electronics

MDPI

Article

A Significant Wave Height Prediction Method Based on Improved Temporal Convolutional Network and Attention Mechanism

Ying Han [1], Jiaxin Tang [2,*], Hongyun Jia [2], Changming Dong [3] and Ruihan Zhao [2]

[1] School of Electronics and Electrical Engineering, Wuhan Textile University, Wuhan 430200, China; 2023088@wtu.edu.cn
[2] School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China; 002337@nuist.edu.cn (H.J.); 202212490085@nuist.edu.cn (R.Z.)
[3] School of Marine Science, Nanjing University of Information Science and Technology, Nanjing 210044, China; dcm@nuist.edu.cn
* Correspondence: 202212490024@nuist.edu.cn

Abstract: Wave prediction is crucial for ensuring the safety and disaster mitigation of coastal areas, helping to support marine economic activities. Currently, many deep learning models, such as the temporal convolutional network (TCN), have been applied to wave prediction. In this study, a prediction model based on improved TCN-Attention (ITCN-A) is proposed. This model incorporates improvements in two aspects. Firstly, to address the difficulty of calibrating hyperparameters in traditional TCN models, a whale optimization algorithm (WOA) has been introduced to achieve global optimization of hyperparameters. Secondly, we integrate dynamic ReLU to implement an adaptive activation function. The improved TCN is then combined with the attention mechanism to further enhance the extraction of long-term features of wave height. We conducted experiments using data from three buoy stations with varying water depths and geographical locations, covering prediction lead times ranging from 1 h to 24 h. The results demonstrate that the proposed integrated model reduces the *RMSE* of prediction by 12.1% and *MAE* by an 18.6% compared with the long short-term memory (LSTM) model. Consequently, this model effectively improves the accuracy of wave height predictions at different stations, verifying its effectiveness and general applicability.

Keywords: significant wave height; temporal convolutional network; attention mechanism

check for updates

Citation: Han, Y.; Tang, J.; Jia, H.; Dong, C.; Zhao, R. A Significant Wave Height Prediction Method Based on Improved Temporal Convolutional Network and Attention Mechanism. *Electronics* **2024**, *13*, 4879. https://doi.org/10.3390/electronics13244879

Academic Editor: Domenico Rosaci

Received: 12 November 2024
Revised: 1 December 2024
Accepted: 3 December 2024
Published: 11 December 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Waves are a fundamental component of the ocean system and have significant impacts on human activities and the ecological environment. Even some extreme wave phenomena, such as storms or tsunamis, can cause serious natural disasters [1,2]. Waves can also pose significant threats to various maritime and coastal activities. For instance, rough sea conditions can endanger the safety of maritime navigation, increasing the risk of vessel capsizing, collisions, or delays in shipping routes [3]. In ports, high waves can disrupt loading and unloading operations, compromise structural stability, and lead to economic losses due to downtime [4]. Furthermore, the unpredictability of wave patterns presents a challenge for marine fisheries, affecting the safety of fishing vessels and the sustainability of fish stocks [5]. Therefore, accurate prediction of wave height is crucial to ensure the safety of human activities [6–9]. In addition, in the field of new energy, wave energy converters can utilize the kinetic and potential energy of waves to generate electricity, and accurate wave forecasting is beneficial for improving energy production efficiency [10–13]. Of course, waves also bear the responsibility of nutrient exchange and sediment transport in marine ecology, playing a crucial role in the sustainable growth of marine organisms [14].

Historically, such predictions have predominantly relied on numerical models such as the wave model (WAM) [15], simulating waves nearshore (SWAN) [16], and Wavewatch III (WW3) [17]. These numerical models are grounded in energy balance equations, describing

physical processes including wind stress, wave–wave nonlinear interactions, and bottom dissipation. However, their reliance on grid data, encompassing inputs like wind and wave fields, poses a significant impediment when applying them to local wave prediction at individual buoy stations [18]. Furthermore, during the parameterization process, approximate functions are utilized, which inevitably introduce errors into the numerical models [19]. Hence, it is imperative to identify a model tailored to enhance the accuracy of significant wave height (SWH) prediction at individual buoy stations.

The rapid development of artificial intelligence (AI) technology offers novel avenues for related research. Distinct from conventional numerical models, AI-based methodologies afford greater flexibility, facilitating predictions using merely historical series data sourced from buoy stations. AI is a generalized concept that includes machine learning and deep learning. Established machine learning algorithms such as support vector regression (SVR) [20], extreme learning machine (ELM) [21], nonlinear function-on-function model [22], and random forest [23] have demonstrated utility in prediction tasks. Simultaneously, the accelerated evolution of deep learning has prompted researchers to leverage these techniques for SWH prediction [24]. Deo et al. [25] employed a straightforward three-layered artificial neural network (ANN) to predict the wave height across three different locations along the Indian coast and achieved satisfactory results. Due to the parallel relationship between different inputs of ANN, the temporal dependencies within the input data are ignored. In contrast, the recurrent neural network (RNN) incorporates the recursive input structure, which can effectively capture the temporal features of input data [26]. Notably, LSTM [27–30], an advanced variant of RNN, address challenges like long-term dependencies and gradient vanishing in RNN. Fan et al. [31] employed an LSTM model to predict wave heights for both 1 h and 6 h intervals, using historical wind speed and wave height data as input parameters. When tested at 10 different buoy sites, it was found that the prediction accuracy and stability were superior to other deep learning models such as ANN. In addition, building upon LSTM, bidirectional LSTM (Bi-LSTM) was applied to the estimation of tidal level, with a stronger capability in handling long-term dependencies [32]. The AI-based models mentioned above revolve around time series forecasting, which ingest historical data like SWH, wind speed, and direction to extrapolate future wave heights.

LSTM and other recurrent neural networks have temporal dependencies during computation, resulting in performance bottlenecks in parallel computing. Convolutional neural networks (CNNs), with their sliding filtering structure, greatly improve the efficiency of parallel computing, and many researchers have applied CNN networks to regional wave prediction [33–35]. In addition, CNN networks and other neural networks or time–frequency decomposition techniques are integrated to construct prediction models [36,37]. On the basis of CNNs, the TCN uses causal dilation convolutional layers to perform convolution operations on input sequences, which not only facilitates parallel computing but also enhances the extraction of temporal features. Ji et al. [38] proposed an effective wave height prediction model based on variational mode decomposition and the TCN and established a prediction model using the TCN and Bayesian hyperparameter optimization, which has been improved in multi-step prediction. Huang et al. [39] tested the predictive performance of ANN, LSTM, and TCN models in China's offshore waters using a multi-station data fusion training strategy. The results showed that multi-station data can improve the prediction results. Lou et al. [40] combined the TCN with empirical mode decomposition (EMD) and applied this hybrid model to buoy observation data. The effectiveness of EMD-TCN in wave height prediction has been verified, and the lag problem in previous wave height prediction research has been eliminated, improving the accuracy of wave height prediction.

In recent years, the attention mechanism has received increasing attention. As the core module of Transformer, the attention mechanism has the ability to extract global features. The attention mechanism has been widely applied in prediction fields such as traffic flow and wind power [41–44]. Zhang et al. [45] applied the TCN-Attention model to ship motion prediction. Different weights were assigned to the original features through the attention mechanism, and it effectively improves the prediction accuracy. Luo et al. [46] combined

Bi-LSTM and attention to predict wave heights in the Atlantic hurricane zone, selecting four data features collected at five buoy stations (wave height, wind speed, wind direction, and wave direction) as model inputs and future wave heights as model outputs. Compared with the benchmark model, the BLA model has a more stable predictive performance.

However, to the best of our knowledge, few researchers have applied the TCN-Attention model to the field of wave prediction. The volatility of waves necessitates strong global feature extraction capabilities in the model. The TCN-Attention model achieves efficient feature learning through global receptive fields and attention weight allocation. Thus, this paper employs the TCN-Attention model to predict wave heights. Additionally, to address the time-consuming challenge of determining hyperparameters in traditional deep learning models, we introduce WOA for global hyperparameter optimization, making the process more efficient. Furthermore, the fixed input and output modes of the plain ReLU activation function limit the model's expressive power. To overcome this limitation, a dynamic ReLU activation function is introduced, which adjusts the parameters of the activation function based on the input sequence, thereby enhancing the model's representation flexibility.

This paper is organized as follows. Section 2 introduces the principles of each model. Section 3 introduces the dataset information. Section 4 presents the predicted results of each model and discusses them. Finally, Section 5 provides a conclusion.

2. Method

2.1. Temporal Convolutional Network

TCN is deep learning architecture designed for processing sequences of data, and it is particularly well suited for tasks involving temporal dependencies [47]. The TCN model has shown considerable success in a variety of sequential data tasks, including natural language processing, audio processing, and time series analysis.

