. All reported physical error metrics are calculated after inverse transformation.

For the three stations with long continuous records, i.e., Zhujiajian, Wenzhou (Dongtou), and Dachen, the observations are divided chronologically into training, validation, and test periods using a 70:20:10 ratio. Chronological splitting is adopted rather than random sampling to ensure that future observations do not leak into model training. For the eight stations with short records (8,976 hourly samples each), temporally blocked cross-validation is used. Each fold preserves temporal ordering between the training and evaluation periods, thereby providing a more realistic assessment of out-of-sample forecasting performance than random K-fold partitioning.

Two principal forecasting horizons are considered: a short-range 6-h forecast and an extended 96-h forecast, corresponding to four days for the hourly observations. Performance is evaluated both over the complete test datasets and over high-wave subsets representing extreme conditions. Because extreme H_{sig} values occur much less frequently than ordinary sea states, the latter evaluation is used to assess whether improvements in overall accuracy persist under the conditions most relevant to coastal hazards.

The 2017–2025 dataset contains 12 historical typhoon-affected periods, including events producing observed significant wave-height values exceeding 4–6 m, such as Typhoon Gaemi, at exposed stations. These observations are highly imbalanced: high-wave conditions constitute only a small fraction of the complete dataset. The resulting combination of strong temporal non-stationarity, event rarity, spatial heterogeneity, and rapid wave growth and decay provides a stringent test for data-driven forecasting models.

Figure 2 shows the trajectories of tropical cyclones affecting the study region during 2017–2025 based on the China Meteorological Administration (CMA) Best-Track dataset, together with the eleven buoy locations. Figure 3 presents representative hourly wave height time series from eight selected buoys with relatively continuous observations, illustrating the contrast between routine sea states and episodic typhoon-induced peaks.

An additional 2026 dataset is used exclusively for independent temporal evaluation. These observations are completely withheld from model development and are not used for training, validation, hyperparameter selection, or estimation of normalization statistics. This independent period therefore provides an additional test of model transfer to previously unseen events beyond the 2017–2025 development dataset.

(figure 2)

(table 1)

(figure 3)

3.2 Baselines, Ablation Design, and Evaluation Metrics

The Time-LLM framework [29], from which TW-LLM is derived, serves as the representative data-driven baseline. Where temporally and spatially matched simulations are available, predictions are additionally compared with physics-based numerical simulations from the SWAN model [6]. Numerical outputs are interpolated to the corresponding buoy locations and hourly timestamps. Because numerical and data-driven forecasting systems differ substantially in their forcing requirements and computational configuration, these comparisons are interpreted as operational reference benchmarks rather than strictly equivalent model-to-model experiments. For fair comparison among all data-driven models, identical training, validation, and test periods and, where supported by the architecture, identical input information are used. Hyperparameters are selected using only the training and validation data. No test observations are used for model selection. Each neural forecasting experiment is repeated using five independent random seeds, and performance is reported as the mean together with the corresponding variability across runs.

Ablation experiments are designed to isolate the contributions of the major TW-LLM components. Starting from the complete model, we separately remove (i) cross-variable covariate interaction, (ii) domain-informed prompting, and (iii) physics-informed regularization. Additional ablations separate the wind-consistency and temporal-rate components of the physical objective where appropriate. These experiments allow the contribution of representation coupling, semantic domain information, and numerical physical regularization to be evaluated independently.

Forecasting accuracy is primarily assessed in physical units using root-mean-square error (RMSE) and mean absolute error (MAE). Normalized mean squared error (MSE) and MAE are additionally reported where needed for comparison with previous time-series forecasting studies. Model robustness under extreme conditions is assessed using MAE computed on high-wave samples (above a predefined threshold). Pattern consistency is evaluated using Taylor diagrams [40].

4 Results

4.1 Overall Forecasting Performance

TW-LLM consistently improves significant-wave-height forecasts across both short- and long-term horizons. The improvement is modest under relatively regular sea states but becomes more pronounced for long-horizon and high-wave conditions, for which conventional data-driven models show greater error accumulation and peak attenuation.

For the 6-h forecasting task, TW-LLM achieves normalized MSE and MAE values of 0.023 and 0.104, respectively, corresponding to a physical RMSE of ~0.15 m. For the 96-h forecasting task, the corresponding normalized MSE and MAE are 0.09 and 0.133, respectively, with a physical RMSE of ~0.3 m. These results indicate that the model retains useful predictive skill as the forecasting horizon increases from several hours to four days.

Relative to the Time-LLM baseline, TW-LLM shows the largest improvements for extended-lead forecasts. In the comparison yielding the largest performance difference, RMSE is reduced by approximately 60%. Performance under extreme conditions is further evaluated on a subsample of the test set where the observed H_{sig} exceeds the 93rd percentile. In this regime, TW-LLM achieves a 40% reduction in RMSE to 0.23 m, suggesting enhanced robustness in capturing high-impact wave dynamics.

Figure 4 summarizes model performance using a Taylor diagram, where angular position denotes the Pearson correlation coefficient and radial distance represents normalized standard deviation, while the distance to the reference point corresponds to centered RMSE. TW-LLM consistently lies closest to the observed reference, with correlation coefficients exceeding 0.94, compared to approximately 0.77 for the

vanilla Time-LLM. In addition, the standard deviation ratio of TW-LLM remains close to 1.0, indicating that the model better reproduces the observed variability of wave heights. In contrast, SWAN tends to underestimate variability during peak events, while Time-LLM exhibits larger centered RMSE. These results suggest that the proposed framework provides a more balanced representation of both amplitude variability and temporal coherence under complex typhoon forcing.

(figure 4)

4.2 Event-Based Performance and Ablation Analysis

Figure 5 examines time series predictions during selected typhoon-affected periods at the Dachen and Nanji stations: Figure 5(a) and (c) show predictions during Typhoons FRANCISCO and CO-MAY in July 2025 from the held-out evaluation data, while Figure 5(b) and (d) show predictions for the independent 2026 period from August 26 to September 4, which includes the influence of Typhoon SAUDEL from August 26 to 28.

Across these events, TW-LLM reproduces the major phases of wave evolution, including pre-storm growth, peak development, and subsequent decay, and follows the magnitude and timing of the observed response more closely than the vanilla Time-LLM, which tends to produce smoother trajectories and attenuate extreme peaks.

The independent 2026 evaluation provides a more stringent test, as neither the observations from this year nor the corresponding events are included in model development. TW-LLM retains useful forecasting skill during this period, suggesting out-of-sample temporal generalization within the related regional wave climate rather than universal transferability across ocean basins.

