Prediction of significant wave height based on feature decomposition and enhancement

Prediction of significant wave height based on feature decomposition and enhancement

Lin Wu [a], Yi An [a,b,c], Pan Qin [a], Huo-Sheng Hu [d]

[a] *School of Control Science and Engineering, Dalian University of Technology, Dalian 116023, China*
[b] *Key Laboratory of Intelligent Control and Optimization for Industrial Equipment of Ministry of Education, Dalian University of Technology, Dalian 116023, China* [c] *School of Electrical Engineering, Xinjiang University, Urumqi 830046, China*
[d] *School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, UK*

A B S T R A C T

Predicting significant wave height (SWH) are crucial for maritime activities, including offshore operations, ship navigation, and meteorological forecasting. However, the complexity, non-stationarity, and distribution shifts of SWH result in relatively low prediction accuracy. Additionally, the inadequate use of local information in many prediction models further hinders accuracy improvements. To solve these problems, this paper proposes a novel multimodal feature enhancement transformer (MFET) method for SWH prediction. The method primarily consists of a signal decomposition module, an encoder stack, and a decoder stack. The signal decomposition module uses the sparrow search algorithm-variational mode decomposition (SSA-VMD) method to optimally decompose SWH signals. The decomposed signals are combined with wave features to form 3D data, which is then input into the encoder and decoder stacks for prediction. Each stack contains six encoders and six decoders respectively. Each encoder comprises a squeeze-and-excitation (SE) attention module, a multi-head convolutional attention (MHCA) module, and a multi-layer perceptron (MLP) module, while each decoder includes a multi-head self-attention (MHSA) module, a cross-attention (CA) module, and an MLP module. The SE attention mechanism dynamically adjusts the influence of each channel by selectively enhancing or suppressing their contributions. A parallel convolution layer is proposed in MHCA to effectively capture local wave feature within each channel. Furthermore, the reversible instance normalization (RevIN) method is used to eliminate distribution shifts. The MFET improves prediction accuracy by optimally decomposing the SWH signal, dynamically enhancing channel information, and extracting local features in parallel. Experimental results show that MFET achieves MSE of 0.0062, 0.0019, and 0.0073, along with R^2 of 98.51%, 98.93%, and 95.09% on the three datasets. Code is available at this repository: https://github.com/wulin777/SWH-Prediction

Keywords:
Prediction of significant wave height Sparrow search algorithm-variational mode decomposition; Squeeze-excitation attention; Parallel convolution

1. Introduction

Significant wave height (SWH) is the average height of the highest one-third of the wave heights observed in a given period (Hashim et al., 2016), which is a crucial ocean wave feature. Generally, buoys are main instruments used for measuring wave features, such as wave height, wave period, and wave direction. These wave features provide a data foundation for the prediction of SWH. Accurate prediction of SWH are essential for ensuring navigation safety, supporting offshore operations, and enhancing meteorological predictions (Clauss, 2002; Foster et al., 2014).

The primary methods for predicting SWH currently are divided into two categories (Huang & Cui, 2023): traditional models prediction methods and intelligent models prediction methods. Traditional prediction methods typically rely on statistical models or numerical simulations of physical processes (Agrawal & Deo, 2002; Kumar et al., 2015). However, these methods, which are based on establishing mathematical models or simulating physical processes, often rely too heavily on statistical models and have limited generalization capabilities. Therefore, these methods cannot further enhance prediction accuracy.

In recent years, the rapid advancement of artificial intelligence has brought traditional machine learning-based approaches for predicting SWH time series to the forefront. Models like the ridge regression model, k-nearest neighbors (kNN), artificial neural networks (ANN), and support vector machines (SVM) have been extensively applied in SWH prediction (Berbić et al., 2017; Chowdary et al., 2023; Domala &

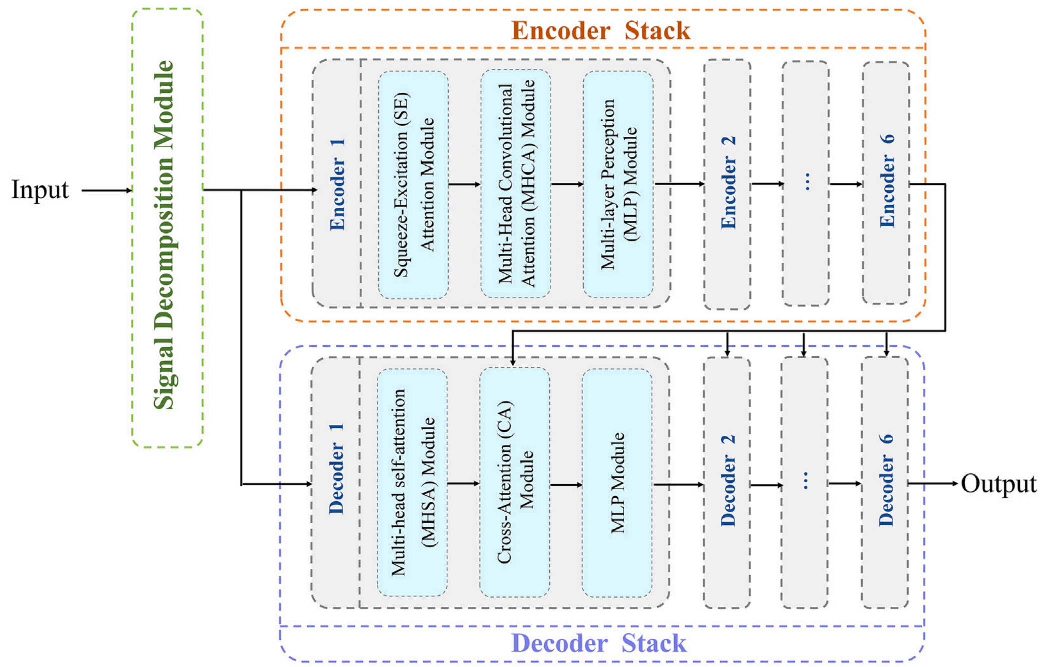


Fig. 1. Structure of the MFET model.

Kim, 2022). Despite their widespread use, these methods increasingly suffer from predicted inaccuracies due to their limitations in data collection and reliance on manual feature extraction. Consequently, researchers have turned their attention towards deep learning. Prediction models, such as the convolutional neural networks (CNN), long short-term memory networks (LSTM), recurrent neural networks (RNN), and sequence to sequence (S2S) have been widely used. These methods have contributed to advancing long sequence time series forecasting (Minuzzi & Farina, 2023; Raj & Prakash, 2024; Wu et al., 2024; Zhang et al., 2024). Especially the transformer model (Vaswani et al., 2023), which can effectively understand the complex spatiotemporal relationships in time series due to its valuable attention mechanism. However, early deep learning models face challenges in processing the nonlinear and non-stationary characteristics of SWH signals, particularly in capturing local temporal dependencies and multi-scale patterns, which limits further advancements in prediction accuracy.

To solve the problems of non-stationarity in SWH signals and the neglect of local features, we propose the multimodal feature enhancement transformer (MFET) method. MFET (see Fig. 1) is primarily composed of three modules: a signal decomposition module, an encoder stack, and a decoder stack. The first module is used for processing signals, the encoder and decoder are used for prediction. To achieve more regular mode decomposition, we apply variational mode decomposition (VMD) to break down the original SWH signal into multimodal signals. During this stage, the sparrow search algorithm (SSA) is used to determine the optimal parameters for effective decomposition. Each decomposed signal is concatenated with other wave features to form 3D data for prediction. In the encoder, considering the different importance of each channel in the 3D data, we use a squeeze-excitation (SE) attention mechanism to selectively enhance and suppress them. Furthermore, to ensure the model captures both global and local features, we propose the multi-head convolutional attention (MHCA) module, which involves adding parallel convolutional layer prior to the multi-head self-attention module. Additionally, the use of the reversible instance normalization (RevIN) effectively addresses the temporal distribution shifts caused by the prolonged data collection process. The main contributions of the research are as follows:

- The novel MFET is proposed to accurately predict SWH. By decomposing the SWH signal and enhancing wave features, the model effectively improves prediction accuracy.
- SSA-VMD decomposes the SWH signal into more regular and predictable components. The components are concatenated and reconstructed into 3D data, which enables the first successful prediction of 3D time series.
- SE attention is used for the first time to weight the components of the decomposed SWH signal, selectively emphasizing key components while suppressing less contributive ones to enhance prediction.
- The MHCA is proposed for the first time, which enhances the understanding of both local and global wave features through the addition of parallel convolutions.

Although our model has shown significant improvement in prediction accuracy and its robustness has been validated on three different datasets, the data preparation phase prior to prediction is time-consuming. This is due to the need for optimized decomposition during the construction of the 3D dataset. Additionally, the model comprises a substantial parameter set. We will work on addressing these issues in future research.

The rest of this paper is organized as follows: Section 2 presents a brief literature review of existing methods. Section 3 presents the process of wave feature measurement and processing. Section 4 provides a detailed introduction to the proposed MFET model. Section 5 introduces the setup of experiment evaluation and analyzes the experimental results. Section 6 provides the conclusion and future work.

2. Related work

The main prediction methods in marine field are divided into two categories: (1) traditional models prediction methods, (2) intelligent models prediction methods.