TCN is composed of a dilated causal convolutional network and a residual network, with the same length of input and output. The convolutional network used by TCN is a causal convolutional network, which means that the current output is only related to historical inputs and not to future inputs, avoiding the problem of information leakage. In addition, this architecture can map sequences of any length to output sequences of the same length using a process similar to RNN and LSTM. Figure 1 shows the causal convolutional network adopted by TCN, which achieves a larger receptive field through dilated convolution.

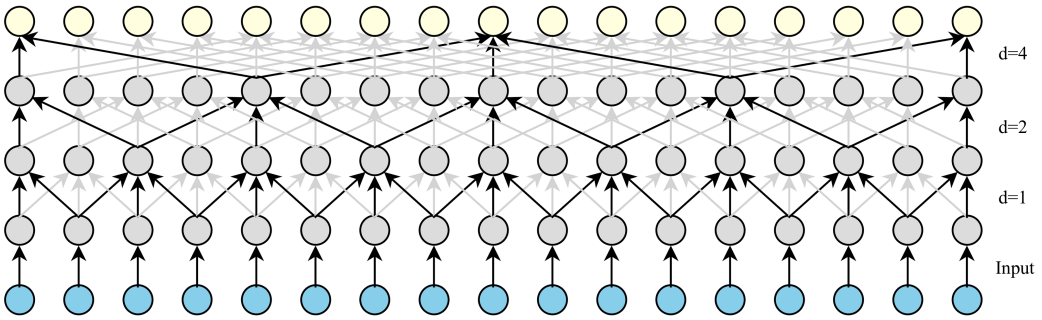


Figure 1. Causal convolutional structure figure.

Dilated convolution adds an dilation factor d on the basis of ordinary convolutional networks, which increases the size of the time series observed by the network without increasing computational complexity. The calculation formula for dilated causal convolution is as follows:

$$Y(s) = (x * f)(s) = \sum_{i=0}^{k-1} f(i) \cdot x_{s-d \cdot i} \quad (1)$$

In the formula: x represents the input sequence, f represents the convolution kernel, s represents the sequence index, k is the size of the convolution kernel, and d represents the dilation coefficient. It increases with the increase in the network layer, and the receptive field is expanded by controlling the size of d and k.

Residual networks are used to solve the problem of performance degradation in traditional deep learning networks when the network layers are very deep. In the TCN model, the superposition of causal convolution and dilated convolution leads to a deeper number of layers. In order to avoid gradient explosion and vanishing during training, residual connections are introduced to fuse the input into the output of the convolutional network. The formula is as follows:

$$o = \varphi(x + f(x)) \tag{2}$$

In the formula, x represents the input, $f(x)$ represents the output of the convolutional layer, and φ represents the activation function.

TCN consists of two dilated causal convolutions and nonlinear mapping layers, as shown in Figure 2. The convolutional layer adopts one-dimensional dilated causal convolution and adjusts the receptive field by adjusting the convolution dilation size through the dilation coefficient. Next, we use dynamic ReLu as the activation function, which dynamically adjusts the parameters of ReLu based on the convolutional output. Finally, we add a dropout layer after each dilated convolution to avoid overfitting during training.

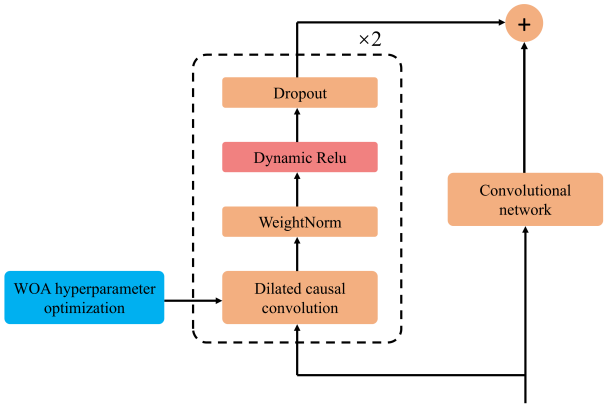


Figure 2. Improved TCN block.

TCN can be efficiently parallelized across time steps, making them computationally efficient, especially when compared with RNN that process sequences sequentially. Due to dilated convolutions, TCNs can capture long-range dependencies in sequences without relying on recurrent connections, but this also brings about the problem of vanishing gradients. The use of residual connections effectively avoids this problem.

Unlike traditional TCN networks, the improved TCN model uses dynamic ReLu [48] as the activation function. The current mainstream ReLu activation function $y = max\{x, 0\}$ is static and performs the same operation on all input sequences. Its parameters are generated by a hyperfunction, which is related to the input sequence. The core of dynamic ReLu is to encode the global sequence as a hyperfunction and adjust the piecewise linear activation function accordingly. Compared with static activation functions, the additional computational cost of dynamic ReLu can be ignored, but it has greater expressive power. Dynamic ReLu can be represented by parameterized piecewise functions:

$$y_c = f_{\theta(x)}(x_c) = \max_{1 \le k \le K} a_c^k(\mathbf{x}) x_c + b_c^k(\mathbf{x}) \tag{3}$$

a_c^k and b_c^k are dynamically adjusted based on the input sequence $x = x_c$, $\theta(x)$ is a hyperfunction, K is the number of piecewise functions, C is the number of channels, and the activation function parameters are related to the input x_c.

2.2. Attention Mechanism

The attention mechanism is inspired by human visual attention and was first applied in the field of computer vision [49]. The basic idea of the attention mechanism is to select information that is more critical to the current task goal from the input information. It assigns corresponding weights based on the importance of the input information, thereby capturing more valuable information. With the continuous development and improvement of the attention mechanism, they have been applied to time series prediction and have achieved good results.

The structure of the attention mechanism is shown in Figure 3. Firstly, a multi-layer perceptron is used to calculate the similarity weight of the data at each time step. The calculation formula is as follows:

$$e_t = W_b tanh(W_a h_t) \tag{4}$$

Among them, W_a represents the connection weight between the input layer and hidden layer of MLP, W_b represents the connection weight between the hidden layer and output layer of MLP, and h_t represents the hidden state matrix. Next, use the softmax function to normalize these similarity weights, and the calculation formula is as follows:

$$a_t = softmax(e_t) = \frac{exp(e_t)}{\sum_{j=t-1}^{t+i}(e_j)} \tag{5}$$

Among them, exp() represents an exponential function based on the natural constant e. Finally, the normalized similarity weights and corresponding data are weighted and summed to obtain the self-attention output matrix R. The calculation formula is as follows:

$$R = \sum_{j=t-1}^{d} a_j h_j \tag{6}$$

Among them, h_j represents the hidden state vector at time j, and a_j represents the normalized similarity weight of h_j. In wave prediction, the self-attention mechanism can use MLP to calculate similarity weights for hourly wave height data. The self attention mechanism can assign corresponding weights to the prediction results based on the importance of historical wave data at different times, thereby capturing more valuable information.

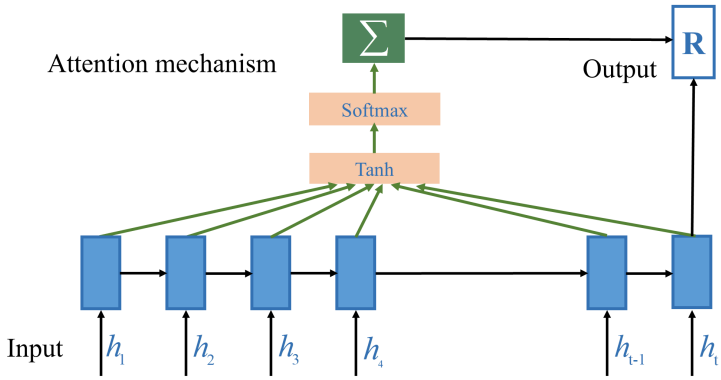


Figure 3. Attention mechanism figure.

2.3. Whale Optimization Algorithm

The WOA is a biological heuristic algorithm proposed by Mirjalili and Lewis [50] that aims to simulate the behavior of group hunting. Compared with other optimization algorithms such as particle swarm optimization, the whale optimization algorithm has faster optimization speed and supports distributed computing, stronger global optimization ability, and fewer parameter adjustments. The whale optimization algorithm aims to

simulate three behaviors of humpback whales: encircling prey, spiral bubble hunting, and searching for prey.