(figure 5)

Figure 6 further evaluates prediction behavior through predicted-versus-observed scatter plots and residual distributions. Under non-typhoon conditions (Figure 6(b) and (d)), the fitted relationship is close to the 1:1 line with $R^2 = 0.99$, indicating minimal systematic bias over the dominant range of ordinary wave conditions. For the high-wave subset (Figure 6(a) and (c)), the model retains a high coefficient of determination (with $R^2 = 0.98$), but it also reveals a remaining amplitude-dependent bias. In particular, the slope below unity indicates that the largest observed wave heights tend to be underestimated, despite the substantial improvement over the baseline models. This residual peak attenuation reflects the difficulty of extrapolating to sparsely sampled extremes. The corresponding residual distributions (Figure 6(c) and (d)) remain approximately centered on zero, with non-typhoon residuals confined within roughly 0.4 m and a longer negative tail under typhoon extremes, consistent with the amplitude-dependent underestimation of the largest wave heights.

Ablation experiments further isolate the contribution of each component by comparing the complete model with variants that remove cross-variable covariate interaction (w/o Structural Cross-variable Attention, SCA+L_{phys}), domain-informed prompting (w/o Prompt), and physics-informed regularization (w/o L_{phys}), together with the original Time-LLM architecture. Experiments are performed under both short-term (6-h) and long-term (96-h) forecasting settings, and the reported results are averaged over five random seeds. For visualization, the error metrics are transformed using $-\log_{10}$, with higher values indicating better performance.

As shown in Figure 6(e) and 6(f), the complete TW-LLM consistently achieves the highest transformed scores across both forecasting horizons. Removing L_{phys} produces a pronounced degradation for short-term prediction but only a limited decline in long-term forecasting, indicating that the physical constraint is most effective at suppressing prediction errors at short lead times.

Removing the prompt mechanism or the cross-variable interaction module also leads to noticeable performance degradation, confirming the value of semantic physical priors and explicit cross-variable coupling. All TW-LLM variants improve substantially over the vanilla Time-LLM, indicating that adapting a general-purpose LLM without domain-specific design is insufficient for complex marine time-series forecasting.

(figure 6)

4.3 Model Interpretation and Generalization

KernelSHAP analysis [41] is applied to the complete forecasting pipeline to provide an input-output interpretation of TW-LLM predictions. The trained model is treated as a black-box predictor, and SHAP values quantify the change in model output associated with individual input variables relative to the selected background distribution. This analysis does not reveal the internal reasoning process of the frozen LLM; rather, it provides a post hoc attribution of the final forecast to the numerical model inputs.

Figure 7(a) presents the global SHAP analysis over the test dataset. The mean absolute SHAP values summarize the overall importance of each input variable, while the beeswarm distribution illustrates the variation and direction of feature contributions among individual samples. Historical wave height variables exhibit the largest contributions, indicating that temporal persistence plays a dominant role in short-term wave evolution. Meanwhile, meteorological and wave-state variables, such as wind speed and wave period, provide additional contributions, demonstrating that TW-LLM effectively incorporates external forcing information beyond autoregressive dependencies. The identified feature importance hierarchy agrees with established oceanographic understanding, where wave conditions are jointly controlled by previous sea states and wind-driven energy input. These results suggest that the proposed framework is able to utilize physically meaningful variables while maintaining the forecasting capability of the LLM-based architecture.

To further investigate individual prediction behavior during extreme events, a case study is conducted for a typhoon-induced wave peak at a representative coastal buoy station during Typhoon Gaemi (2024). As shown in Figure 7(b), the SHAP force plot shows how contributions from individual input variables shift the prediction from the expected prediction value (H_{sig}) to the final forecast. Historical wave height provides the dominant contribution, while wind-related and wave-related variables adjust the prediction according to the evolving environmental conditions.

Overall, KernelSHAP provides an interpretation of TW-LLM predictions at the input-output level. The consistency between feature contributions and known wave-generation processes supports the reliability and interpretability of the proposed framework for typhoon-induced wave forecasting.

(figure 7)

5 Discussion

The results demonstrate the potential of physics-guided large language models for typhoon-induced wave forecasting. By extending the Time-LLM framework with structural covariate interaction, physics-informed prompts, and physically constrained optimization, the proposed model improves the representation of nonlinear relationships between wave evolution and atmospheric forcing. The KernelSHAP analysis further provides an input-output interpretation of model predictions, showing that historical wave conditions dominate the forecasting process while meteorological variables provide additional information consistent with wave-generation mechanisms.

The improvements achieved by TW-LLM highlight the benefit of integrating physical knowledge into data-driven forecasting frameworks. The structural covariate interaction module addresses the limitation of conventional channel-independent modeling by enabling information exchange among related

environmental variables before LLM-based representation learning. The weighted Huber objective improves robustness against the highly skewed distribution of typhoon-induced wave observations, where extreme events constitute only a small fraction of the dataset [42]. Meanwhile, the physics-informed loss constrains unrealistic temporal variations and improves stability during long-horizon forecasting. These findings support the growing potential of physics-informed machine learning approaches for complex geophysical forecasting tasks [43, 44, 45].

The independent 2026 evaluation provides encouraging evidence of temporal generalization to previously unseen typhoon conditions. However, such temporal transfer should not be conflated with geographical generalization. All training data originate from a single regional buoy network, and the model has not yet been evaluated across fundamentally different tropical-cyclone basins, shelf geometries, or wave climates. Demonstrating cross-basin transfer—for example, between the western North Pacific, Atlantic, and Indian Ocean—would provide a substantially stronger test of whether the learned representation captures general tropical-cyclone wave dynamics rather than regional statistical structure.

From an operational perspective, TW-LLM provides an efficient alternative for coastal wave prediction by combining forecasting accuracy with model interpretability. The ability to forecast significant wave height several days in advance can support coastal hazard assessment and marine emergency management. Furthermore, the explainable prediction framework allows users to identify the dominant environmental factors associated with individual forecasts, improving the transparency of AI-assisted decision-making.

Several limitations remain. First, extreme observations are still sparse, and the largest H_{sig} values remain more difficult to predict than routine conditions. Second, the present physical regularization uses simplified wind–wave and temporal-rate proxies rather than the complete wave-action balance. More physically grounded constraints based on wave energy, spectral characteristics, wave age, fetch, and propagation may further improve extrapolation while reducing reliance on empirical coefficients. Third, the availability of future atmospheric forcing must be treated explicitly for operational long-horizon prediction. In practical 96-h forecasting, future wind information should originate from operational atmospheric forecasts rather than future observations. Quantifying the sensitivity of TW-LLM to errors in forecast wind forcing is therefore an important next step. Fourth, the present uncertainty analysis based on multiple random initializations captures model-training variability but does not represent the full predictive uncertainty associated with atmospheric forcing, observational error, and distribution shift.