2.1. Traditional models prediction methods

Traditional wave prediction methods are divided into two approaches: physics-based models, which solve hydrodynamic equations through numerical simulations, and statistical models, which use historical data to infer wave characteristics probabilistically. The numerical simulation methods based on physical models simulate the propagation and evolution of waves by solving basic physical equations describing wave motion. For example, Li et al. (2021) used the weather research and forecasting (WRF) model to conduct a sensitivity experiment on the marine wind model in the Baltic Sea area. They tested different settings in the model to see how well its predictions matched real data. However, the WRF model's accuracy depends on the settings chosen and the detail of its data. Its natural uncertainty also makes its predictions less reliable. Rekha Sankar and Panchapakesan (2024) combined mathematical statistics, optimization algorithms, and signal processing methods to predict wind speed. Katalinić and Parunov (2020) predicted extreme waves using initial distribution, extreme value, and peak threshold methods. They improved their predictions by fitting math models to the data. They tried these methods on data from the Adriatic Sea. Similarly, Orimolade et al. (2016) used these methods to study extreme waves in the NORA10 database. But these methods rely too much on data and mistakes can add up, making their predictions not very accurate.

2.2. Intelligent models prediction methods

To further enhance prediction accuracy, machine learning has been applied to ocean related prediction tasks. Domala and Kim (2022) utilized various models such as ridge regression, SVM, and kNN to predict SWH using multiple wave features, including sea surface temperature, mean period, and wind speed. Etemad-Shahidi and Mahjoobi (2009) compared M5 model trees and neural networks for predicting SWH in lakes. Duan et al. (2016) employed empirical mode decomposition (EMD) in conjunction with SVM for the short-term SWH prediction. Additionally, Demetriou et al. (2021) analyzed various machine learning methods for SWH prediction and proposed an improved ANN method. However, these models often need to tackle issues such as heavy computational demands that slow down performance, high sensitivity to input data which undermines their generalization capabilities, the inability to avoid linear assumptions and strong reliance on handcrafted features.

Modern deep learning models address these limitations by automatic feature learning, which eliminates the need for manual feature engineering. By using attention mechanisms, these models adaptively focus on key patterns in raw data while suppressing noise. Furthermore, techniques like dropout and batch normalization enhance robustness to input variations, improving generalization across diverse marine environments. For example, Pan et al. (2024) proposed a method that uses Conv-LSTM as the unit structure and adopts an encoder–decoder framework to predict global sea surface temperature. However, such models using a recurrent structure tend to gradually weaken in their ability to capture long-range dependencies when handling long time series. This means that the model performs poorly when dealing with excessively long data. Moreover, Obakrim et al. (2023) utilized a CNN-LSTM model to explore the spatiotemporal relationship between SWH, but it struggled with the non-stationary nature of the data. Wang et al. (2024) proposed a SWH prediction system that integrates feature extraction, model selection, and weight optimization. However, the prediction models used in the system are basic, which limits the accuracy. Additionally, the system's complexity leads to low operational efficiency. Wu et al. (2023) combined human cognition with deep models to predict wave height. However, the article only implements single-step prediction, which limits its practical value. Liang et al. (2024) used graph attention network (GAT) and transformer to predict the natural climate phenomenon ENSO. Daniel and Adytia (2023) utilized the transformer to predict significant wave height. However, transformer faces the issue of overlooking the relationships between local features.

Overall, existing models face significant challenges in simultaneously capturing the intricate relationships between local and global features. Furthermore, for large, complex, and highly non-stationarity datasets, these models often fail to achieve further improvements in prediction accuracy. We propose the MFET to solve the problems in current research. Through feature decomposition and enhancement, the method demonstrates good predictive accuracy and performance.

Table 1
Details of measurement.

Dataset	Interval	Total number	SWH related features
Australia	0.5 h	16 759	*Hmax*, *Tp*, *Tz*, *Drip*, *Tsea*
Ours	3 min	28 756	*Ss*, *Tp*, *Tz*, *Tdw1*, *Qp*
NDBC	0.5 h	12 620	*WSPD*, *GST*, *DPD*, *APD*, *MWD*

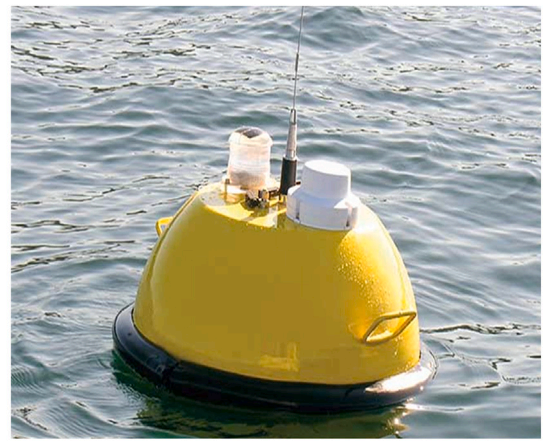


Fig. 2. The MK III buoy.

3. SWH related features

In our study, we used three distinct datasets to validate the robustness of the model for predicting SWH, as detailed in Table 1:

(1) Australia dataset: This dataset was collected using a Waverider buoy off the Gold Coast, Australia. The sampling interval was 0.5 h (in meters). The data collection period was from 00:00 on 01/01/2021 to 23:30 on 31/08/2022, with a total of 16,759 data points.

(2) Our dataset: This dataset was collected using a MK III buoy (Vries et al., 2003) (see Fig. 2) in the Yellow Sea, China (38°51′N, 121°39′E). The sampling interval was 3 min (in meters). The data was collected from 08:00 on 26/02/2023 to 17:59 on 09/05/2023, with a total of 28,756 data points.

(3) NDBC dataset: The data were collected using large moored buoys from the National Data Buoy Center (NDBC), with a 0.5-h sampling interval (in meters). The data covers two periods: from 00:40 on 01/01/2020 to 00:40 on 15/09/2020, and from 00:10 on 01/01/2023 to 23:40 on 31/05/2023, totaling 12,620 data points.

In order to accurately analyze the trend dynamics of SWH and other wave features, Pearson analysis (Benesty et al., 2009) and preprocessing were conducted on the data. The calculation process of Pearson analysis is expressed in Eq. (1):

$$\begin{aligned}\rho_{x,y} &= \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y} \\ &= \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \cdot \sigma_Y} \\ &= \frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\cdot\sqrt{E(Y^2)-E^2(Y)}}\end{aligned} \tag{1}$$

where, $\operatorname{cov}(x,y)$ is the covariance of variables x and y; σ_x and σ_y are the standard deviations of variables x and y, respectively. E denotes the expectation; μ_X and μ_Y are the means of variables X and Y, respectively.

Through this approach, we selected the five most suitable features from each dataset as inputs for the model:

(1) The features of Australia dataset: *Hmax*, *Tp*, *Tz*, *Drip*, and *Tsea*. *Hmax* is the maximum wave height. *Tp* is the peak wave period, which is the period of the maximum energy density in the wave energy spectrum. *Tz* is the average zero-crossing period, which is the average

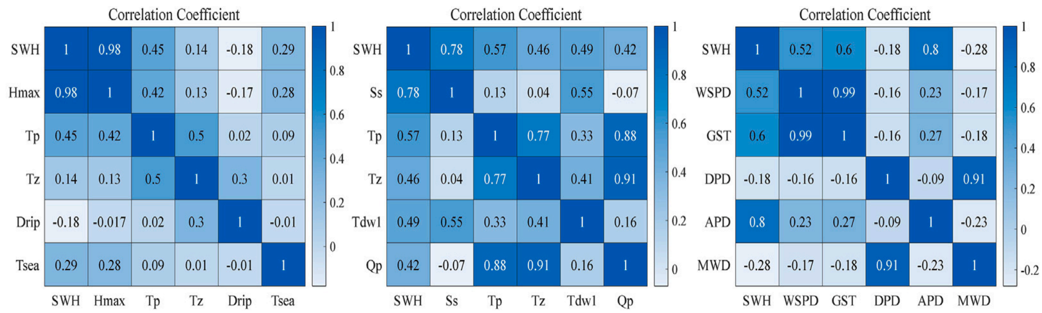


Fig. 3. Results of Pearson analyses for the three datasets.

time interval between zero-crossings of the wave. *Drip* is the direction of wave propagation, and *Tsea* is the average sea surface temperature.

(2) The features of our dataset: *Ss*, *Tp*, *Tz*, *Tdw1*, and *Qp*. *Ss* is the significant wave steepness, which is the ratio of SWH to wavelength. *Tdw1* is the dominant wave period, which is the period with the maximum energy concentration in the wave energy spectrum. *Qp* is the Gotoh wave height, which is a wave calculation method proposed by Goda (Wiebe et al., 2014).

(3) The features of NDBC dataset: *WSPD*, *GST*, *DPD*, *APD*, and *MWD*. *WSPD* is the wind speed. *GST*, gust wind speed, is the maximum wind speed within a short period of time. *DPD* consistent with *Tdw1*, which is the dominant wave period. *APD* is the average period of all waves, and *MWD* is the average direction of wave propagation.

The specific analysis results are shown in Fig. 3. Highly correlated features exhibit stronger intrinsic relationships, which suggests that these features contribute more effectively to prediction.

4. SWH prediction

In the task of SWH prediction, the generation and propagation of SWH signals are influenced by various meteorological and oceanic phenomena at different scales. This results in SWH signals containing multiple time-scale components, which makes them nonlinear and non-stationary. Additionally, the prolonged collection time results in a shift in the data distribution. These problems make SWH difficult to predict. Existing research often focuses on contextual information while neglecting the relationships between input features. These problems limit the accuracy of wave prediction. The MFET is designed to solve these problems effectively.