(1) Surrounding prey

A group of humpback whales will find the direction of their prey and surround them. During this process, whales will determine the position of their prey based on the optimal direction in the current optimization space and continuously approach the prey. In this iterative cycle, the position of whales will also be constantly updated. The formula for surrounding behavior is described as follows:

$$\vec{D} = |\vec{C}.\vec{X}^*(t) - \vec{X}(t)| \tag{7}$$

$$\vec{X}(t+1) = \vec{X}^*(t) - \vec{A} \cdot \vec{D} \tag{8}$$

Among them, t represents the current number of iterations, $\vec{A}, \vec{C}$ is the coefficient vector, X^* is the optimal solution in the current optimization space, updated if a better solution appears after each iteration, $\vec{X}$ is the orientation vector of the whale group, and $\cdot$ is the elementwise multiplication of the vector. The coefficient vector $\vec{A}, \vec{C}$ is defined as follows:

$$\vec{A} = 2\vec{a} \cdot \vec{r} - \vec{a} \tag{9}$$

$$\vec{C} = 2 \cdot \vec{r} \tag{10}$$

In the formula, $\vec{a}$ linearly decreases from 2 to 0 during continuous iteration, while $\vec{r}$ is a random vector that fluctuates in the interval [0, 1].

(2) Spiral bubble hunting (mining stage)

Spiral bubble hunting involves two behavioral mechanisms: contraction encirclement and spiral orientation update, with effective probabilities of 0.5 each. The random number p determines which mechanism to adopt to attack the prey. The mathematical description of the contraction enclosure mechanism is shown in Formula (8), where the whale's possible new position will be at any position between its original position and the optimal position. Unlike the square contraction circle formed by contraction and encirclement behavior, the spiral position update mechanism allows whales to perform spiral motion between their current position and prey. The mathematical model of spiral bubble hunting is shown as follows:

$$\vec{X}(t+1) = \begin{cases} \vec{X}^*(t) - \vec{A} \cdot \vec{D} & if\ p < 0.5 \\ \vec{D}\prime \cdot e^{bl} \cdot cos(2\pi l) + \vec{X}^*(t) & if\ p \geq 0.5 \end{cases} \tag{11}$$

Among them, $\vec{D}\prime = |\vec{X^*(t)} - \vec{X(t)}|$ represents the distance between the whale and its prey (i.e., the distance between the whale and the current optimal optimization agent), the constant b is used to define the logarithmic helix, l is a random number within the interval $[-1, 1]$, and p is a random number within the interval [0, 1].

(3) Search for predation (exploration stage)

Unlike the development phase, the search for prey (exploration phase) is based on randomly selected search agent locations rather than the location of the optimal agent. This model design helps the optimization agent to jump out of local optima and converge towards a more optimal direction. The search and predation process is described using a mathematical model as follows:

$$\vec{D} = |\vec{C} \cdot \overrightarrow{X_{rand}} - \vec{X}| \tag{12}$$

$$\vec{X}(t+1) = \vec{X_{rand}} - \vec{A} \cdot \vec{D} \tag{13}$$

Among them, $\vec{X_{rand}}$ is the position of a randomly selected search agent (i.e., a humpback whale) in the current iteration calculation. The value range of $\vec{A}$ is different from

the previous stage. In order to expand the search range and find the global optimum, the random number of vectors $\vec{A}$ is greater than 1 or less than -1, that is, $|\vec{A}| > 1$, and the update amplitude of the search agent coordinates will be significantly increased.

2.4. Model Structure

The ITCN-A prediction model proposed in this article is shown in Figure 4. The model can be divided into three stages: data preprocessing, WOA hyperparameter optimization, and TCN-Attention model prediction.

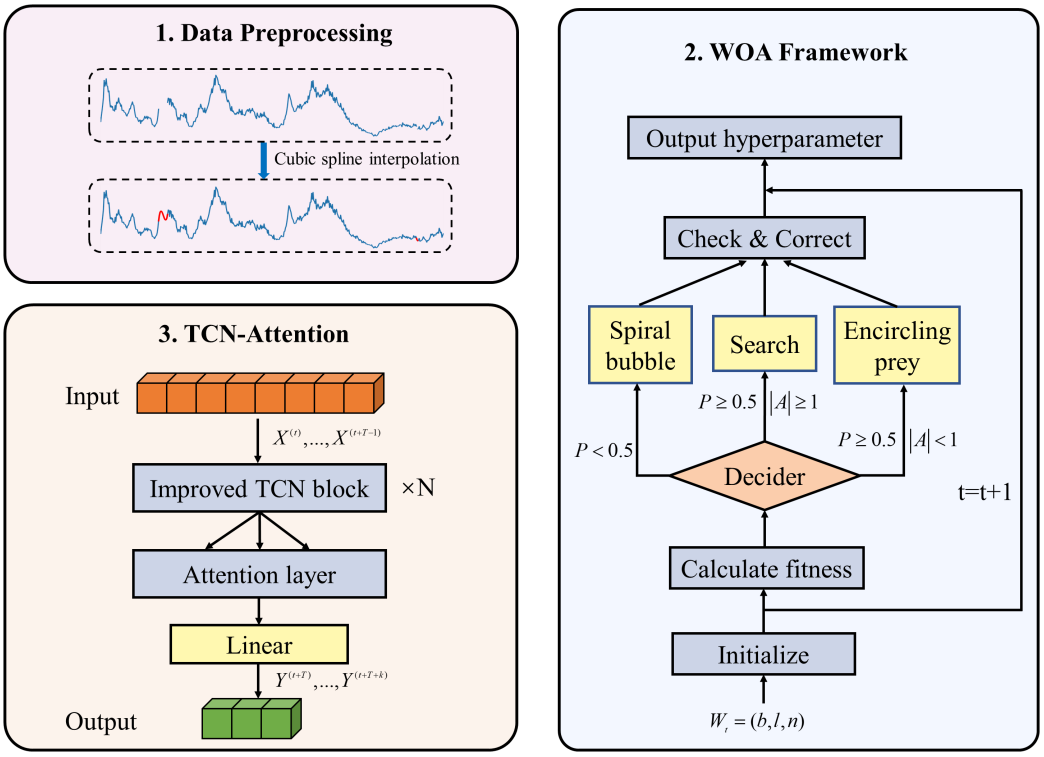


Figure 4. Structure of proposed method.

Stage 1: Firstly, cubic spline interpolation is used to fill in the missing values of the original time series. Non-nodal boundaries are selected as boundary conditions, and the interpolation is solved by combining equations based on the condition that the original function, first-order derivative, and second-order derivative are all continuous.

Stage 2: Construct the optimization space with the hyperparameters of the TCN, including the number of convolutional kernels, batch size, and learning rate. The WOA first randomly generates the initial position of search agents, and then these search agents will adaptively choose different hunting modes to update their coordinate positions. When the termination condition is met, the WOA will output the globally optimal hyperparameter combination.

Stage 3: Build a prediction model through the TCN-Attention network. The TCN uses dilated causal convolution to extract information from the input sequence, with consistent input and output sizes. Then, the TCN output sequence is input into a multi-head attention mechanism to further obtain global information. Finally, a fully connected network is constructed behind the attention mechanism to obtain the predicted wave height value. Unlike traditional TCN networks, the TCN network used in this paper adapts the dynamic ReLu activation function, which dynamically adjusts the activation function parameters based on the input sequence, making it very flexible.

3. Study Area

The dataset is from the National Data Buoy Center (NDBC, https://www.ndbc.noaa.gov/, accessed on 1 November 2024). The NDBC is a branch of the National Oceanic and Atmospheric Administration (NOAA) that plays a crucial role in monitoring and collecting data from a network of buoys and coastal stations across the United States. These buoys are strategically placed in oceans, coastal areas, and Great Lakes to gather real-time environmental data, including information on waves, wind, currents, and other marine conditions. The NDBC provides valuable datasets related to waves. These datasets can include: wave height (the average height of the highest one-third of waves in a given period), wave period (the time it takes for successive wave crests to pass a fixed point), and wave direction (the compass direction from which the waves are coming).

The experiment selected three buoy stations, 41008, 42055, and 46083, for research. Table 1 lists the geographic information of these stations. Figure 5 shows the locations of these buoy stations. The time range for selecting the dataset is from 2018 to 2022, with a collection frequency of once per hour. The data from 2018 to 2021 are selected as the training set, and the data from 2022 are used as the testing set.

Table 1. Geographic information of buoy stations.

Station ID	Latitude	Longitude	Water Depth
41008	31°24′0″ N	80°51′59″ W	16 m
42055	22°8′24″ N	94°6′42″ W	3608 m
46083	58°16′12″ N	138°1′8″ W	129 m

In addition, many values in the dataset are 99.0, which represents missing values that need to be filled or removed during the data preprocessing stage. This process requires the use of Python's Pandas library, which provides rich functions for data reading and preprocessing.