Although the proposed framework effectively utilizes numerical oceanographic variables, additional multimodal information, such as satellite observations and numerical wave model outputs, may further improve forecasting capability and uncertainty estimation. In addition, deployment on distributed marine observation platforms requires further optimization of computational efficiency. Future studies may investigate multimodal fusion, hybrid numerical-AI forecasting systems, and adaptive learning strategies across heterogeneous buoy networks.

6 Conclusions

This study develops TW-LLM, a physics-informed large language model framework for forecasting typhoon-induced significant wave height. Building on Time-LLM, TW-LLM combines explicit cross-variable interaction, domain-informed semantic conditioning, extreme-aware learning, and soft physical regularization to address the coupled, imbalanced, and highly non-stationary characteristics of extreme wave evolution. Evaluation using hourly observations from eleven coastal buoys spanning 2017–2025 demonstrates improved forecasting performance across both 6-h and 96-h horizons, with particularly pronounced gains under high-wave conditions. Ablation experiments further show that cross-variable interaction and physics-informed regularization provide complementary benefits across both short- and extended-lead forecasting. An independent 2026 evaluation also demonstrates useful predictive skill for previously unseen typhoon events within the regional wave climate. These results suggest that combining

pretrained time-series representations with explicit environmental interactions and physically motivated constraints provides a promising approach to rapid extreme-wave forecasting. Further evaluation across different ocean basins and integration with operational atmospheric and numerical wave forecasts will be important for establishing broader geographical generalization and operational applicability.

Data availability

The code and data in this study are available at https://github.com/Xufeeg/Experimental_dataset_for_TW-LLM.git.

References

1. Wang, S. & Toumi, R. Recent migration of tropical cyclones toward coasts. *Science* **371,** 514–517 (2021). https://doi.org/10.1126/science.abb9038
2. Yang, Z., Ye, Q., Yang, B., Shi, W. & Zhang, J. Characteristics of measured typhoon waves on the central coast of Zhejiang province. *Front. Mar. Sci.* **11,** 1494107 (2024). https://doi.org/10.3389/fmars.2024.1494107
3. Gao, Y., Li, X., Chen, X. & Wang, L. Extreme wave and storm surge characteristics in the southeastern coastal and offshore regions of China. *Sci. Rep.* **15,** 26915 (2025). https://doi.org/10.1038/s41598-025-09737-x
4. IPCC. Climate Change *2022: Impacts, Adaptation, And Vulnerability. Contribution Of Working Group II To The Sixth Assessment Report Of The Intergovernmental Panel On Climate Change.* (Cambridge University Press, 2022). https://doi.org/10.1017/9781009325844
5. Jing, R. et al. Global population profile of tropical cyclone exposure from 2002 to 2019. *Nature* **626,** 549–554 (2024). https://doi.org/10.1038/s41586-023-06963-z
6. Booij, N., Ris, R. C. & Holthuijsen, L. H. A third-generation wave model for coastal regions: 1. Model description and validation. *J. Geophys. Res. Oceans* **104,** 7649–7666 (1999). https://doi.org/10.1029/98JC02622
7. Deb, M. & Ferreira, C. Simulation of cyclone-induced storm surges in the low-lying delta of Bangladesh using coupled hydrodynamic and wave model (SWAN + ADCIRC). *J. Flood Risk Manag.* **11,** S750–S765 (2018). https://doi.org/10.1111/jfr3.12254
8. Lodge, C. & Weaver, R. J. Coupling a parametric wave solver into a hydrodynamic circulation model to improve efficiency of nested estuarine storm surge predictions. *J. Mar. Sci. Eng.* **10,** 1117 (2022). https://doi.org/10.3390/jmse10081117
9. Pringle, W. J., Wirasaet, D., Roberts, K. J. & Westerink, J. J. Global storm tide modeling with ADCIRC v55: Unstructured mesh design and performance. *Geosci. Model Dev.* **14,** 1125–1161 (2021). https://doi.org/10.5194/gmd-14-1125-2021
10. Feng, X., Yin, B. & Yang, D. Development of an unstructured-grid wave-current coupled model and its application. *Ocean Model.* **104,** 213–225 (2016). https://doi.org/10.1016/j.ocemod.2016.06.007
11. Ding, T., Wu, D., Li, Y., Shen, L. & Zhang, X. A hybrid CEEMDAN-VMD-TimesNet model for significant wave height prediction in the South Sea of China. *Front. Mar. Sci.* **11,** 1375631 (2024). https://doi.org/10.3389/fmars.2024.1375631
12. Huang, W. et al. Reinforcement learning-based multi-model ensemble for ocean waves forecasting. *Front. Mar. Sci.* **12,** 1534622 (2025). https://doi.org/10.3389/fmars.2025.1534622
13. Janssen, P. A. E. M. *The Interaction Of Ocean Waves And Wind.* (Cambridge University Press, 2004). https://doi.org/10.1017/CBO9780511525018
14. Breiman, L. Random forests. *Mach. Learn.* **45,** 5–32 (2001). https://doi.org/10.1023/A:1010933404324
15. Vapnik, V., Golowich, S. & Smola, A. Support vector method for function approximation, regression estimation and signal processing. In *Advances in Neural Information Processing Systems* (eds M. C. Mozer, M. Jordan & T. Petsche) **9,** 281–287 (MIT Press, 1997).

16. Huang, W., Zhao, X., Huang, W., Hao, W. & Liu, Y. A training strategy to improve the generalization capability of deep learning-based significant wave height prediction models in offshore China. *Ocean Eng.* **283,** 114938 (2023). https://doi.org/10.1016/j.oceaneng.2023.114938

17. Liu, Y. et al. A fast, high-precision deep learning model for regional wave prediction. *Ocean Eng.* **288,** 115949 (2023). https://doi.org/10.1016/j.oceaneng.2023.115949

18. Akshay, K., Grace, G., Gunasekaran, K. & Samikannu, R. Power consumption prediction for electric vehicle charging stations and forecasting income. *Sci. Rep.* **14,** 5897 (2024). https://doi.org/10.1038/s41598-024-56507-2

19. Lawal, Z., Yassin, H., Lai, D. T. C. & Idris, A. Understanding the dynamics of ocean wave-current interactions through multivariate multi-step time series forecasting. *Appl. Artif. Intell.* **38,** 2393978 (2024). https://doi.org/10.1080/08839514.2024.2393978

20. Rummel, K., Strauß, T., Lauer, F. & Gräwe, U. Real-time prediction of salt intrusion in tidal estuaries using long short-term memory networks. *J. Geophys. Res. Mach. Learn. Comput.* **2,** e2025JH000768 (2025). https://doi.org/10.1029/2025JH000768

21. Sun, Y., Yu, L. & Zhu, D. A hybrid deep learning model based on FFT-STL decomposition for ocean wave height prediction. *Appl. Sci.* **15,** 5517 (2025). https://doi.org/10.3390/app15105517