As shown in Fig. 4, the MFET primarily consists of a signal decomposition module, an encoder stack, and a decoder stack. The encoder stack and decoder stack each consist of six encoders and decoders, respectively. The original input is $X = [\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3, \boldsymbol{a}_4, \boldsymbol{a}_5, \boldsymbol{h}] \in \mathbb{R}^{l\times f}$, where l represents the total length of the wave time series and $f = 6$ is the number of features (five wave features and SWH). The wave features $[\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3, \boldsymbol{a}_4, \boldsymbol{a}_5]$, most correlated with SWH as mentioned in Table 1, vary across datasets. The SWH feature $\boldsymbol{h}$ is optimally decomposed into nine components $[\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_9]$ by SSA-VMD. To align the dimensions of the decomposed components with the original features, we replicate the five wave features $[\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3, \boldsymbol{a}_4, \boldsymbol{a}_5]$ for each of the nine SWH components. Each SWH component $\boldsymbol{h}_i$ is then concatenated with the replicated wave features, resulting in nine distinct channels. After features are concatenated, the encoder input $X_e = [\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3, \boldsymbol{a}_4, \boldsymbol{a}_5, \boldsymbol{h}_i]$, where $1 \le i \le 9$. Every channel contains the original five wave features and one SWH component, forming a 9-channel input structure. For the SWH prediction task, we use a sliding window of fixed length 48 to generate inputs. Specifically, at each current time t, the encoder input $X_e \in \mathbb{R}^{9\times 48\times 6}$ consists of continuous data from $t-47$ to t.

4.1. MFET framework

Each encoder include the SE attention module, MHCA module, and multi-layer perceptron (MLP) module. Following normalization by the RevIN layer, X_e is fed into the first encoder. The SE attention module determines the weights of each channel through squeeze and excitation operations, thereby scaling X_e to selectively enhance or suppress various channel information. The MHCA module mainly comprises parallel convolutional layers and a multi-head self-attention (MHSA) mechanism. This module extracts contextual information and local information between features. The MLP module produces output. As shown in Fig. 4, each subsequent encoder takes the output of the previous encoder as its input and iteratively executes this process. The final encoder output will then be used in the subsequent decoding process.

Each decoder include the MHSA module, cross-attention (CA) module, and MLP module. For the first decoder, it not only receives the encoder output as part of input but also requires an additional specific input to help for forecasting future values. This supplementary input, termed the prior target sequence, encompasses prior information about future time series. We use p to represent the future prediction horizon. This study is designed to execute predictions at multiple horizons: 1, 2, 6, 12, 24, and 48, with varying p values for each prediction task. The prior target sequence is set to match the horizon p, which means that when predicting the SWH for p horizon, the decoder will input prior information of the same length. For example, at the current time t, when $p = 12$, the resulting predictions cover the time from $t+1$ to $t+12$. And the decoder input consists of historical data from time $t-11$ to t, which represents the most relevant 12 time steps. This input is represented as $X_d \in \mathbb{R}^{9\times p\times 6}$. X_d is input into the MHSA module. The output of the MHSA module, along with the encoder output, is fed into the CA module. This design enables the decoder to focus more precisely on the input sequence most related to the currently generated output. The first decoder output is obtained after the MLP module. As shown in Fig. 5, the output from the previous decoder is input into the next decoder along with the encoder output. The final decoder output undergoes dimension mapping by a linear layer and a Softmax layer. At last, The predicted SWH $Y \in \mathbb{R}^{p\times l'}$ is obtained by denormalization (De-ReVIN) and reconstruction operations, where l' is the length of output time series.

4.2. SWH decomposition module

The SWH signal contains components of various scales, such as wind, waves, and tides. This diversity results in a high complexity in the signal. These different scale components and their interactions make the SWH signal temporally unstable, which pose significant prediction challenges. VMD (Dragomiretskiy & Zosso, 2014) effectively addresses these challenges by decomposing the signal into intrinsic mode functions (IMFs) and a residual signal (Res) based on the different time–frequency characteristics. Each decomposed IMF represents a specific scale or frequency component of the SWH signal. The Res is

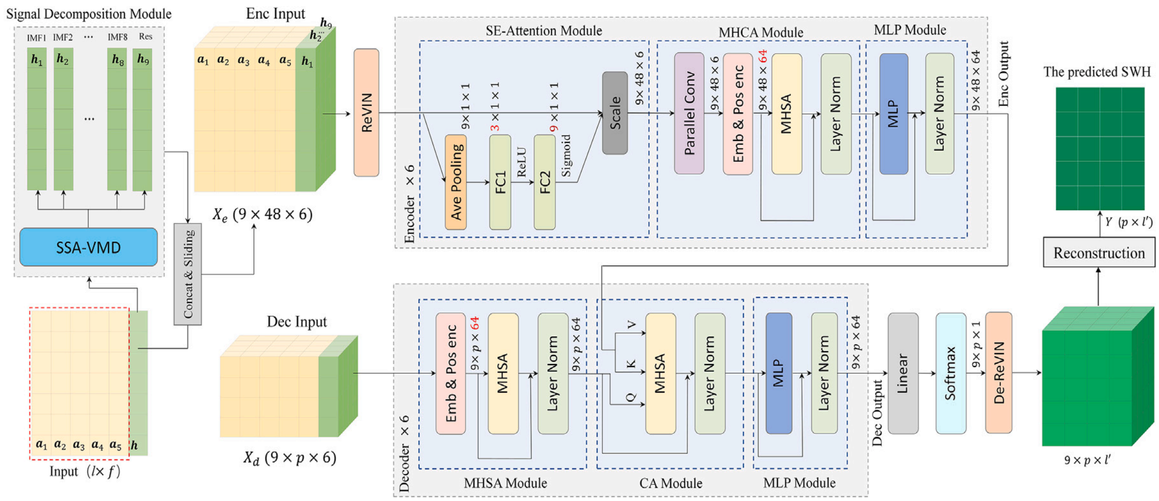


Fig. 4. The MFET framework.

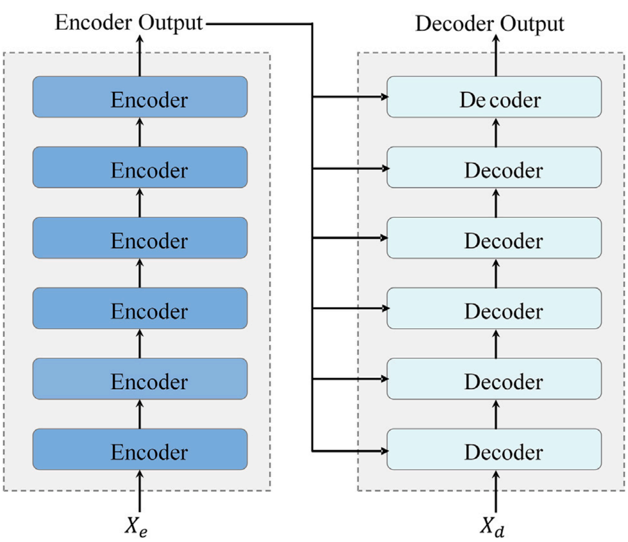


Fig. 5. Encoder and Decoder processing flowchart.

the difference between the sum of all IMFs and the original SWH signal, typically including noise, high-frequency oscillations, and possible outliers.

Through using VMD, the information of the SWH signal is effectively decomposed. The decomposed IMF exhibits greater regularity and stability, which makes them easier to understand and predict. The calculation process of VMD is expressed in Eq. (2):

$$x(t) = \sum_{k=1}^{K} u_k(t) + r(t) \tag{2}$$

where $x(t)$ is the original SWH signal, $u_k(t)$ is the kth modal function, K is the number of modes, and $r(t)$ is the residual term.

To best represent a specific frequency component of the SWH signal, the VMD algorithm finds the optimal modal function and center frequency by solving iteratively:

$$\min \|u_k(t) - c_k(t) \cdot u_k(t)\|_2^2 + \alpha \cdot \Phi(u_k(t)) \tag{3}$$

where $c_k(t)$ is the center frequency for the kth modal function, α is the regularization parameter controlling the smoothness and bandwidth of the modal functions, $\Phi(u_k(t))$ is the bandwidth of the modal function, and $\|u_k(t) - c_k(t) \cdot u_k(t)\|_2^2$ is the square L_2 norm of the reconstruction

Table 2
Details of optimization algorithm.

Dataset	K	α	MEE	MSE	R^2
Australia		4865	9.41	8.72×10^{-4}	99.79%
Ours	9	2353	9.72	3.26×10^{-5}	99.93%
NDBC		3029	8.23	3.15×10^{-3}	99.40%

error and aims to minimize the difference between the modal function and its corresponding center frequency. Among these, the most critical parameters are α and K.