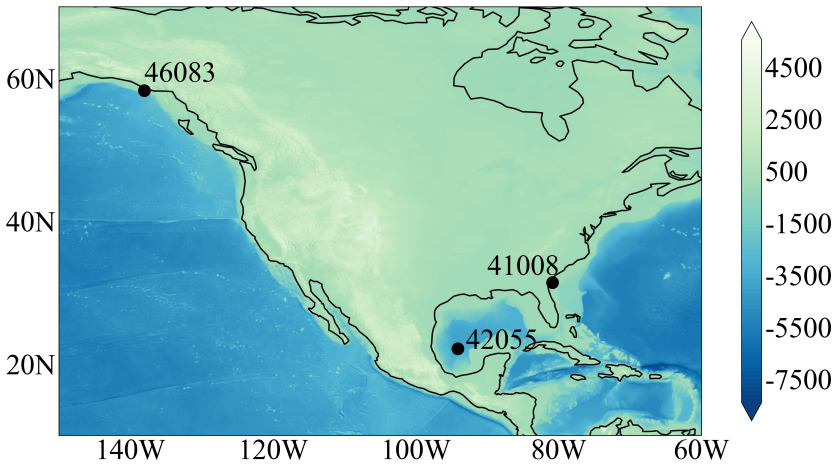


Figure 5. Selected buoy station location.

The model used in this paper are all built using the Pytorch framework. Pytorch is an open-source machine learning library for Python, developed by Facebook's AI Research lab (FAIR). It provides a flexible and dynamic computational graph computation paradigm, which makes it particularly well suited for research and experimentation in deep learning.

In order to evaluate the performance of the model prediction, we also selected scientific evaluation metrics, including root mean square error ($RMSE$), mean absolute error (MAE), and R^2. Their expressions are as follows:

$$\begin{aligned} RMSE &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} \\ MAE &= \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}| \\ R^2 &= 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \end{aligned} \tag{14}$$

In Equation (14), $\widehat{y_i}$ is the observed value (i.e., the true value), y_i is the predicted value of the model, $\bar{y}$ is the average of the observed values, and n represents the total number of samples.

4. Results and Analysis

4.1. Hyperparameter Optimization

The hyperparameters of the TCN are determined using the WOA. Firstly, it is necessary to determine the range of these hyperparameters. The range of batch size is [32, 96], the range of the ADAM optimizer learning rate is [0.001, 0.03], and the range for the numbers of convolution kernels in the TCN is [2, 16]. When initializing the position coordinates of the humpback whale, we randomly generate the positions of 10 humpback whales within the set range as the starting state of the iteration. Subsequently, the whale will calculate the updated direction based on its current position and optimal position, gradually approaching the position of the best fitness.

The iterative changes in the value of the best fitness function are shown in Figure 6. The fitness function iteration curve in Figure 6 was obtained by running on data from the 41008 station, and the fitness function values recorded the changes in $RMSE$ and MAE corresponding to the 1 h prediction. In the initial situation, the $RMSE$ value of the fitness function was 0.109, as indicated by the starting point of the blue line on the graph. This high initial value reflects the model's error before optimization. Over the first few iterations, there is a significant decrease in $RMSE$, which quickly drops to around 0.097 within the first five iterations. This rapid decline demonstrates the effectiveness of the initial optimization steps. After the initial sharp drop, the $RMSE$ curve begins to plateau, showing a more gradual decrease over the subsequent iterations. By the 10th iteration, the $RMSE$ value has decreased to 0.091, marking a 16.5% reduction from the initial value. This point indicates the convergence of the $RMSE$, as further iterations beyond the 10th do not show a significant reduction, suggesting that the model has reached its optimal configuration for the hyperparameters.

Similarly, the MAE value, represented by the red line, starts at approximately 0.072. There is a noticeable reduction in the first few iterations, where the MAE drops to around 0.063. This initial drop highlights the model's improvement in prediction accuracy as the optimization process begins. By the 10th iteration, the MAE has decreased to around 0.059, reflecting an 18% reduction from its initial value. After this point, the MAE also shows convergence, with minimal changes being observed in subsequent iterations.

The optimized parameters are as follows: batch size of 32, learning rate of 0.003, and a number of TCN convolution kernels of 6. The optimization process was supported by a Tesla T4 graphics card. It took approximately 26 h to process 20 epochs of whale optimization. These optimized hyperparameters result in a more accurate and efficient model for prediction. The convergence of both $RMSE$ and MAE after 10 iterations suggests that WOA effectively finds the optimal hyperparameters within a relatively short number of

iterations. This efficiency is crucial for practical applications where computational resources and time are limited.

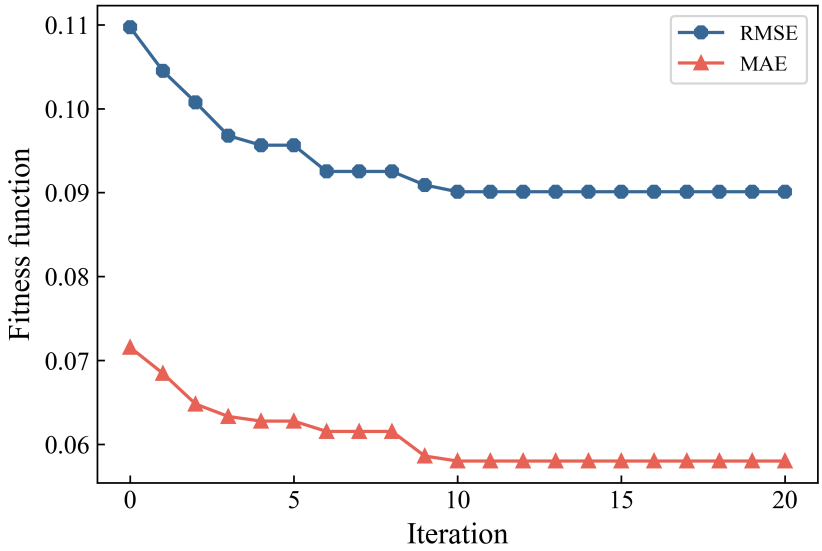


Figure 6. Fitness value iteration curve.

4.2. Model Comparison

After obtaining the optimal hyperparameters through whale optimization algorithm, they were applied to the prediction model ITCN-A and tested at the 41008 site for 1 h, 3 h, and 6 h of prediction results. At the same time, they were compared with the LSTM, TCN, and TCN-Attention models, as shown in Figure 7. In the 1 h prediction, the *RMSE*, *MAE*, and R^2 values predicted by LSTM for wave height are 0.091, 0.07, and 0.954, respectively. However, the LSTM model adopts a temporal structure, and the continuous updating of cell and hidden states can lead to a certain degree of gradient vanishing and gradual disappearance of historical information, which affects the prediction accuracy. The *RMSE* and *MAE* of the TCN model decreased to 0.084 and 0.064, and the R^2 value increased to 0.958. The improvement in accuracy is attributed to the dilated causal convolution structure of the TCN. The dilated causal convolution expands the temporal receptive field to all input sequences, fully extracting the temporal features of the input sequences, avoiding the situation of vanishing gradients. In addition, the residual connection of the TCN can greatly improve the lower limit of prediction accuracy. In order to further enhance the extraction of input features, the TCN and attention mechanisms are coupled. The attention mechanism can extract global temporal features and also avoid the phenomenon of gradient vanishing. In 1 h wave prediction, the *RMSE* and *MAE* of TCN-Attention are 0.079 and 0.059, which are 13.2% and 15.7% lower than LSTM, respectively, and the R^2 value is 0.964, an increase of one percentage point. The traditional TCN model uses the ReLu activation function, which is fixed regardless of the numerical value of the input data and is not flexible enough. Therefore, we switched to dynamic ReLu, where the parameters of the piecewise function are determined by the input data's hyperfunction, making it more flexible and improving accuracy to a certain extent. The *RMSE*, *MAE*, and R^2 values are 0.075, 0.051, and 0.972, respectively, with *RMSE* and *MAE* being reduced by 17.6% and 27.1% compared with the LSTM model. Overall, in terms of the 1 h prediction, there is not much difference in prediction accuracy among the four models, but TCN-DyReLu-Attention has the highest prediction accuracy.