22. Thet, P., Tao, A., Lv, T. & Zheng, J. Wave height forecasting in the Bay of Bengal using multivariate hybrid deep learning models. *J. Mar. Sci. Eng.* **13,** 1412 (2025). https://doi.org/10.3390/jmse13081412

23. James, S. C., Zhang, Y. & O'Donncha, F. A machine learning framework to forecast wave conditions. *Coast. Eng.* **137,** 1–10 (2018). https://doi.org/10.1016/j.coastaleng.2018.03.004

24. Naeini, S. & Snaiki, R. A physics-informed machine learning model for time-dependent wave runup prediction. *Ocean Eng.* **296,** 116986 (2024). https://doi.org/10.1016/j.oceaneng.2024.116986

25. Baggio, R. et al. Local wind speed forecasting at short time horizons based on numerical weather prediction and observations from surrounding stations. *J. Geophys. Res. Mach. Learn. Comput.* **2,** e2025JH000709 (2025). https://doi.org/10.1029/2025JH000709

26. Ehlers, S., Stender, M. & Hoffmann, N. Bridging ocean wave physics and deep learning: Physics-informed neural operators for nonlinear wavefield reconstruction in real-time. Preprint at https://arxiv.org/abs/2508.03315 (2025). https://doi.org/10.48550/arXiv.2508.03315

27. Naeini, S. S., Snaiki, R. & Wu, T. Advancing spatio-temporal storm surge prediction with hierarchical deep neural networks (Version 1). Preprint at https://arxiv.org/abs/2410.12823 (2024). https://doi.org/10.48550/arXiv.2410.12823

28. Zhu, Z. et al. Physics informed neural network modelling for storm surge forecasting — A case study in the Bohai Sea, China. *Coast. Eng.* **197,** 104686 (2025). https://doi.org/10.1016/j.coastaleng.2024.104686

29. Jin, M. et al. Time-LLM: Time series forecasting by reprogramming large language models. Preprint at https://arxiv.org/abs/2310.01728 (2024). https://doi.org/10.48550/arXiv.2310.01728

30. Abdullahi, S. et al. Time-series large language models: A systematic review of state-of-the-art. *IEEE Access* **13,** 30235–30261 (2025). https://doi.org/10.1109/access.2025.3535782

31. Hu, Y., Liao, H., Wu, M. & Yuan, L. SST-LLM: Time series forecasting based on large language models. In *Seventh International Conference on Computer Information Science and Artificial Intelligence (CISAI 2024)* **13681,** 136810Q (SPIE, 2025). https://doi.org/10.1117/12.3073528

32. Chen, T. et al. OpenTSLM TeeMoE: A unified time-series language model for forecasting, contextual prediction, and reasoning. Preprint at https://arxiv.org/abs/2609.40265 (2026). https://doi.org/10.48550/arXiv.2609.40265

33. Sun, Y. et al. Adapting LLMs to time series forecasting via temporal heterogeneity modeling and representation alignment. Preprint at https://arxiv.org/abs/2508.07195 (2026). https://doi.org/10.48550/arXiv.2508.07195

34. Nie, Y., Nguyen, N. H., Sinthong, P. & Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. Preprint at https://arxiv.org/abs/2211.14730 (2022). https://doi.org/10.48550/arXiv.2211.14730

35. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* **30,** 5998–6008 (2017).

36. Huber, P. J. Robust estimation of a location parameter. *Ann. Math. Stat.* **35,** 73–101 (1964). https://doi.org/10.1214/aoms/1177703732

37. Cavaleri, L. & Malanotte-Rizzoli, P. Wind wave prediction in shallow water: Theory and applications. *J. Geophys. Res. Oceans* **86,** 10961–10973 (1981). https://doi.org/10.1029/JC086iC11p10961

38. Holthuijsen, L. H. *Waves In Oceanic And Coastal Waters.* (Cambridge University Press, 2007). https://doi.org/10.1017/CBO9780511618536

39. Jiang, L. L., Yao, R. S., Wang, Y. & Tu, X. P. Analysis of typhoon wave characteristics on the sea surface of Zhejiang coastal area based on buoy stations and ERA5 reanalysis data. *J. Ocean Technol.* **43,** 1–9 (2024). https://doi.org/10.3969/j.issn.1003-2029.2024.01.001

40. Taylor, K. E. Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res. Atmos.* **106,** 7183–7192 (2001). https://doi.org/10.1029/2000JD900719

41. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30,** 4765–4774 (2017). https://doi.org/10.48550/arXiv.1705.07874

42. Gong, Y., Dong, S. & Wang, Z. Forecasting of typhoon wave based on hybrid machine learning models. *Ocean Eng.* **266,** 112934 (2022). https://doi.org/10.1016/j.oceaneng.2022.112934

43. Donnelly, J., Daneshkhah, A. & Abolfathi, S. Physics-informed neural networks as surrogate models of hydrodynamic simulators. *Sci. Total Environ.* **912,** 168814 (2024). https://doi.org/10.1016/j.scitotenv.2023.168814

44. Sharma, P., Chung, W., Akoush, B. & Ihme, M. A review of physics-informed machine learning in fluid mechanics. *Energies* **16,** 2343 (2023). https://doi.org/10.3390/en16052343

45. Shen, C. et al. Differentiable modelling to unify machine learning and physical models for geosciences. *Nat. Rev. Earth Environ.* **4,** 552–567 (2023). https://doi.org/10.1038/s43017-023-00450-9

Author contributions

Zhiguo He: Conceptualization; Methodology; Supervision; Project administration; Funding acquisition; Writing – review & editing.

Yiting Zhao: Methodology; Software; Formal analysis; Investigation; Validation; Visualization; Writing – original draft.

Dongrui Han: Data curation; Investigation; Validation; Visualization; Funding acquisition; Writing – review & editing.

Hui Shen: Resources; Investigation; Validation; Writing – review & editing.

Ye Zhu: Data curation; Validation; Investigation; Writing – review & editing.

Xiangbing Kong: Methodology; Formal analysis; Validation; Writing – review & editing.

Li Li: Methodology; Validation; Writing – review & editing.

Acknowledgments

This work is supported by the National Key R&D Program of China (2023YFC3008100), the National Natural Science Foundation of China (12572470), the Postdoctoral Fellowship of CPSF (GZC20251267, 2025M781799) and Zhejiang Province (ZJ2025075), and the Zhejiang Provincial Natural Science Foundation (LMS26E090004).

Additional information

Competing interests

The authors declare that there are no conflicts of interest for this manuscript.