SSA (Liu et al., 2023) is an optimization method inspired by the foraging and predator evasion behaviors of sparrows in their natural environment. It is applied to solve a variety of optimization problems. Given the differences between datasets and the challenges posed by large-scale data, this study adapts SSA for single-objective optimization. It specifically targets the optimization of the VMD problem. In this algorithm, we set the parameter $K = 9$, which aims to identify the optimal parameter α across various datasets to effectively decompose the best SWH signal. Table 2 presents a detailed overview of the optimization results for each dataset. In this study, the minimum envelope entropy (MEE) is employed as the fitness function to evaluate the uniformity of the solutions generated by the algorithm. Lower entropy values indicate a more concentrated set of solutions, while higher entropy values suggest greater dispersion. For the single-objective optimization problem addressed in this research, we aim to maintain lower MEE values to ensure the solutions for parameter α are both concentrated and precise. The MEE values we obtained from different datasets are 9.41, 9.72, and 8.23, respectively. These lower MEE values demonstrate a highly concentrated parameter optimization process, validating the decomposition efficacy of our proposed method. All the MSE and R^2 values clearly demonstrate that the reconstructed signals closely match the original signals. For example, the MSE of the reconstructed Australia dataset (with a scale of [0.277, 4.867] m) is 8.72×10^{-4}, and the R^2 value reaches 99.79% (see Table 2). These results robustly confirm the optimization algorithm's effectiveness in mitigating decomposition-induced prediction errors. The original and decomposed signals of the Australian dataset are presented in Fig. 6.

4.3. Encoder stack

The encoder stack consists of 6 encoders. Each encoder includes the SE attention module, MHCA module, and MLP module. The normalized X_e is input into the first encoder, and its output serves as the input for the next encoder. Each encoder performs iterative encoding tasks

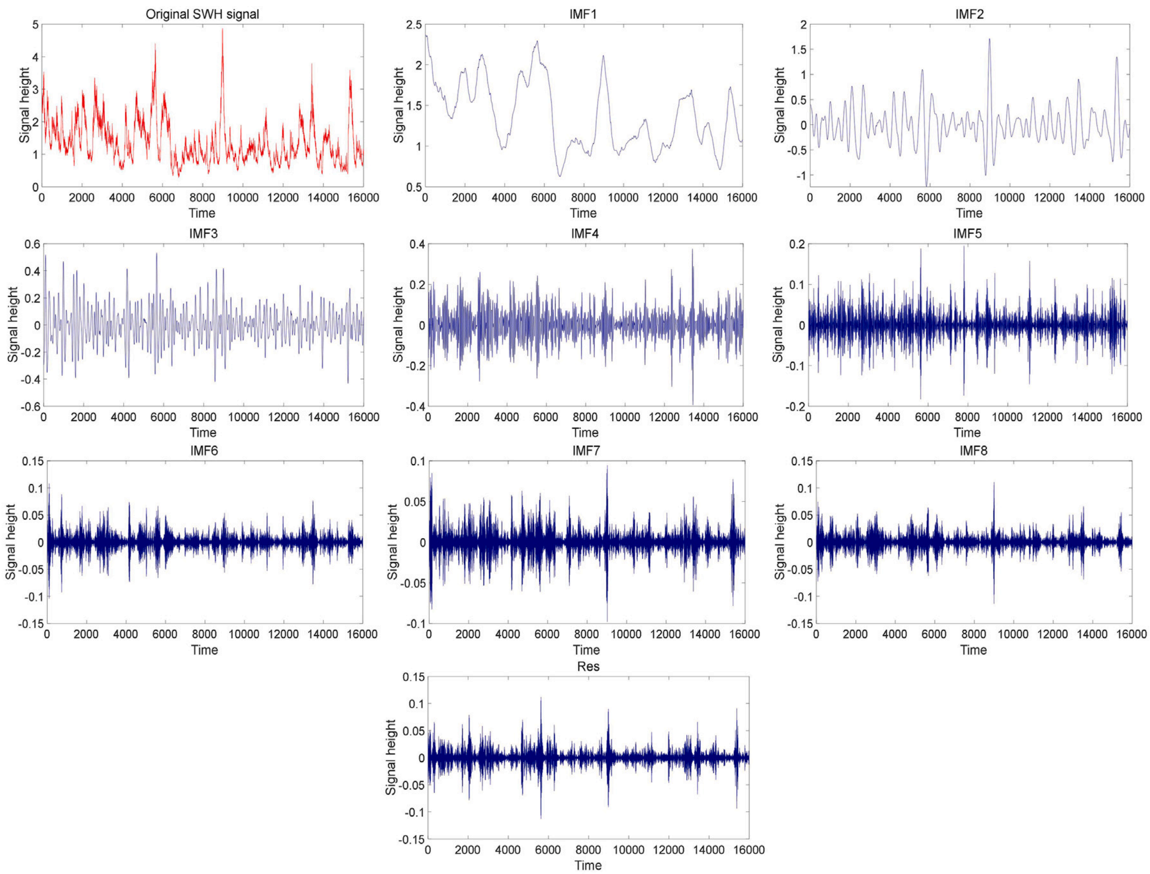


Fig. 6. The result of SSA-VMD in the Australia dataset.

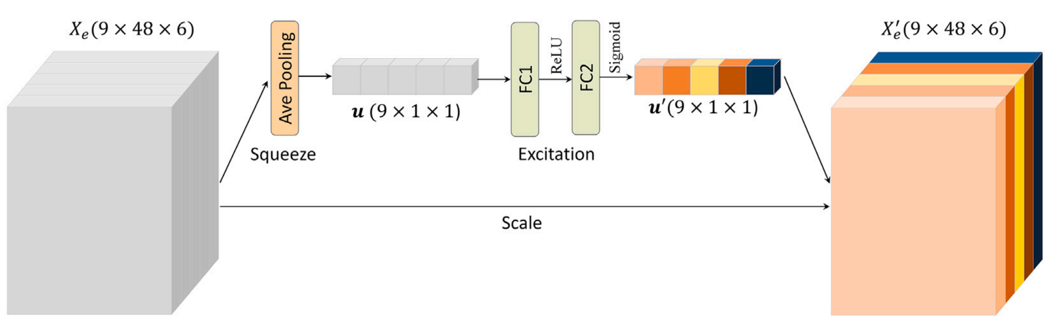


Fig. 7. SE attention module.

to produce the final encoder output, which is used for the subsequent decoding phase.

(1) SE attention module

Since the decomposed signals in different channels have varying importance, we use the SE attention module (Hu et al., 2018) to focus more on the information from important channels. As shown in Fig. 7, this module adopts the squeeze-and-excitation mechanism to enhance the ability of expressiveness by dynamically adjusting the relative importance of each feature channel. First, the input features $X_e \in \mathbb{R}^{9\times48\times6}$ undergo spatial compression by global average pooling and produce a channel feature map $\boldsymbol{u} \in \mathbb{R}^{9\times1\times1}$. This significantly reduces the spatial dimension and focuses on the channel information. Subsequently, two fully connected layers analyze and learn the inter-channel dependence from wave features. The first fully connected layer (FC1) uses a scaling factor $\boldsymbol{r} = 3$ to reshape $\boldsymbol{u}$, which reduces the computational burden. Then it undergoes a ReLU activation function to fit more complex mapping relationships. The second fully connected layer (FC2) then restores the dimensions back to their original form and uses a Sigmoid activation function to output weight factor $\boldsymbol{u}'$ for each channel. This weight factor is crucial for finely tuning the responses of the original features. At last, $\boldsymbol{u}'$ is multiplied with X_e for each channel to obtain the output $X_e' \in \mathbb{R}^{9\times48\times6}$ of this module, which recalibrates the response strength of each channel.

(2) MHCA module

The MHCA module is the core of the encoders. It can capture both global and local information and independently process information in different channels. This module includes the parallel convolution layer, position encoding and embedding layers, MHSA and layer normalization (LN). The parallel convolution layer uses nine 1×3 convolutional kernels to extract local information from different channels in parallel. Details of the convolution process are described in Fig. 8. The X_e' is input into the parallel convolution layer and outputs $E_i \in \mathbb{R}^{48\times6}$ for ith channel, where $i \in [1, 9]$. After convolution, the position encoding and embedding layers are used to further enrich the feature representation.

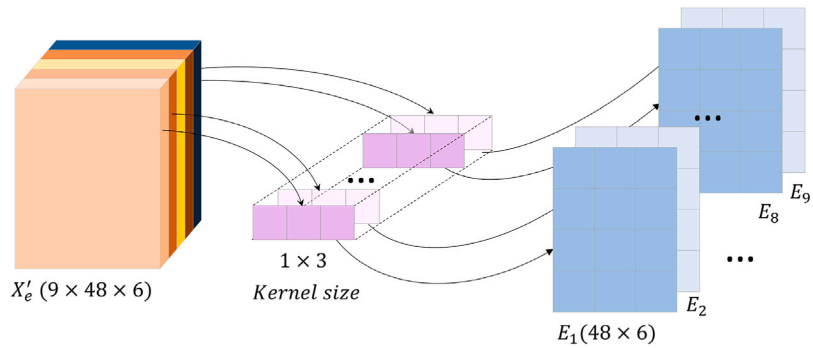


Fig. 8. The parallel convolution. $E_1, \ldots, E_9$ are the outputs of each channel.