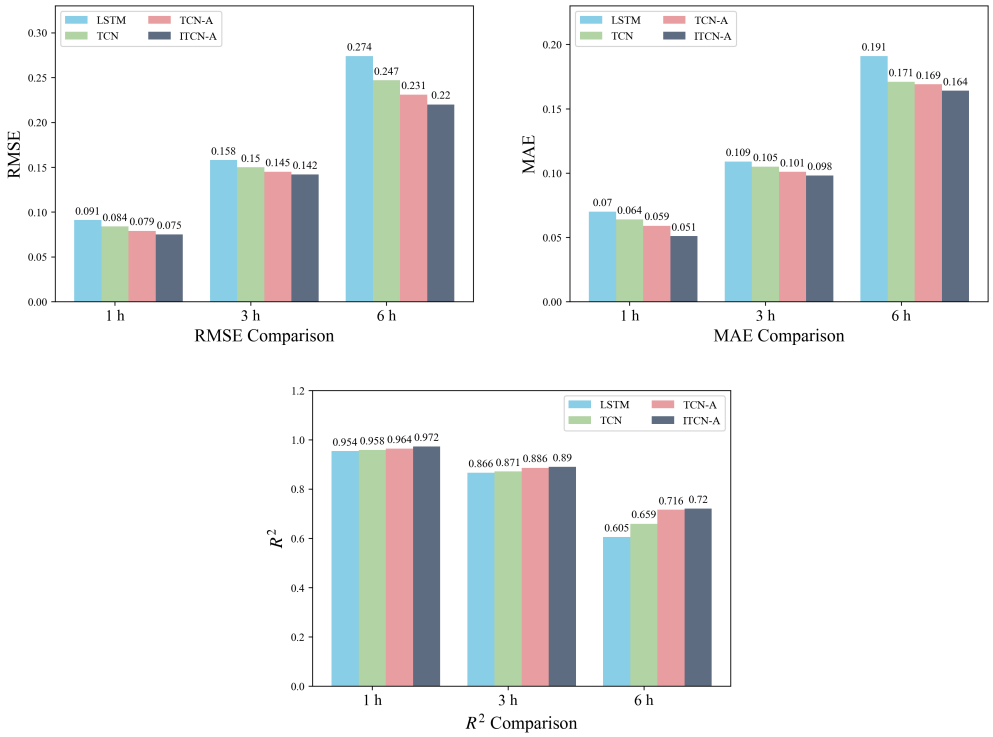


Figure 7. Prediction performance of different models at station 41008.

By increasing the prediction time, in a 3 h prediction, the TCN model compared with LSTM reduced *RMSE* by 5.1%, reaching 0.15, and increased R^2 by 0.5 percentage, reaching 0.871. The error metrics of the TCN and TCN-Attention models are not significantly different, indicating that the TCN model has effectively extracted temporal features, and the improvement of the coupled ensemble model is not significant. Compared with the LSTM model, ITCN-A reduces *RMSE* and *MAE* by 10.1% and 10%, respectively. In the 3 h prediction, there is still not much difference among the four models.

Extending the prediction time to 6 h and conducting tests at the 41008 buoy station, the results showed that the proposed integrated model reduced *RMSE* and *MAE* by 19.7% and 14.1%, respectively, compared with LSTM and increased R^2 by 11.8 percentage points. It can be seen that although the gap between models is small at a low prediction lead time, as the prediction time increases, the leading advantage of the integrated model gradually expands, and the lag of traditional LSTM gradually becomes apparent.

In order to present the model prediction results more vividly, the 1 h predicted values of the ITCN-A, TCN-Attention, TCN, and LSTM models on the 41008 buoy station's test dataset were plotted as continuous curves, as shown in the Figure 8. It can be seen that the LSTM model has a significant deviation between the predicted values and the true values, and overall, it is lower than the true values, indicating a serious underestimation phenomenon, which is difficult to accept in actual ocean prediction. After using the TCN model, the deviation from the true value decreases, and many time points are larger than the true value, resulting in a slight overestimation phenomenon. After incorporating the attention mechanism and dynamic ReLu activation function, the predicted curve of the integrated model is closest to the true value curve and has the lowest error.

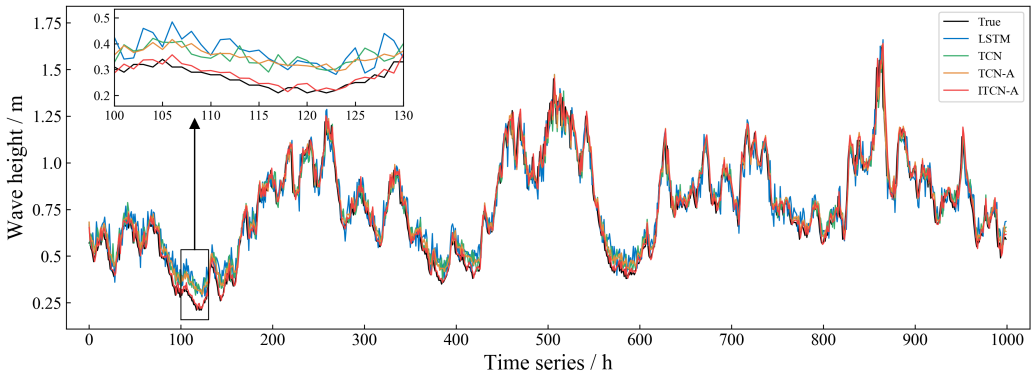


Figure 8. Comparison of 1 h predicted values of different models at station 41008.

At station 41008, the ITCN-A, TCN-Attention, TCN, LSTM models were applied for 3 h prediction, and their predicted values were compared, as shown in the Figure 9. It can be seen that the predicted values of the LSTM model and TCN model differ significantly from the actual values. The attention mechanism to some extent enhances the extraction of input features. Dynamic ReLu further improves the representation ability of the combination model, and its predicted curve is closer to the true value curve. The Figure 10 shows the 6 h prediction curves of different models, and it can be seen that all models have a certain degree of underestimation, with LSTM being the most obvious. Although the integrated model proposed in this article also underestimates the actual value, it is closest to the actual value.

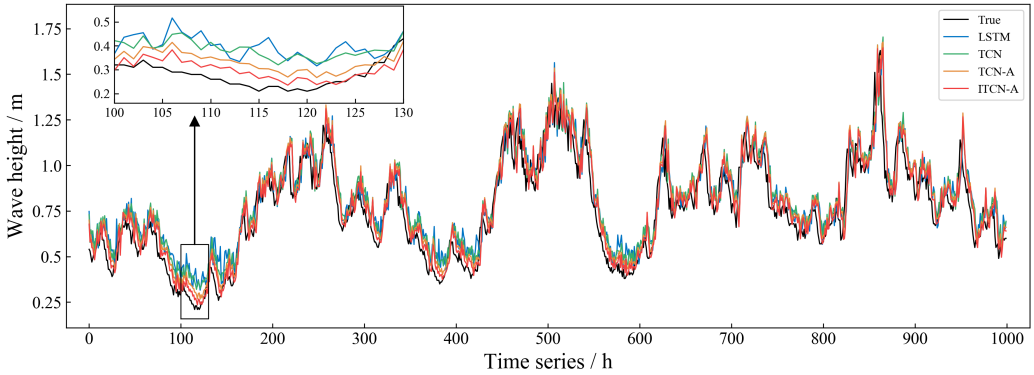


Figure 9. Comparison of 3 h predicted values of different models at station 41008.

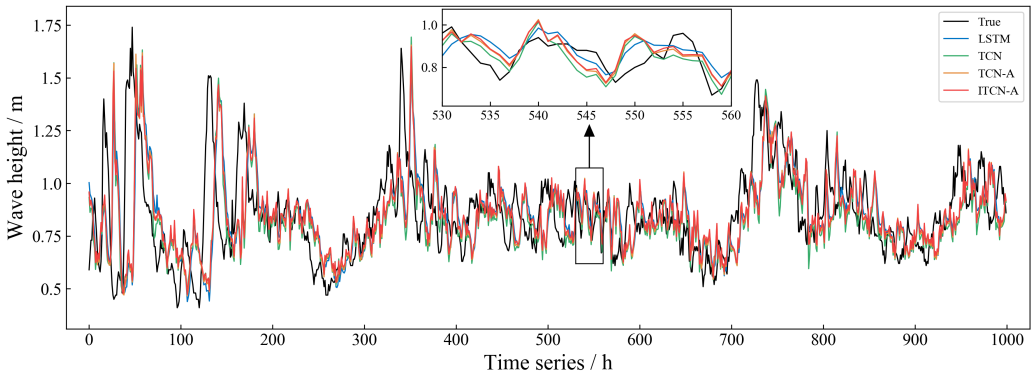


Figure 10. Comparison of 6 h predicted values of different models at station 41008.