Figure legends

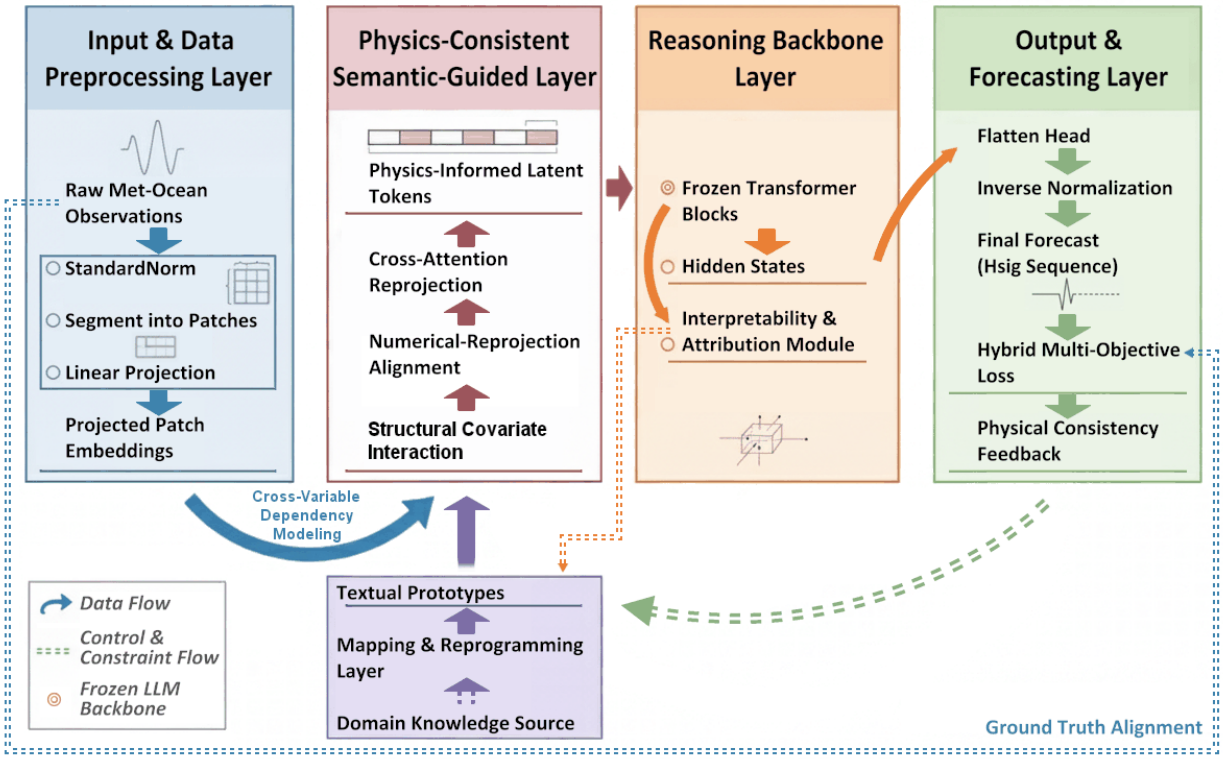


Figure 1. Schematic architecture of the TW-LLM framework, illustrating the four functional stages: (I) Temporal and Covariate Representation, (II) Cross-Modal Reprogramming and Domain Grounding, (III) Pre-trained LLM Representation, and (IV) Forecasting and Physics-informed Optimization. The architecture combines a numerical data-flow pathway and a complementary domain-guidance and physical-constraint pathway.

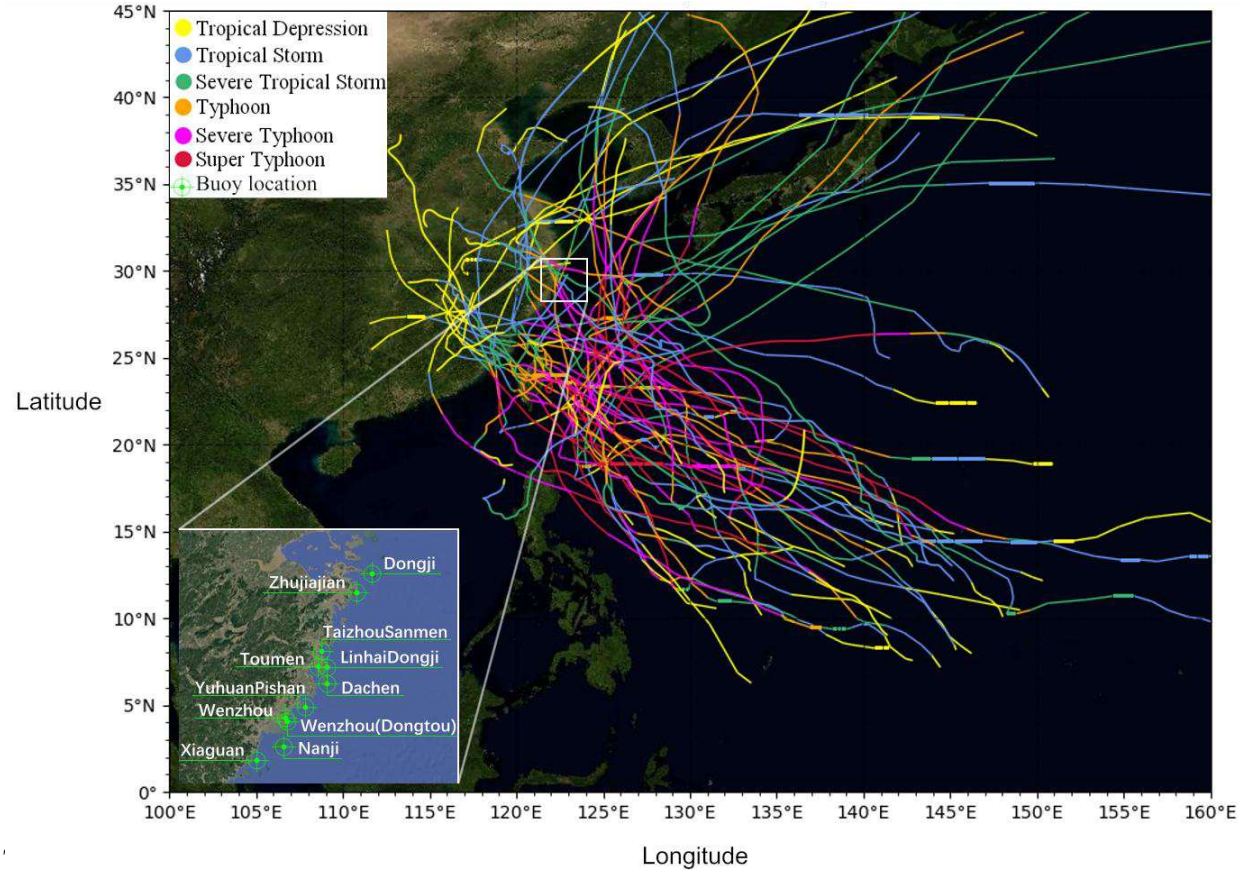


Figure 2. Location of the eleven coastal buoy stations and tracks of tropical cyclones affecting the study region during 2017–2025. Tropical-cyclone tracks are obtained from the China Meteorological Administration's Best-Track Database and categorized according to storm intensity.