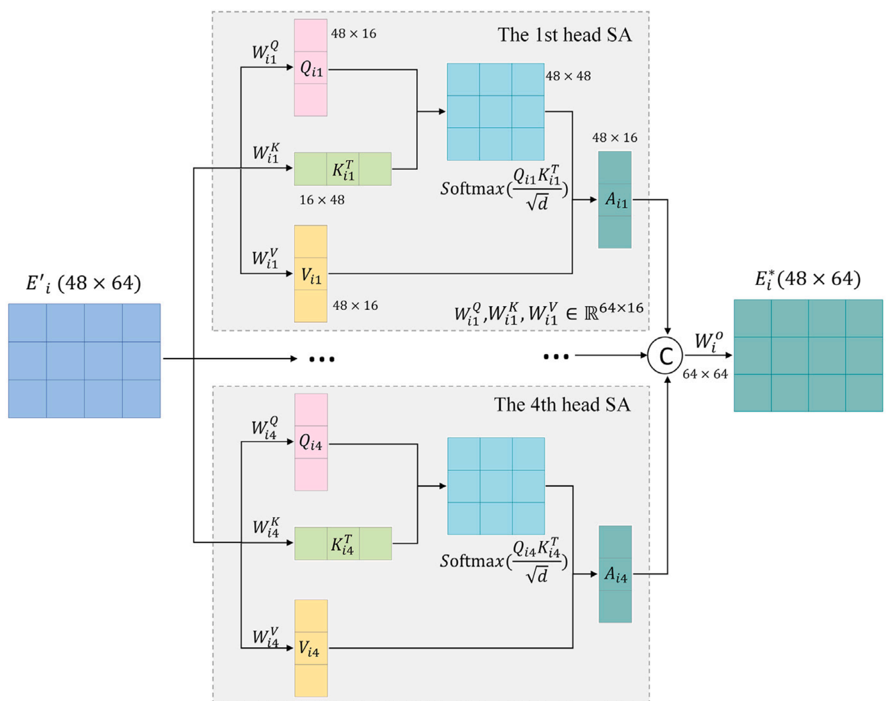


Fig. 9. The MHSA mechanism.

The position encoding ensures that the temporal information of wave features is accurately understood, and the embedding is used to expand the feature dimensions from 6 to 64. The output after embedding is $E_i' \in \mathbb{R}^{48\times64}$.

Following this, the MHSA mechanism processes the information of each channel in parallel and captures the features from different attention heads. The MHSA mechanism builds on the self-attention framework and divides the feature dimensions after embedding based on the number of attention heads. Then each head conducts self-attention mechanism independently before the results are merged. In this study, the number of attention heads is 4. After evenly dividing the 64 dimensions, each head processes a 16-dimensional wave feature. The specific changes in dimensions during this process are shown in Fig. 9, and the calculation process is given in Eqs. (4)–(6) (Vaswani et al., 2023). The input E_i' is multiplied by the weight matrices W_{ij}^Q, W_{ij}^K, and W_{ij}^V respectively, to obtain the queries Q_{ij}, keys K_{ij}, and values V_{ij}. After multiplying the queries by the transposed keys and scaling, they are passed through the Softmax activation function to obtain the attention score matrix. This matrix is then multiplied by the values to obtain the output A_{ij}. The outputs from all four heads are then merged and linearly re-mapped to the original dimensions, which results in the final output $E_i^* \in \mathbb{R}^{48\times64}$. Finally, E_i^* undergoes LN to maintain the stability of the training process and accelerate convergence.

$$(Q_{ij}, K_{ij}, V_{ij}) = (E_i' W_{ij}^Q, E_i' W_{ij}^K, E_i' W_{ij}^V) \tag{4}$$

$$A_{ij} = \mathrm{Softmax}\left(\frac{Q_{ij}K_{ij}^T}{\sqrt{d}}\right)V_{ij} \tag{5}$$

$$E_i^* = \mathrm{Concat}(A_{1j}, \ldots, A_{4j})W_i^o \tag{6}$$

where, i denotes the ith channel, while j denotes the jth head. Q_{ij}, K_{ij}, and V_{ij} represent the queries, keys, and values, respectively. W_{ij}^Q, W_{ij}^K, and W_{ij}^V are their respective weights. The feature dimension per head is $d = 16$. A_{ij} is the output of the jth head, and W_i^o is the weight for the combined total outputs.

(3) MLP module

The MLP module is necessary because it effectively enhances the ability to process complex features. After E_i^* is input into this module, it passes through two fully connected layers. The first, FC1, expands the dimensionality of the input features from 64 to 1024. The second, FC2, reduces these dimensions back to the original 64. A nonlinear activation function links these layers. Finally, it undergoes LN to obtain

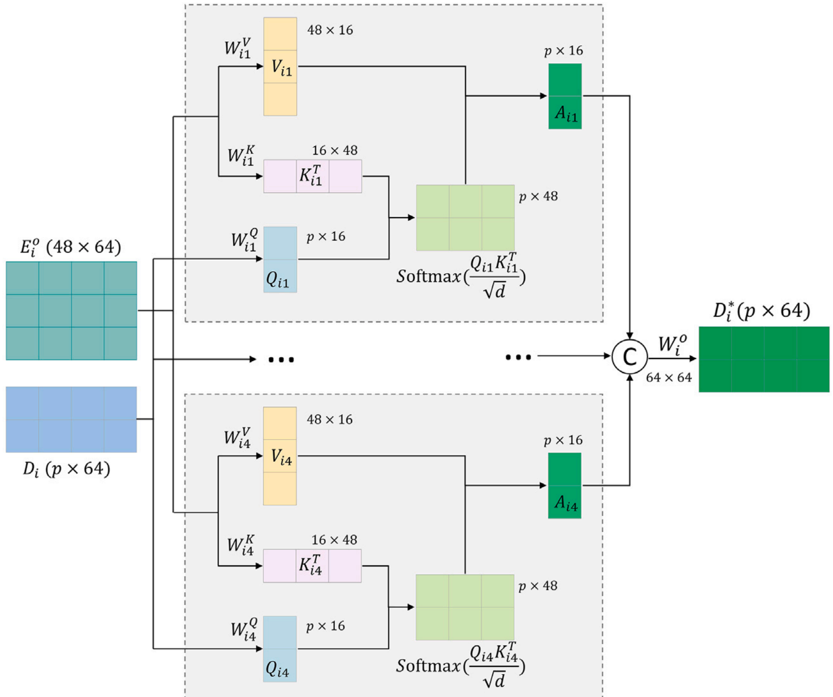


Fig. 10. The process of the CA module. E_i^o is the encoder output for the ith channel, and D_i is the ith channel of X'_d. A_{i1} is the 1th head output. D_i^* is the output of the CA module.

the encoder output. Using the MLP module can effectively boost the representation ability of model.

4.4. Decoder stack

Similar to the encoder stack, the decoder stack have 6 decoders. Each decoder contains the MHSA module, CA module and MLP module. Each decoder receives two parts as input: one part comes from the final encoder output; the other part varies—for the first decoder, it is the prior target sequence $X_d \in \mathbb{R}^{9 \times p \times 6}$, where p is the prediction horizon; for subsequent decoders, it is the output from the previous decoder.

When X_d is input into the MHSA module, it first undergoes dimensional expansion and temporal dependency capture by the embedding and positional encoding layers, and then $X'_d \in \mathbb{R}^{9 \times p \times 64}$ is generated by the MHSA mechanism. The MHSA module processes the information of each channel in parallel. The number of attention heads in MHSA is also set to 4. This module can capture the dependencies between the current time step and the previous time steps. Subsequently, X'_d along with the encoder output is input into CA module. Specifically, X'_d provides the queries, while the encoder output provides the keys and values. This module also uses the MHSA mechanism to process the information of different channels in parallel, and its output can effectively uses the contextual information from the encoder to help the decoding process. The MHSA in this module also have 4 attention heads. Fig. 10 illustrates the process of the CA module. E_i^o (the ith channel of the encoder output) and D_i (the ith channel of X'_d) are input into the CA module and outputs D_i^*. After being processed by this module, D_i^* is input into the MLP module and generates the decoder output. These modules allow the decoder to effectively assimilate and utilize the information provided by the encoder.

After the decoder stack, the linear layer and Softmax activation function are used to map the feature dimension from 64 to 1, which represents the SWH feature. Furthermore, the RevIN (Kim et al., 2022) is used to better solve the problem of temporal distribution shifts caused by the long span of data collection. The RevIN primarily includes normalization and denormalization processes. The normalization dynamically adjusts the mean and standard deviation to adapt to changes in data distribution, and this process is carried out before encoding. The denormalization is performed after the Softmax layer. It restores the data to its original scale and distribution through label denormalization. Following this, the data from each channel is reconstructed to obtain the final SWH prediction output Y.

5. Experiment and results

We conducted ablation and comparison experiments using three datasets (Australia, our dataset, and NDBC). Detailed information about the datasets is provided in Table 1. Each experiment was performed with predictions for horizons of 1, 2, 6, 12, 24, and 48. This means, for example, Australian dataset with a sampling frequency of 0.5 h can achieve predictions in {0.5 h, 1 h, 3 h, 6 h, 12 h, 24 h}. All datasets used in the experiments are divided into training, validation, and testing sets in a 7:1.5:1.5 ratio. In the ablation study, we utilized each transformer variant model to effectively evaluate the performance of each module. In the comparison study, we compare our model with four marine prediction models and three other advanced models to demonstrate the superiority of our approach.