4.3. Different Lead Time

Long-term forecasting of significant wave height is crucial for preventing marine disasters; therefore, it is necessary to measure the model's ability to make medium-term and long-term predictions. At station 41008, the LSTM, TCN, TCN-Attention and ITCN-A models were constructed to predict the wave heights for the next 12 h, 18 h, and 24 h. The variation of various error metrics with the prediction time is shown in Figure 11. From Figure 11, it can be seen that in terms of 12 h, 18 h, and 24 h prediction, the *RMSE* and *MAE* metrics of the LSTM model are slightly higher than those of the TCN model. Specifically, for a 12 h lead time, the *RMSE* of the LSTM model is around 0.32, while the TCN model achieves a slightly lower *RMSE* of approximately 0.31. For *MAE*, the LSTM model starts at around 0.07 and the TCN model at 0.063. This trend continues as the lead time increases, highlighting the consistent performance of the TCN over LSTM for medium-term forecasts.

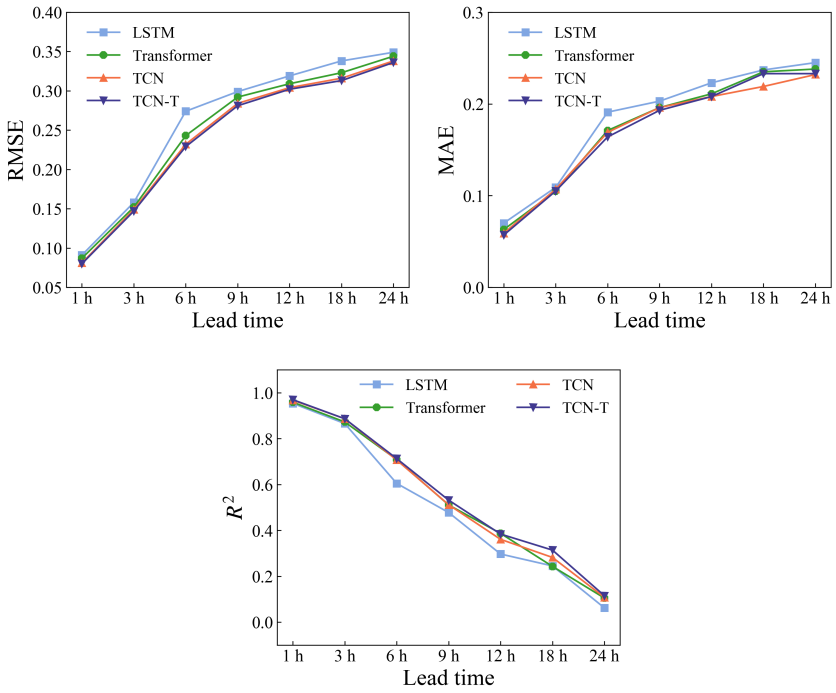


Figure 11. Long-term predictive performance of different models.

As the prediction horizon extends to 18 h and 24 h, the benefits of integrating the attention mechanism and using the dynamic ReLU activation function become more evident. The *RMSE* and *MAE* for the TCN-Attention and ITCN-A models show a noticeable reduction compared with the plain TCN model. For instance, at the 24 h prediction mark, the ITCN-A model achieves an *RMSE* of 0.336 and an *MAE* of 0.233, significantly lower than the corresponding values for the LSTM model.

Moreover, we examined the R^2 metric, which measures the proportion of variance explained by the model. For short-term predictions (1 h to 6 h), all models, including LSTM, TCN, and its variants, exhibit high R^2 values close to 1.0, indicating strong predictive power. However, as the lead time increases to 12 h, 18 h, and 24 h, the R^2 values decrease, reflecting the increasing difficulty of long-term forecasting. Notably, the ITCN-A model shows a higher R^2 compared with the LSTM model, especially for the 24 h lead time, where it achieves an R^2 improvement of 5.3 percentage points over the LSTM model.

This enhancement indicates that the attention mechanism can effectively extract long-term sequence features, and the dilated convolution of TCN, which has a global receptive field, enhances the model's long-term prediction ability. The integration of dynamic ReLU further refines the activation function, providing better adaptability and improving the model's performance on longer lead times.

4.4. Multi-Station Analysis

In order to verify the universality of the model, it is necessary to apply each model to multiple different buoy stations. The significant wave height data from stations 42055 and 46083 were selected for model testing, covering 1 h, 3 h, 6 h, 9 h, 12 h, 18 h, and 24 h predictions. Based on the predicted and true values, scatter charts can be drawn, as shown in Figure 12.

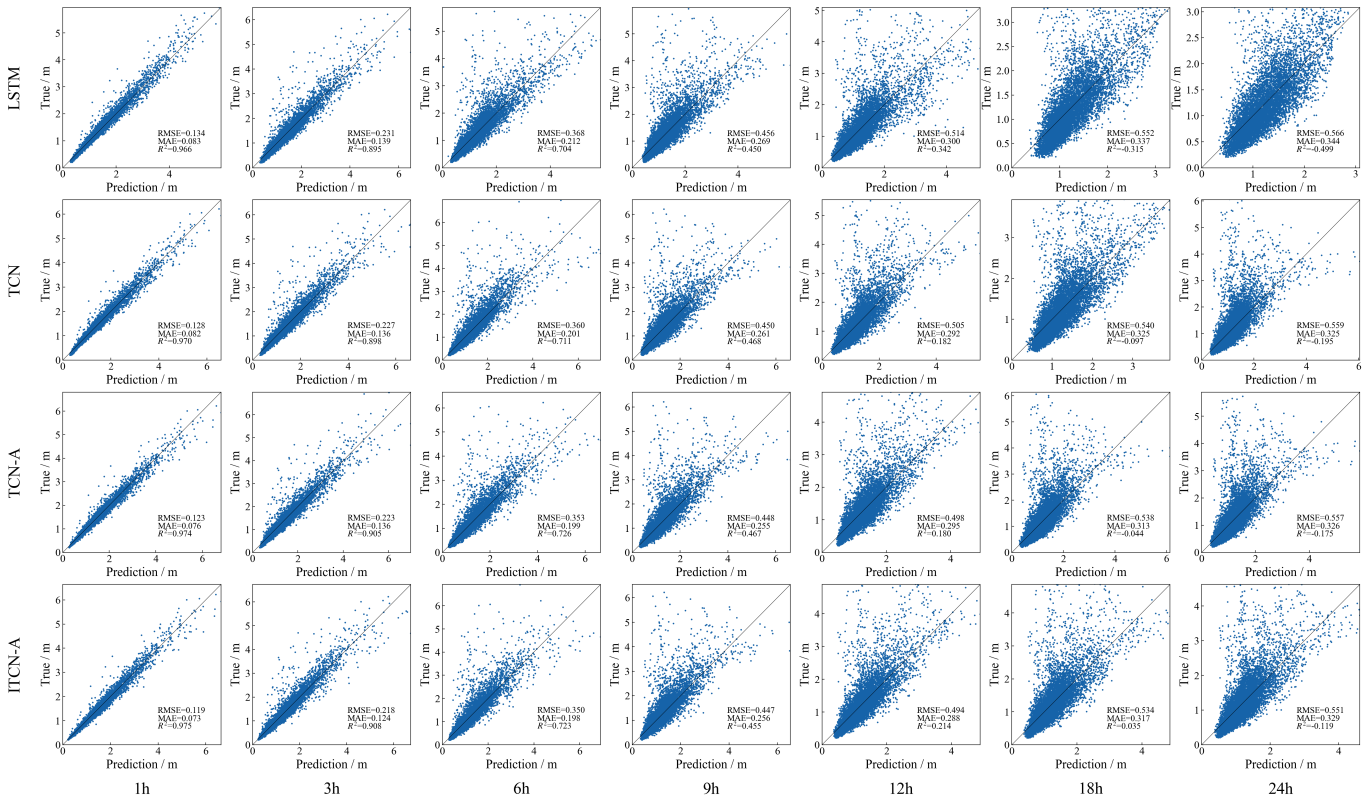


Figure 12. Scatter plot of station 42055's prediction results.

The chart shows the scatter plots of predicted values for different models at different lead times at station 42055. From the chart, it can be seen that as the lead time increases, the prediction accuracy continuously decreases. The LSTM model has the largest error, and the scatter points are more dispersed compared with other models. This is evident in the 24 h lead time, where the LSTM's scatter points deviate significantly from the $y = x$ line, indicating a higher prediction error ($RMSE$ = 0.566, MAE = 0.344, R^2 = 0.499). The TCN model effectively improves prediction accuracy, thanks to the global receptive field of dilated causal convolution. Even at longer lead times, scatter points are concentrated around the $y = x$ line, though there is a noticeable underestimation trend. For instance, at the 24 h lead time, TCN's scatter points are closer to the $y = x$ line compared with LSTM, with improved $RMSE$ and MAE values ($RMSE$ = 0.559, MAE = 0.325, R^2 = 0.195). This demonstrates the model's ability to maintain better accuracy over extended prediction horizons.