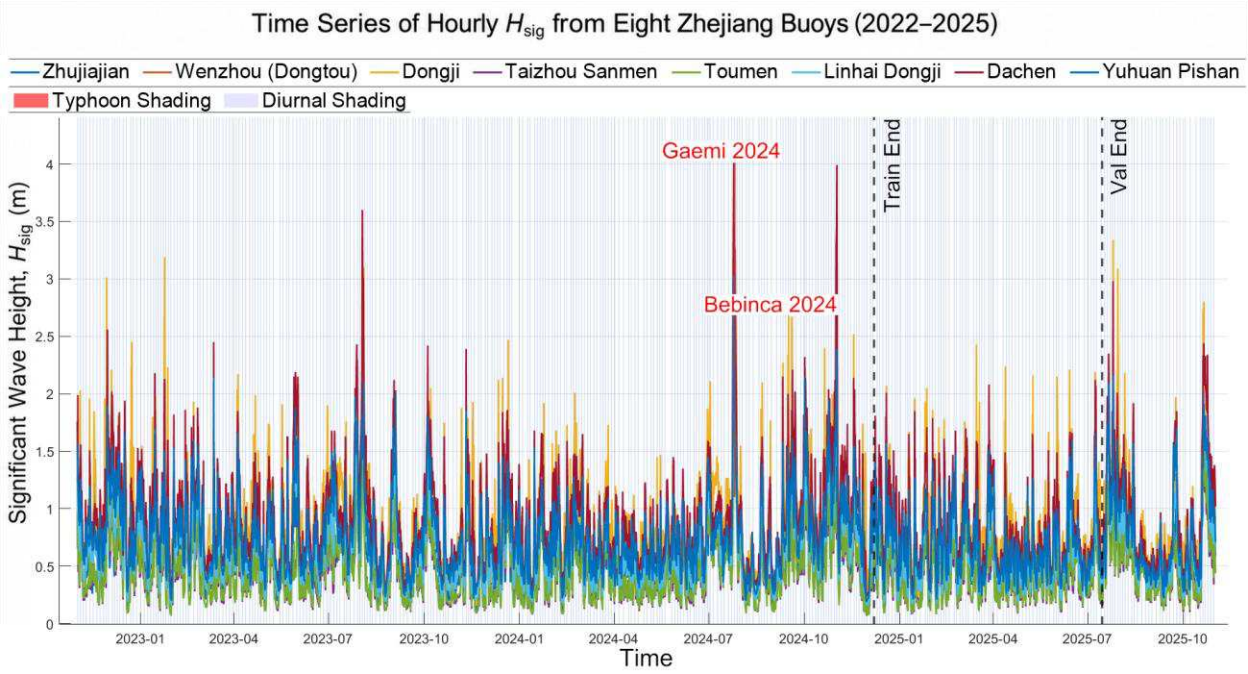


Figure 3. Representative hourly H_{sig} records from eight buoy stations (2022–2025). Black dashed vertical lines indicate the training and validation period endpoints where applicable.

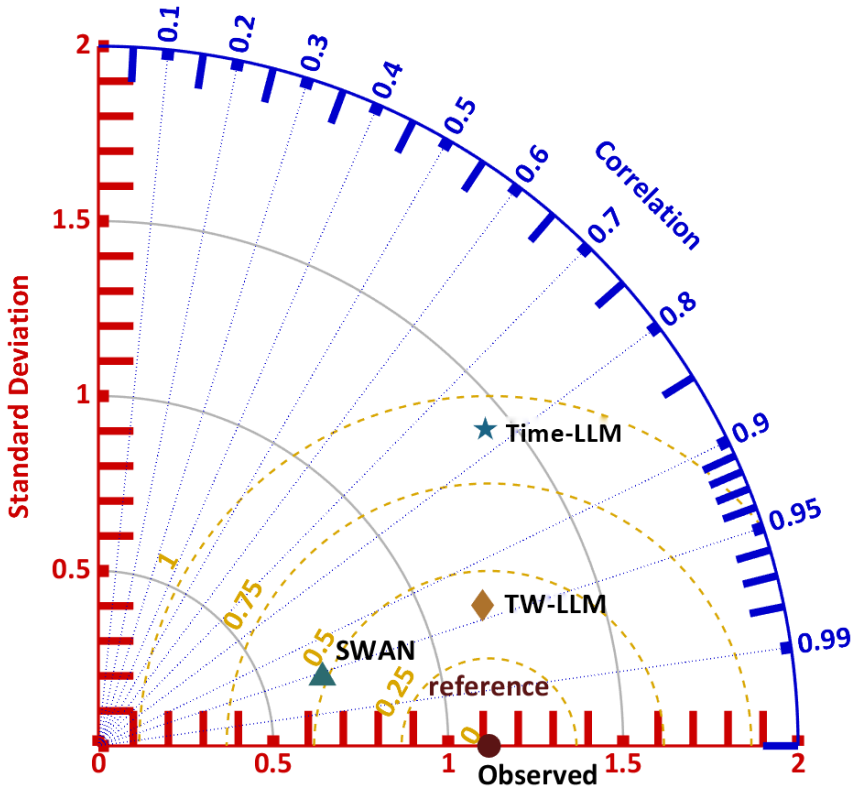
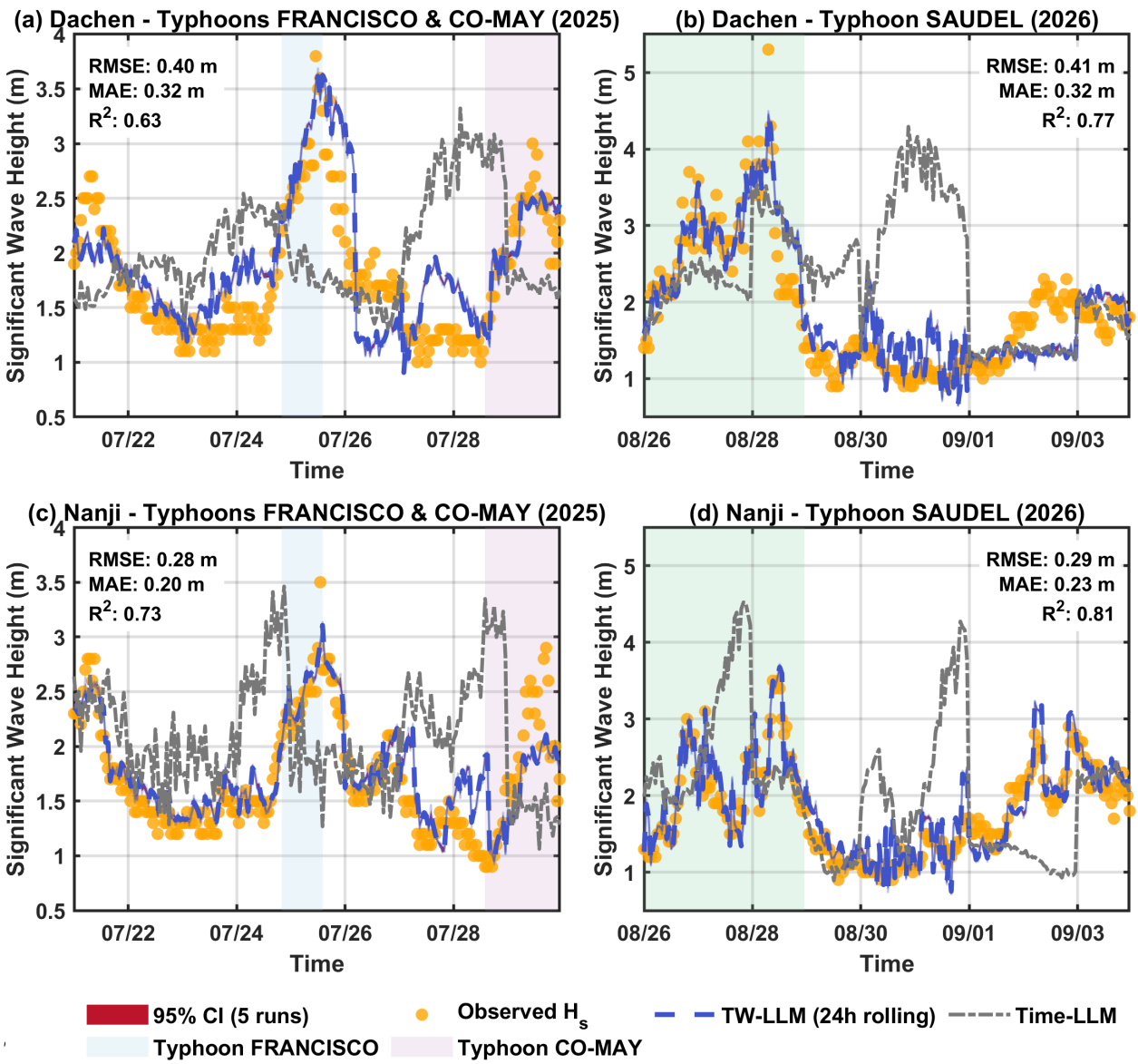