5.1. Evaluation metrics

To assess the performance of the models, the following evaluation metrics are used: the mean squared error (MSE), mean absolute error (MAE), coefficient of determination R^2 and mean absolute relative error (MARE). The equations for these metrics are shown in Eqs. (7)–(10) (Chicco D, 2021; Hodson, 2022; Robeson, 2023). MARE is a metric that measures the degree of deviation between the predicted value and the actual value, with a value closer to 0 indicating a smaller deviation. The performance improvement in MSE ($\Delta\%$) has been calculated in Eq. (11), which means the innovative model achieves a $\Delta\%$ reduction in MSE compared to the baseline model. This better reflects the enhancement of the new model.

$$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n}(y_t - +\hat{y}_t)^2 \tag{7}$$

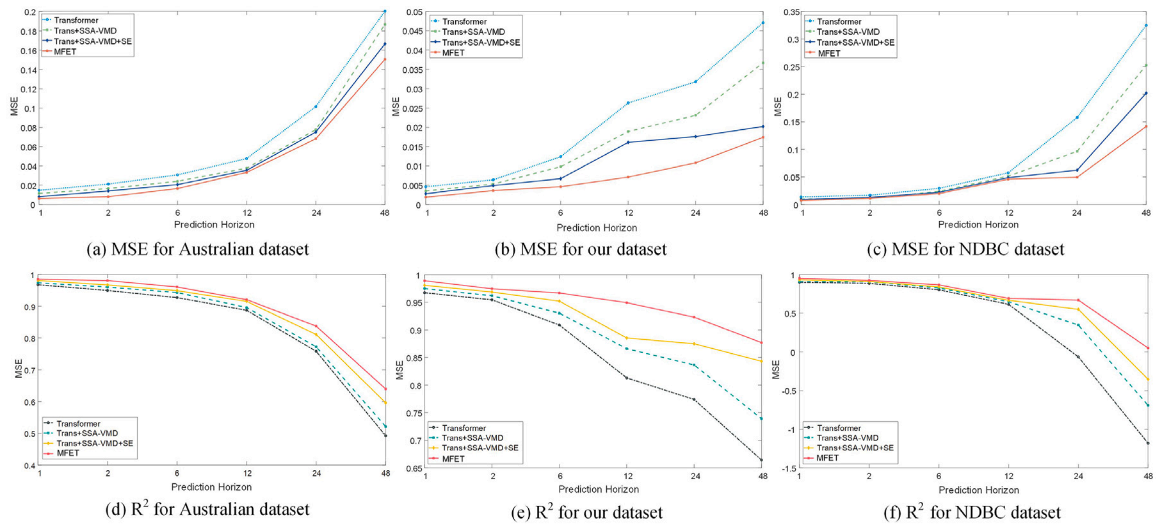


Fig. 11. The trends of MSE and R^2 for different prediction horizons across various datasets.

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} |y_t - \hat{y}_t| \tag{8}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y}_t)^2} \tag{9}$$

$$\mathrm{MARE} = \frac{1}{n}\sum_{t=1}^{n} \frac{|y_t - \hat{y}_t|}{y_t} \times 100\% \tag{10}$$

$$\Delta\% = \frac{\mathrm{MSE}_A - \mathrm{MSE}_B}{\mathrm{MSE}_A} \times 100\% \tag{11}$$

where $\hat{y}_t$ is the predicted SWH value at the time t, y_t is the actual SWH value at the time t, $\bar{y}_t$ is the mean of the actual values at the time t, and n is the length of the time series for the predicted SWH. A is the baseline model and B is the innovative model.

5.2. Ablation study

In this experiment, transformer is used as the backbone. To evaluating the performance of the SSA-VMD, SE attention, and parallel convolution modules in the MFET.

(1) Parameter Settings

The study uses PyTorch version 1.11.0 to build the MFET. The batch size is set to 64, and the loss function is specified as MSE. The learning rate is initialized at 0.01, and the weight decay parameter is set to 0.001. To achieve better experimental results, the stochastic gradient descent (SGD) optimizer is set to 0.7, and the forward propagation dimension is set to 1024.

(2) Results and Analysis

Table 3 details the performance of model across various datasets under different modules, and Fig. 11 illustrates trends in MSE and R^2 values. The results demonstrate that the MFET substantially outperformed the baseline transformer. For example, at the horizon of 12, the proposed MFET yields 30.2% (0.0477→0.0333), 73% (0.0263→0.0071), and 20.1% (0.0575→0.0459) MSE reduction on Australian dataset (with a scale of $[0.277, 4.867]$ m), our dataset ($[0.099, 2.912]$ m), and NDBC ($[0.25, 6.65]$ m) compared to transformer. It is particularly noteworthy that the MFET excels in the prediction horizon of 1, which shows an exceptionally high R^2 across three datasets, with values of 98.51%, 98.93%, and 94.59%. The prediction results at the 1-horizon across three datasets are shown in Fig. 12.

The addition of the SSA-VMD module significantly enhances the accuracy of the transformer. At the horizon of 6, this module achieves MSE reduction of 21.5% (0.0307→0.0241), 20.97% (0.0124→0.0098), and 18.8% (0.0293→0.0238) across three datasets. This enhancement stems from the module's ability to generate regular and stable subcomponents by reducing noise while preserving long-term trends, thereby improving the predictability of both IMFs and Res.

The SE attention module also contributes to performance enhancements across prediction horizons and datasets. At the horizon of 12, this module has MSE decrease of 5.8% (0.0377→0.0355), 14.8% (0.0189→0.0161), and 5.3% (0.0513→0.0486) across three datasets. These improvements confirm its effectiveness in automatically prioritizing critical features across channels, especially in identifying sporadic yet crucial wave patterns.

The parallel convolution layer addresses local feature extraction limitations through parallel filters processing. At the horizon of 48, this module achieves MSE reduction of 9.6% (0.1667→0.1507), 13.9% (0.0202→0.0174), and 30% (0.2020→0.1415) across three datasets. This design captures local features through parallel filters with different receptive fields, while simultaneously combining contextual information for global feature extraction. This capability is particularly important in long-term predictions.

In conclusion, each module's integration into backbone model has led to significant enhancements in performance, particularly demonstrated by substantial MSE reductions across varied datasets and prediction horizons. In addition, MFET shows strong robustness and generalization capabilities. It maintains high prediction accuracy on datasets with different sampling frequencies, sea state conditions, and related features.

5.3. Comparative experiments

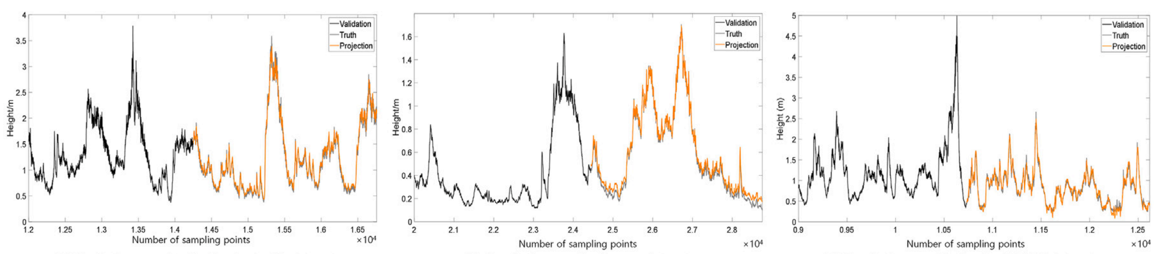
The comparative experiments are conducted with seven models: SWH-CLSTM (Guan, 2020), wave-S2S (Zeng et al., 2020), SWH-Trans (Wei et al., 2024), informer (Zhou et al., 2021), autoformer (Wu et al., 2022), reformer (Kitaev et al., 2020) and ATL-Net (Sun et al., 2024). The first three models and the last one are ocean wave prediction models, while the others are classical prediction models. The performance of these models is measured using the evaluation metrics MSE, MAE and MARE.

(1) Parameter Settings

To maintain consistency with the MFET, we keep certain basic parameters consistent among all baseline models, such as batch size, input length, and loss function. Taking into account the characteristics of the baseline models, the number of attention heads is set to 6 for the SWH-Trans, informer, autoformer, and reformer models. The encoder and decoder layers are kept at their default values, and the

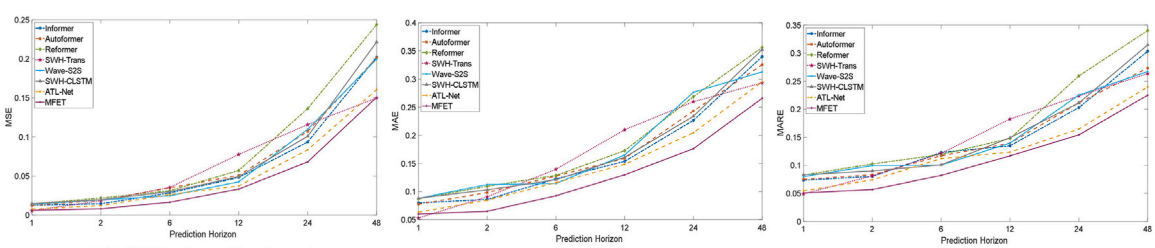
Table 3
Performance assessment of the prediction results of each model.