On the basis of the TCN model, coupling it with the attention mechanism and using dynamic ReLU activation further refines the prediction accuracy. The TCN-Attention and ITCN-A models exhibit scatter distributions that are closest to the $y = x$ line across all lead times. For example, in the 24 h prediction, the ITCN-A model achieves $RMSE$ and MAE values of 0.551 and 0.329, respectively, indicating the lowest error among the models tested. This improvement is attributed to the attention mechanism's ability to effectively extract long-term dependencies in the data and the adaptability of the dynamic ReLU at handling different input sequences.

Next, we analyze station 46083, and its scatter plot is shown in Figure 13. In the 1 h prediction, all four models performed well, with R^2 values around 0.97, indicating high accuracy for short-term forecasts. However, as the prediction horizon extends, the LSTM model exhibits a degree of overestimation, likely due to issues related to gradient vanishing and exploding. These issues cause instability in input features during the chain calculation process, leading to inaccurate predictions. In contrast, the TCN adopts residual connections, effectively mitigating the problems of vanishing or exploding gradients. From Figure 13, it can be observed that the scatter points predicted by the TCN model are more symmetrically distributed around the $y = x$ line. The TCN-Attention and ITCN-A models further enhance prediction accuracy. For example, in the 18 h prediction, the ITCN-A model achieves $RMSE$ and MAE values of 0.379 and 0.281, respectively, which are 16.2% and 6.1% lower than the corresponding metrics of the LSTM model.

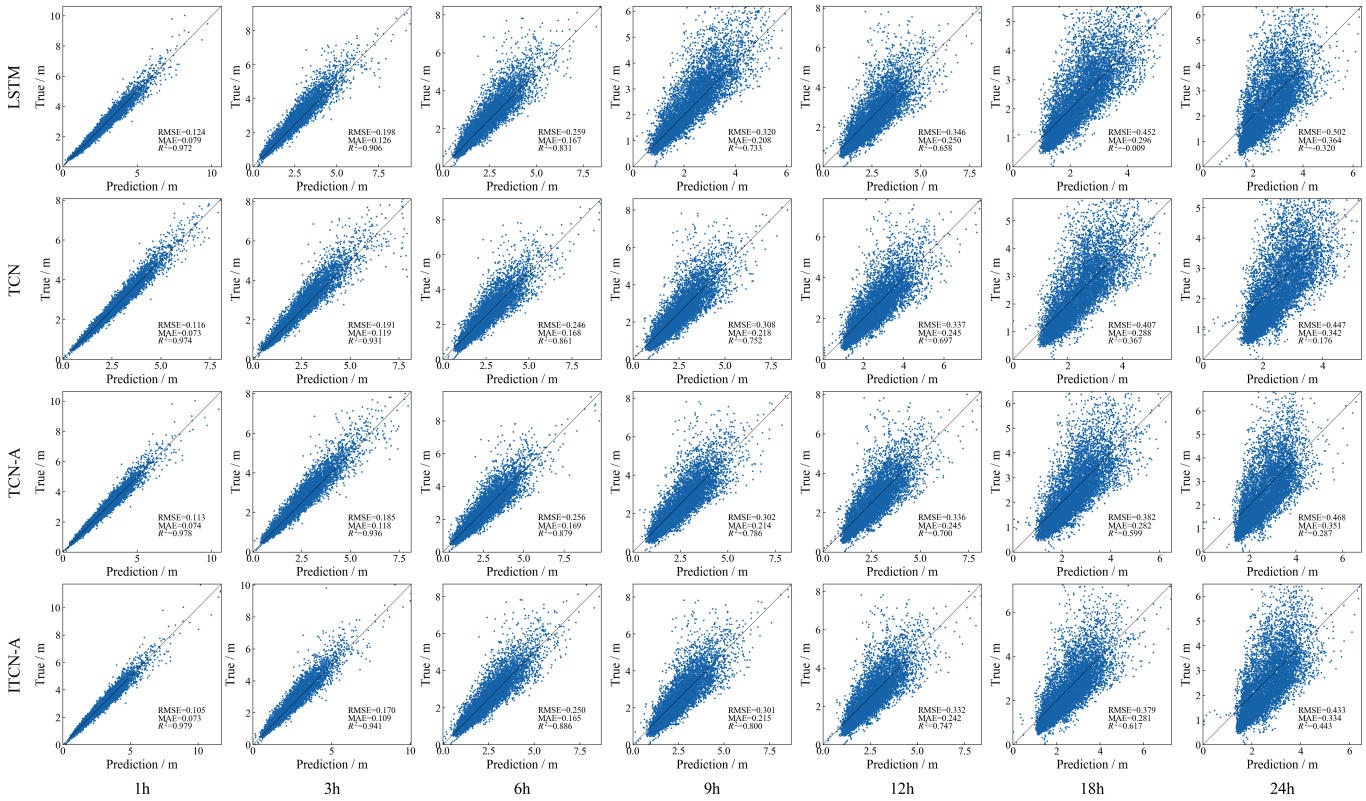


Figure 13. Scatter plot of station 46083's prediction results.

5. Conclusions

An ITCN-A prediction model is proposed to address the difficulty of global feature extraction in traditional deep learning models. Firstly, the hyperparameters of the TCN model are optimized using the WOA, and the activation function in the TCN adopts dynamic ReLu. Then, the TCN and attention mechanism are coupled to achieve the prediction of wave height. The following conclusions are drawn:

(1) With the help of dilated causal convolution and residual connections, the TCN model can effectively avoid the problem of vanishing or exploding gradients. In multi-layer chain computing, input features can also be fully preserved. Adopting a dynamic ReLu activation function makes the model representation more flexible.

(2) The introduction of the WOA can adaptively determine the hyperparameters of the TCN, with fast convergence speed and effective avoidance of getting stuck in local optima.

(3) By using the attention mechanism to dynamically allocate the weights of input features, the ability to capture global features is improved, and the performance of the model is improved.
(4) The proposed ITCN-A prediction model can effectively reduce prediction errors and performs well in the three metrics of $RMSE$, MAE, and R^2. In the testing of multiple buoy stations, the model has improved the prediction accuracy of wave height and demonstrated good generalization ability. At the same time, different prediction time steps were tested, and it was found that the proposed combination model also has strong long-term prediction ability. In future research, we will integrate multiple data sources such as satellites and ocean buoys to further improve prediction accuracy.

Author Contributions: Conceptualization, Y.H. and J.T.; methodology, Y.H.; software, J.T.; validation, Y.H., J.T., H.J. and R.Z.; formal analysis, Y.H.; investigation, Y.H.; resources, Y.H.; data curation, J.T.; writing—original draft preparation, Y.H.; writing—review and editing, Y.H.; visualization, J.T.; supervision, C.D.; project administration, C.D.; funding acquisition, C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (2023YFC3008203) and the National Natural Science Foundation of China (62076136).