Figure 4. Taylor diagram comparing TW-LLM with the Time-LLM baseline and SWAN numerical simulations.



745 Figure 5. Observed and predicted significant wave height at the Dachen and Nanji buoy stations during selected
746 typhoon-affected periods in 2025 and 2026. Orange dots denote observed significant wave height, blue curves
747 represent the TW-LLM forecasts with a 24-h rolling prediction, gray dashed curves indicate the Time-LLM baseline,
748 shaded bands around the TW-LLM curves denote the 95% confidence intervals estimated from five independent
749 runs, and shaded regions indicate typhoon-affected periods (blue for Francisco, purple for Co-May, and green for
750 Saudel). The 2026 period is used exclusively as an independent temporal evaluation and is not included in model
751 development.

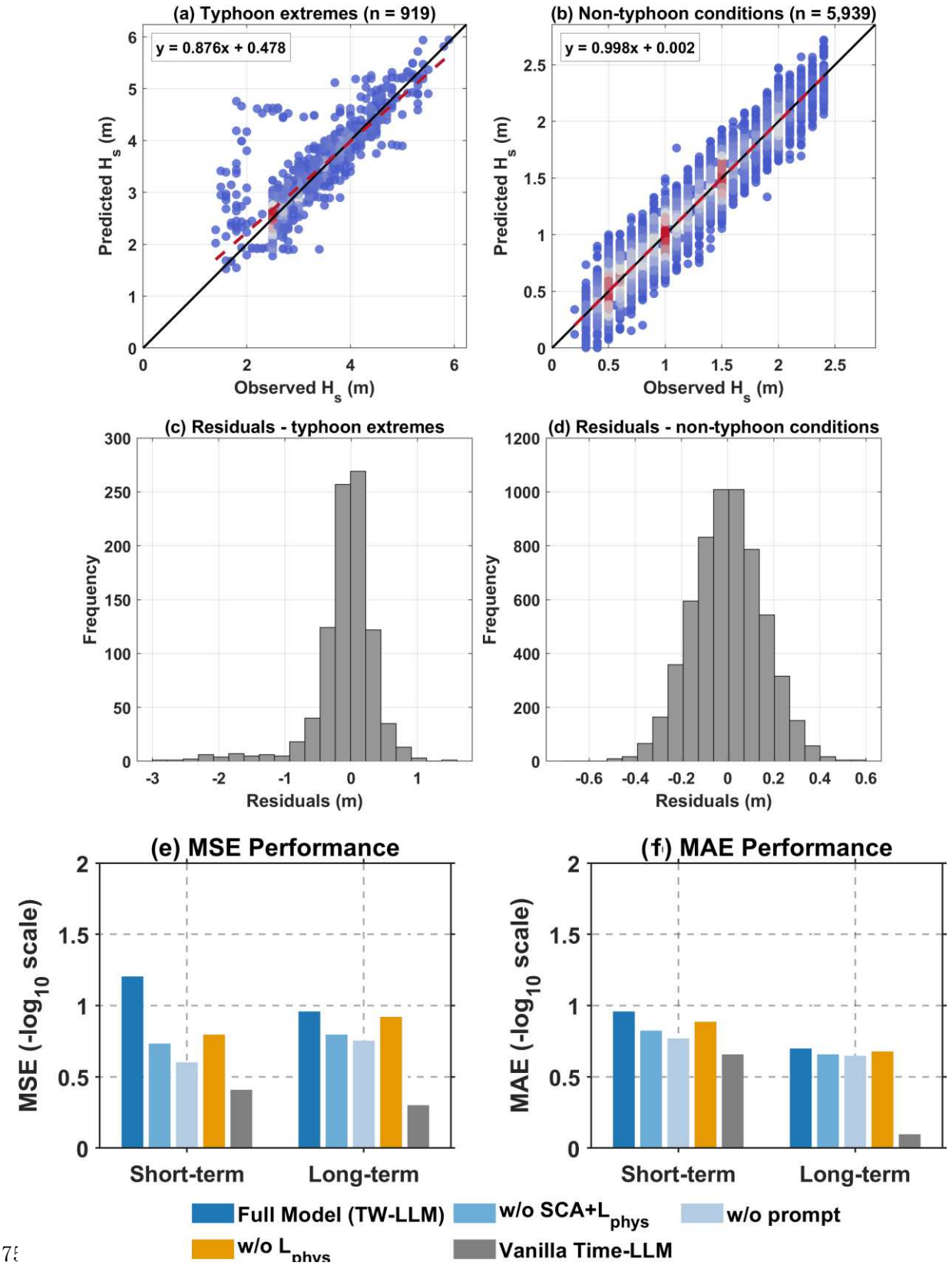


Figure 6. Predicted versus observed significant wave height for the Dachen test dataset (n = 6,858 samples) under (a) typhoon extremes and (b) non-typhoon conditions, with the corresponding residual distributions shown in (c) and (d). The solid black line denotes the 1:1 reference, and the red dashed regression line quantifies systematic prediction bias. (e, f) Ablation analysis of TW-LLM and its variants for short-term (6-h) and long-term (96-h) forecasting horizons, evaluated using the MSE and MAE metrics.