Horizon	Module				Australian datasets			Our datasets			NDBC datasets		
	Trans	SSA-VMD	SE-Att	Par-Conv	MSE	MAE	R^2	MSE	MAE	R^2	MSE	MAE	R^2
1	✓	×	×	×	0.0147	0.0865	0.9677	0.0046	0.0544	0.9677	0.0138	0.0968	0.9007
	✓	✓	×	×	0.0115	0.0816	0.9739	0.0035	0.0466	0.9751	0.0098	0.0910	0.9090
	✓	✓	✓	×	0.0081	0.0731	0.9811	0.0028	0.0405	0.9810	0.0090	0.0864	0.9321
	✓	✓	✓	✓	**0.0062**	**0.0601**	**0.9851**	**0.0019**	**0.0328**	**0.9893**	**0.0081**	**0.0644**	**0.9459**
2	✓	×	×	×	0.0212	0.1001	0.9497	0.0064	0.0653	0.9547	0.0169	0.0978	0.8867
	✓	✓	×	×	0.0165	0.0932	0.9602	0.0053	0.0579	0.9622	0.0113	0.0944	0.9080
	✓	✓	✓	×	0.0140	0.0871	0.9671	0.0049	0.0487	0.9689	0.0129	0.0899	0.9111
	✓	✓	✓	✓	**0.0081**	**0.0650**	**0.9809**	**0.0034**	**0.0386**	**0.9748**	**0.0100**	**0.0708**	**0.9328**
6	✓	×	×	×	0.0307	0.1435	0.9272	0.0124	0.1087	0.9090	0.0293	0.1237	0.8035
	✓	✓	×	×	0.0241	0.1019	0.9433	0.0098	0.0862	0.9305	0.0238	0.1095	0.8269
	✓	✓	✓	×	0.0206	0.1037	0.9491	0.0067	0.0605	0.9523	0.0221	0.1147	0.8404
	✓	✓	✓	✓	**0.0165**	**0.0926**	**0.9607**	**0.0046**	**0.0468**	**0.9672**	**0.0198**	**0.0994**	**0.8669**
12	✓	×	×	×	0.0477	0.1554	0.8872	0.0263	0.1308	0.8132	0.0575	0.1770	0.6148
	✓	✓	×	×	0.0377	0.1597	0.8959	0.0189	0.1111	0.8661	0.0513	0.1725	0.6490
	✓	✓	✓	×	0.0355	0.1388	0.9153	0.0161	0.1041	0.8855	0.0486	0.1661	0.6672
	✓	✓	✓	✓	**0.0333**	**0.1299**	**0.9208**	**0.0071**	**0.0602**	**0.9495**	**0.0459**	**0.1498**	**0.6923**
24	✓	×	×	×	0.1016	0.2304	0.7582	0.0318	0.1435	0.7741	0.1579	0.3595	−0.0644
	✓	✓	×	×	0.0777	0.2099	0.7723	0.0231	0.1202	0.8367	0.0965	0.2540	0.3465
	✓	✓	✓	×	0.0751	0.1972	0.8111	0.0176	0.1072	0.8751	0.0624	0.1887	0.5505
	✓	✓	✓	✓	**0.0682**	**0.1763**	**0.8375**	**0.0108**	**0.0766**	**0.9233**	**0.0494**	**0.1575**	**0.6696**
48	✓	×	×	×	0.2006	0.3637	0.4923	0.0471	0.174	0.6643	0.3248	0.5051	−1.1809
	✓	✓	×	×	0.1869	0.3472	0.5213	0.0367	0.1568	0.7393	0.2527	0.4186	−0.6912
	✓	✓	✓	×	0.1667	0.3063	0.5959	0.0202	0.1189	0.8434	0.2020	0.3715	−0.3537
	✓	✓	✓	✓	**0.1507**	**0.2657**	**0.6396**	**0.0174**	**0.0999**	**0.8771**	**0.1415**	**0.2807**	**0.0506**



(a) Prediction results for the Australia dataset (b) Prediction results for our dataset (c) Prediction results for the NDBC dataset

Fig. 12. The prediction results at the 1-horizon across three datasets.



(a) MSE for Australian dataset (b) MAE for Australian dataset (c) MARE for Australian dataset

Fig. 13. The trends of MSE, MAE and MARE for different prediction models on Australian dataset.

fully connected layer size was set to 2048. The time feature encoding is introduced by the "timeF" approach. For the wave-S2S model, the LSTM is chosen as the unit structure, with one stacking layer and a hidden layer dimension of 64. The SWH-CLSTM model had a hidden layer dimension of 64 and an output channel of 9. All models use the Adam optimizer.

(2) Results and Analysis

Table 4 presents the experimental results of our model compared to seven baseline models on the Australian dataset. Fig. 13 shows the trend variations for each metric. The results indicate that our model outperforms the comparison models across all metrics. Specifically, for MSE, our model consistently performs better than the other models at almost all prediction horizons, with significant improvements. For example, at the 2-horizon, MFET outperforms SWH-CLSTM, wave-S2S, SWH-Trans, and ATL-Net on MSE by reducing 59.1% (0.0198→0.0081), 58.5% (0.0195→0.0081), 55.7% (0.0183→0.0081), and 33.1% (0.0121→0.0081). MAE also shows similarly favorable results. For instance, at the 24-horizon, the MFET result is 0.1763, much lower than the second-best value of 0.2265 from informer, which demonstrates MFET's superiority in handling extreme values. Additionally, MFET performs well on the MARE metric, with a value of just 22.5% at the 48-horizon, indicating that MFET provides more stable prediction results.

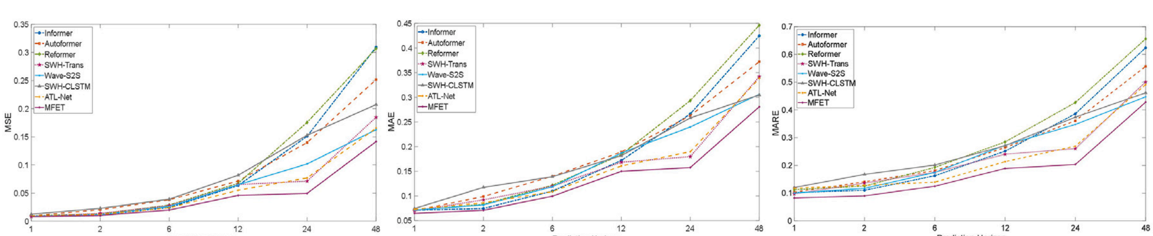
Table 5 presents the experimental results of all models and Fig. 14 is corresponding trend graphs on the NDBC dataset. Our model shows significant advantages on this dataset, outperforming other models across all metrics. The improvement in MSE performance is particularly

Table 4
Results of MSE, MAE, and MARE across various prediction horizons on Australian dataset.

Dataset	Horizon	Metric	MFET	SWH-CLSTM	Wave-S2S	SWH-Trans	Informer	Auto-former	Reformer	ATL-Net
Australian dataset	1	MSE	0.0062	0.0147	0.0137	**0.0057**	0.0125	0.0123	0.0145	0.0077
		MAE	0.0601	0.0882	0.0880	**0.0531**	0.0799	0.0775	0.0870	0.0634
		MARE	5.06%	8.32%	7.98%	**4.91%**	7.34%	7.50%	8.28%	5.44%
	2	MSE	**0.0081**	0.0198	0.0195	0.0183	0.0145	0.0190	0.0219	0.0121
		MAE	**0.0650**	0.1030	0.1125	0.0908	0.0860	0.0982	0.1094	0.0848
		MARE	**5.67%**	9.01%	9.96%	8.11%	8.03%	8.32%	10.25%	7.40%
	6	MSE	**0.0165**	0.0299	0.0248	0.0355	0.0284	0.0349	0.0312	0.0263
		MAE	**0.0926**	0.1211	0.1144	0.1397	0.1227	0.1277	0.1288	0.1160
		MARE	**8.21%**	10.09%	10.04%	12.14%	12.29%	11.71%	11.94%	11.21%
	12	MSE	**0.0333**	0.0493	0.0427	0.0777	0.0482	0.0510	0.0572	0.0374
		MAE	**0.1299**	0.1611	0.1651	0.2099	0.1538	0.1590	0.1727	0.1485
		MARE	**11.69%**	14.91%	13.96%	18.25%	13.48%	13.94%	14.78%	12.35%
	24	MSE	**0.0682**	0.1009	0.1098	0.1158	0.0936	0.1070	0.1363	0.0836
		MAE	**0.1763**	0.2338	0.2766	0.2596	0.2265	0.2432	0.2687	0.2046
		MARE	**15.42%**	21.13%	22.52%	22.37%	20.29%	21.14%	25.93%	16.46%
	48	MSE	0.1507	0.2215	0.1996	**0.1499**	0.2021	0.2013	0.2438	0.1605
		MAE	**0.2657**	0.3524	0.3128	0.2934	0.3399	0.3254	0.3562	0.2951
		MARE	**22.50%**	31.46%	26.73%	26.36%	30.34%	27.34%	34.03%	23.98%

Table 5
Results of MSE, MAE, and MARE across various prediction horizons on NDBC dataset.