Data Availability Statement: The data used in this study come from NDBC, whose website is: https://www.ndbc.noaa.gov/ (accessed on 1 November 2024).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Intergovernmental Panel on Climate Change. *Climate Change*; World Meteorological Organization: Geneva, Switzerland, 2007; Volume 52, pp. 1–43.
2. Shuto, N.; Fujima, K. A short history of tsunami research and countermeasures in Japan. *Proc. Jpn. Acad. Ser. B* **2009**, *85*, 267–275. [CrossRef] [PubMed]
3. Chen, C. Case study on wave-current interaction and its effects on ship navigation. *J. Hydrodyn.* **2018**, *30*, 411–419. [CrossRef]
4. Van Dongeren, A.; De Jong, M.; Van der Lem, C.; Van Deyzen, A.; Den Bieman, J. Review of long wave dynamics over reefs and into ports with implication for port operations. *J. Mar. Sci. Eng.* **2016**, *4*, 12. [CrossRef]
5. Mantari, J.; Ribeiro e Silva, S.; Soares, C.G. Intact stability of fishing vessels under combined action of fishing gear, beam waves and wind. *Ocean Eng.* **2011**, *38*, 1989–1999. [CrossRef]
6. Joerger, M.; Spenko, M. Towards Navigation Safety for Autonomous Cars. Inside GNSS 2017. Available online: https://par.nsf.gov/biblio/10070277 (accessed on 1 November 2024).
7. Liu, C.P.; Liang, G.S.; Su, Y.; Chu, C.W. Navigation safety analysis in Taiwanese ports. *J. Navig.* **2006**, *59*, 201–211. [CrossRef]
8. Abouhalima, M.; das Neves, L.; Taveira-Pinto, F.; Rosa-Santos, P. Machine Learning in Coastal Engineering: Applications, Challenges, and Perspectives. *J. Mar. Sci. Eng.* **2024**, *12*, 638. [CrossRef]
9. Berg, H. Human factors and safety culture in maritime safety. In *Marine Navigation and Safety of Sea Transportation: STCW, Maritime Education and Training (MET), Human Resources and Crew Manning, Maritime Policy, Logistics and Economic Matters*; CRC Press: Boca Raton, FL, USA, 2013; Volume 107, pp. 107–115.
10. Falcao, A.F.d.O. Wave energy utilization: A review of the technologies. *Renew. Sustain. Energy Rev.* **2010**, *14*, 899–918. [CrossRef]
11. Curto, D.; Franzitta, V.; Guercio, A. Sea wave energy. A review of the current technologies and perspectives. *Energies* **2021**, *14*, 6604. [CrossRef]
12. Rusu, E. Evaluation of the wave energy conversion efficiency in various coastal environments. *Energies* **2014**, *7*, 4002–4018. [CrossRef]
13. Guillou, N.; Lavidas, G.; Chapalain, G. Wave energy resource assessment for exploitation—A review. *J. Mar. Sci. Eng.* **2020**, *8*, 705. [CrossRef]
14. Denny, M.W. Ocean waves, nearshore ecology, and natural selection. *Aquat. Ecol.* **2006**, *40*, 439–461. [CrossRef]
15. Group, T.W. The WAM model—A third generation ocean wave prediction model. *J. Phys. Oceanogr.* **1988**, *18*, 1775–1810. [CrossRef]
16. Booij, N.; Ris, R.C.; Holthuijsen, L.H. A third-generation wave model for coastal regions: 1. Model description and validation. *J. Geophys. Res. Ocean.* **1999**, *104*, 7649–7666. [CrossRef]
17. Tolman, H.L. A third-generation model for wind waves on slowly varying, unsteady, and inhomogeneous depths and currents. *J. Phys. Oceanogr.* **1991**, *21*, 782–797. [CrossRef]

18. Londhe, S.; Shah, S.; Dixit, P.; Nair, T.B.; Sirisha, P.; Jain, R. A coupled numerical and artificial neural network model for improving location specific wave forecast. *Appl. Ocean Res.* **2016**, *59*, 483–491. [CrossRef]
19. Allen, M.R.; Kettleborough, J.; Stainforth, D. Model error in weather and climate forecasting. In Proceedings of the ECMWF Predictability of Weather and Climate Seminar, Reading, UK, 9–13 September 2002; pp. 279–304.
20. Li, G.; Li, R.; Hou, H.; Zhang, G.; Li, Z. A Data-Driven Motor Optimization Method Based on Support Vector Regression—Multi-Objective, Multivariate, and with a Limited Sample Size. *Electronics* **2024**, *13*, 2231. [CrossRef]
21. Kumar, N.K.; Savitha, R.; Al Mamun, A. Ocean wave height prediction using ensemble of extreme learning machine. *Neurocomputing* **2018**, *277*, 12–20. [CrossRef]
22. Wang, Q.; Wang, H.; Gupta, C.; Rao, A.R.; Khorasgani, H. A non-linear function-on-function model for regression with time series data. In Proceedings of the 2020 IEEE International Conference on Big Data, Atlanta, GA, USA, 10–13 December 2020; pp. 232–239.
23. Yang, J.; Kim, J.; Ryu, H.; Lee, J.; Park, C. Predicting Car Rental Prices: A Comparative Analysis of Machine Learning Models. *Electronics* **2024**, *13*, 2345. [CrossRef]
24. Zilong, T.; Yubing, S.; Xiaowei, D. Spatial-temporal wave height forecast using deep learning and public reanalysis dataset. *Appl. Energy* **2022**, *326*, 120027. [CrossRef]
25. Deo, M.C.; Jha, A.; Chaphekar, A.; Ravikant, K. Neural networks for wave forecasting. *Ocean Eng.* **2001**, *28*, 889–898. [CrossRef]
26. Sadeghifar, T.; Nouri Motlagh, M.; Torabi Azad, M.; Mohammad Mahdizadeh, M. Coastal wave height prediction using Recurrent Neural Networks (RNNs) in the south Caspian Sea. *Mar. Geod.* **2017**, *40*, 454–465. [CrossRef]
27. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
28. Hao, W.; Sun, X.; Wang, C.; Chen, H.; Huang, L. A hybrid EMD-LSTM model for non-stationary wave prediction in offshore China. *Ocean Eng.* **2022**, *246*, 110566. [CrossRef]
29. Jörges, C.; Berkenbrink, C.; Stumpe, B. Prediction and reconstruction of ocean wave heights based on bathymetric data using LSTM neural networks. *Ocean Eng.* **2021**, *232*, 109046. [CrossRef]
30. Hu, H.; van der Westhuysen, A.J.; Chu, P.; Fujisaki-Manome, A. Predicting Lake Erie wave heights and periods using XGBoost and LSTM. *Ocean Model.* **2021**, *164*, 101832. [CrossRef]
31. Fan, S.; Xiao, N.; Dong, S. A novel model to predict significant wave height based on long short-term memory network. *Ocean Eng.* **2020**, *205*, 107298. [CrossRef]
32. Bai, L.H.; Xu, H. Accurate estimation of tidal level using bidirectional long short-term memory recurrent neural network. *Ocean Eng.* **2021**, *235*, 108765. [CrossRef]
33. Jing, Y.; Zhang, L.; Hao, W.; Huang, L. Numerical study of a CNN-based model for regional wave prediction. *Ocean Eng.* **2022**, *255*, 111400. [CrossRef]
34. Huang, L.; Jing, Y.; Chen, H.; Zhang, L.; Liu, Y. A regional wind wave prediction surrogate model based on CNN deep learning network. *Appl. Ocean Res.* **2022**, *126*, 103287. [CrossRef]
35. Ni, C.; Ma, X. Prediction of wave power generation using a convolutional neural network with multiple inputs. *Energies* **2018**, *11*, 2097. [CrossRef]
36. Guan, X. Wave height prediction based on CNN-LSTM. In Proceedings of the 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 23–25 October 2020; pp. 10–17.
37. Zhang, J.; Xin, X.; Shang, Y.; Wang, Y.; Zhang, L. Nonstationary significant wave height forecasting with a hybrid VMD-CNN model. *Ocean Eng.* **2023**, *285*, 115338. [CrossRef]
38. Ji, Q.; Han, L.; Jiang, L.; Zhang, Y.; Xie, M.; Liu, Y. Short-term prediction of the significant wave height and average wave period based on the variational mode decomposition–temporal convolutional network–long short-term memory (VMD–TCN–LSTM) algorithm. *Ocean Sci.* **2023**, *19*, 1561–1578. [CrossRef]
39. 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.* **2023**, *283*, 114938. [CrossRef]
40. Lou, R.; Lv, Z.; Guizani, M. Wave height prediction suitable for maritime transportation based on green ocean of things. *IEEE Trans. Artif. Intell.* **2022**, *4*, 328–337. [CrossRef]
41. Zheng, C.; Fan, X.; Wang, C.; Qi, J. Gman: A graph multi-attention network for traffic prediction. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 1234–1241. [CrossRef]
42. Ma, Z.; Mei, G. A hybrid attention-based deep learning approach for wind power prediction. *Appl. Energy* **2022**, *323*, 119608. [CrossRef]
43. Zhang, H.; Yan, J.; Liu, Y.; Gao, Y.; Han, S.; Li, L. Multi-source and temporal attention network for probabilistic wind power prediction. *IEEE Trans. Sustain. Energy* **2021**, *12*, 2205–2218. [CrossRef]
44. Qian, R.; Ding, Y. An Efficient UAV Image Object Detection Algorithm Based on Global Attention and Multi-Scale Feature Fusion. *Electronics* **2024**, *13*, 3989. [CrossRef]
45. Zhang, B.; Wang, S.; Deng, L.; Jia, M.; Xu, J. Ship motion attitude prediction model based on IWOA-TCN-Attention. *Ocean Eng.* **2023**, *272*, 113911. [CrossRef]
46. Luo, Q.R.; Xu, H.; Bai, L.H. Prediction of significant wave height in hurricane area of the Atlantic Ocean using the Bi-LSTM with attention model. *Ocean Eng.* **2022**, *266*, 112747. [CrossRef]

47. Bai, S.; Kolter, J.Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv* **2018**, arXiv:1803.01271.
48. Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; Liu, Z. Dynamic relu. In Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 351–367.
49. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
50. Mirjalili, S.; Lewis, A. The whale optimization algorithm. *Adv. Eng. Softw.* **2016**, *95*, 51–67. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.