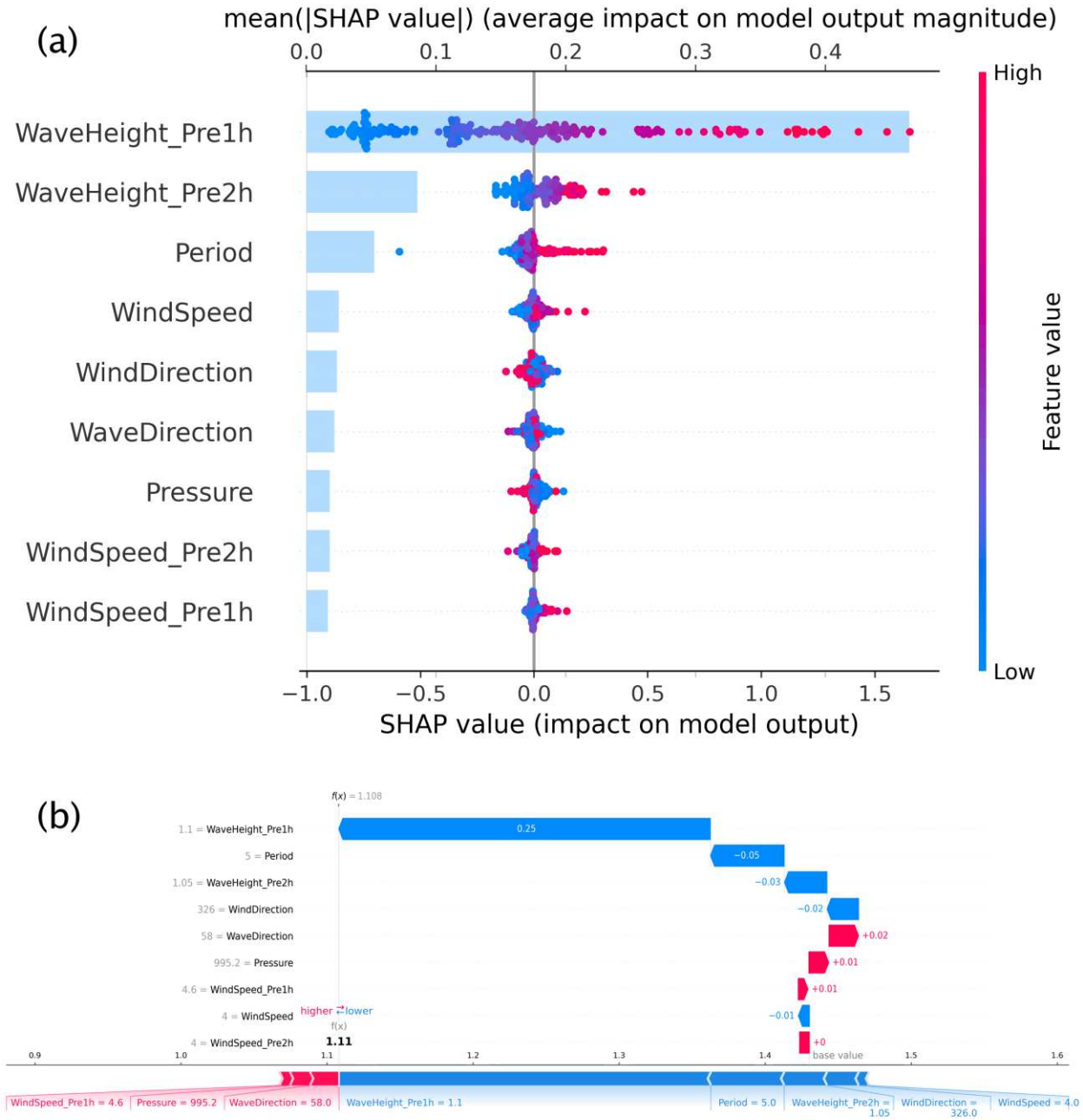


Figure 7. KernelSHAP analysis of TW-LLM: (a) global feature importance and SHAP value distributions over the test dataset; (b) local waterfall and force-plot explanations for a typhoon-induced wave event at a representative coastal buoy station during Typhoon Gaemi (2024), where blue and red bars denote contributions that decrease and increase the prediction, respectively.

763 **Table 1. Summary of the eleven coastal buoy datasets used in this study.**

Buoy ID	Location (Lat/Lon)	Mean H_s (m)	Std H_s (m)	Sample Size (Train/Val/Test)	Typhoon Event Count
Zhujiajian	29.894°N, 122.427°E	1.178	0.843	63,360 (7:2:1)	53
Wenzhou (Dongtou)	27.812°N, 121.159°E	0.654	0.490	53,472 (7:2:1)	47
Dongji	30.188°N, 122.739°E	1.610	0.698	8,976 (K-fold Cross-Validation)	10
TaizhouSanmen	28.958°N, 121.812°E	0.725	0.444	8,976 (K-fold Cross-Validation)	10
Toumen	28.711°N, 121.733°E	0.341	0.590	8,976 (K-fold Cross-Validation)	10
LinhaDongji	28.698°N, 121.919°E	1.364	0.682	8,976 (K-fold Cross-Validation)	10
Dachen	28.44°N, 121.915°E	1.469	0.843	70,632 (7:2:1)	57
YuhuanPishan	28.067°N, 121.53°E	1.566	0.717	8,976 (K-fold Cross-Validation)	10
Wenzhou	27.895°N, 121.157°E	0.996	0.443	8,976 (K-fold Cross-Validation)	10
Nanji	27.425°N, 121.121°E	1.759	0.847	8,976 (K-fold Cross-Validation)	10
Xiaguan	27.207°N, 120.628°E	1.484	0.840	8,976 (K-fold Cross-Validation)	10

764