Dataset	Horizon	Metric	MFET	SWH-CLSTM	Wave-S2S	SWH-Trans	Informer	Auto-former	Reformer	ATL-Net
NDBC dataset	1	MSE	**0.0081**	0.0123	0.0086	0.0083	0.0092	0.0093	0.0096	0.0090
		MAE	**0.0644**	0.0744	0.0711	0.0697	0.0708	0.0700	0.0710	0.0729
		MARE	**8.26%**	12.04%	10.02%	9.73%	10.19%	10.34%	11.23%	11.76%
	2	MSE	**0.0100**	0.0228	0.0111	0.0137	0.0110	0.0204	0.0127	0.0116
		MAE	**0.0708**	0.1173	0.0817	0.0919	0.0745	0.0988	0.0829	0.0859
		MARE	**9.02%**	16.76%	11.74%	13.64%	11.04%	14.12%	12.56%	12.85%
	6	MSE	**0.0198**	0.0391	0.0262	0.0277	0.0241	0.0379	0.0286	0.0238
		MAE	**0.0994**	0.1389	0.1186	0.1196	0.1098	0.1399	0.1216	0.1087
		MARE	**12.49%**	20.13%	17.45%	17.49%	16.27%	18.12%	19.28%	13.96%
	12	MSE	**0.0459**	0.0817	0.0647	0.0651	0.0634	0.0715	0.0674	0.0550
		MAE	**0.1498**	0.1816	0.1869	0.1682	0.1718	0.1890	0.1837	0.1608
		MARE	**18.90%**	27.10%	27.27%	24.02%	25.12%	26.44%	28.46%	21.35%
	24	MSE	**0.0494**	0.1528	0.1018	0.0712	0.1510	0.1398	0.1757	0.0769
		MAE	**0.1575**	0.2582	0.2398	0.1797	0.2663	0.2628	0.2933	0.1896
		MARE	**20.33%**	37.40%	34.76%	26.04%	38.62%	36.11%	42.67%	26.93%
	48	MSE	**0.1415**	0.2073	0.1627	0.1847	0.3094	0.2517	0.3069	0.1654
		MAE	**0.2807**	0.3047	0.3061	0.3418	0.4247	0.3722	0.4461	0.3382
		MARE	**42.79%**	46.16%	44.73%	50.11%	62.43%	55.71%	65.62%	49.14%



(a) MSE for NDBC dataset (b) MAE for NDBC dataset (c) MARE for NDBC dataset

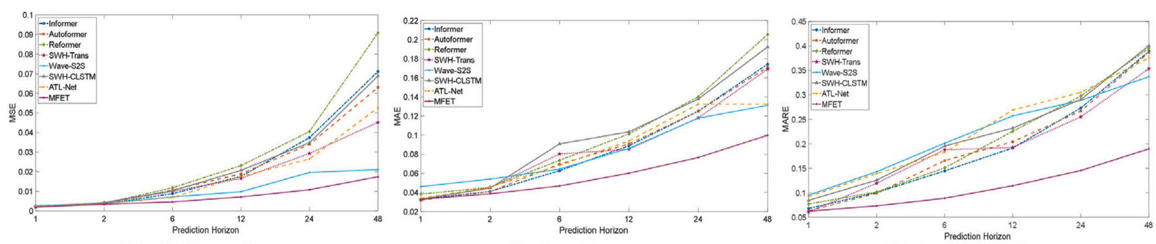
Fig. 14. The trends of MSE, MAE and MARE for different prediction models on NDBC dataset.

noticeable. For example, in the 6-horizon prediction, MFET reduces the MSE by 17.8% (0.0241→0.0198), 47.8% (0.0379→0.0198), 30.8% (0.0286→0.0198) and 16.8% (0.0238→0.0198) compared to informer, autoformer, reformer, and ATL-Net respectively. Additionally, MFET also performs well on the MARE metric, particularly in the first step prediction, where the MARE value is 8.26%, a reduction of 3.78% compared to SWH-CLSTM. These results also indicate MFET's excellent generalization ability, as it maintains stable performance across diverse datasets.

To further validate the generalization ability of the model, we conducted experiments on our dataset with significant sampling differences. Table 6 and Fig. 15 show the experimental results and trend variations for all models on this dataset. The results indicate that our model demonstrates more stable performance, with significant improvements compared to other models. For example, in comparison to ATL-Net, MFET's MSE values decreased by 13.6% (0.0022→0.0019), 8.11% (0.0037→0.0034), 33.33% (0.0069→0.0046), 59.66% (0.0176 →0.0071), 59.25% (0.00265→0.0108), and 66.92% (0.0526→0.0174) at different prediction horizons. Additionally, by observing the values

Table 6
Results of MSE, MAE, and MARE across various prediction horizons on our dataset.

Dataset	Horizon	Metric	MFET	SWH-CLSTM	Wave-S2S	SWH-Trans	Informer	Auto-former	Reformer	ATL-Net
Our dataset	1	MSE	**0.0019**	0.0020	0.0027	0.0021	0.0022	0.0020	0.0026	0.0022
		MAE	0.0328	0.0331	0.0461	0.0324	0.0330	**0.0319**	0.0383	0.0339
		MARE	6.23%	8.45%	9.53%	**6.09%**	6.79%	6.31%	7.72%	9.23%
	2	MSE	**0.0034**	0.0043	0.0038	0.0040	0.0035	0.0036	0.0039	0.0037
		MAE	**0.0386**	0.0446	0.0540	0.0452	0.0411	0.0409	0.0454	0.0456
		MARE	**7.33%**	12.58%	14.30%	11.96%	10.06%	9.90%	10.17%	13.91%
	6	MSE	**0.0046**	0.0103	0.0071	0.0099	0.0088	0.0107	0.0118	0.0069
		MAE	**0.0468**	0.0909	0.0645	0.0803	0.0628	0.0697	0.0737	0.0684
		MARE	**8.89%**	19.56%	20.14%	18.80%	14.47%	16.59%	15.09%	18.27%
	12	MSE	**0.0071**	0.0207	0.0098	0.0166	0.0176	0.0189	0.0231	0.0176
		MAE	**0.0602**	0.1035	0.0862	0.0853	0.0889	0.0909	0.1013	0.0939
		MARE	**11.44%**	23.22%	25.71%	19.27%	19.16%	20.42%	22.58%	26.92%
	24	MSE	**0.0108**	0.0348	0.0196	0.0293	0.0374	0.0340	0.0405	0.0265
		MAE	**0.0766**	0.1379	0.1178	0.1181	0.1252	0.1251	0.1402	0.1325
		MARE	**14.55%**	29.07%	28.89%	25.52%	27.33%	26.71%	29.76%	30.52%
	48	MSE	**0.0174**	0.0688	0.0211	0.0452	0.0711	0.0630	0.0908	0.0526
		MAE	**0.0999**	0.1924	0.1313	0.1692	0.1744	0.1712	0.2055	0.1883
		MARE	**18.98%**	40.04%	33.72%	35.38%	38.89%	38.66%	39.55%	37.56%



(a) MSE for our dataset (b) MAE for our dataset (c) MARE for our dataset

Fig. 15. The trends of MSE, MAE and MARE for different prediction models on our dataset.

Table 7
Performance improvement calculations on two datasets.

Dataset	Our dataset						NDBC					
Horizon	1	2	6	12	24	48	1	2	6	12	24	48
Reformer	0.0026	0.0039	0.0118	0.0231	0.0405	0.0908	0.0096	0.0127	0.0286	0.0674	0.1757	0.3069
MFET	0.0019	0.0034	0.0046	0.0071	0.0108	0.0174	0.0081	0.0100	0.0198	0.0459	0.0494	0.1415
Δ%	26.92%	12.82%	61.02%	69.26%	**73.33%**	**80.84%**	15.62%	21.26%	30.77%	31.90%	**71.88%**	**53.89%**
Informer	0.0022	0.0035	0.0088	0.0176	0.0374	0.0711	0.0092	0.0110	0.0241	0.0634	0.1510	0.3094
MFET	0.0019	0.0034	0.0046	0.0071	0.0108	0.0174	0.0081	0.0100	0.0198	0.0459	0.0494	0.1415
Δ%	13.64%	2.86%	47.73%	59.66%	**71.12%**	**75.53%**	11.96%	0.091%	17.84%	27.60%	**67.28%**	**54.27%**

of MARE and their trends, it can be seen that MFET significantly outperforms other models across all prediction horizons. Specifically, in the long-term predictions at 24 and 48 horizons, the MARE values are lower than the second-best model by 10.97% and 14.74%, respectively. These results demonstrate its strong robustness and generalization ability in long-term predictions.

To further illustrate the superior performance of our model in long-term predictions (at 24 and 48), we present the *Δ%* values under various conditions in Table 7. As shown in this table, on the NDBC dataset, our model outperforms the reformer model by 15.62%, 21.26%, 30.77%, 31.9%, 71.88%, and 53.89% at each prediction horizon. This demonstrates that the performance improvement in long-term forecasts is notably higher than in shorter prediction horizons. This trend is consistently observed across other datasets and comparison models.

6. Conclusions

In this study, we conduct experiments in the Yellow Sea using the MK III buoy to measure various wave features. We utilize the data collected along with two public datasets independently to enhance SWH feature prediction. To solve the problems of signal instability and inadequate local feature learning in SWH prediction, we propose the MFET model. This model introduces a signal decomposition module that employs SSA-VMD technology to break down the SWH signal into 8 IMFs and a Res signal, which represent more stable and regular component signals. These decomposed signals, along with other wave features, are reconstructed into a 3D dataset for subsequent prediction tasks. The prediction model primarily consists of an encoder stack and a decoder stack. In the encoder, we incorporate an SE attention module to dynamically adjust the influence of each channel, based on the varying impact of the different channels on the prediction. Additionally, to overcome the limitations of traditional transformer models in capturing local features, we design an MHCA module, which integrates parallel convolutional layers with the MHSA module to enhance the focus on both global information and local relationships. Furthermore, we use the RevIN layer to manage the distribution shift problem caused by long-term data collection.

We use multiple datasets to conduct extensive ablation experiments on the core modules of the MFET. The experimental design for each dataset includes multi-horizon predictions at 1, 2, 6, 12, 24, and 48, aimed at comprehensively assessing the effectiveness of each module. The results consistently demonstrate that these modules significantly enhance model performance in predictions. Additionally, successful tests on three distinctly different datasets illustrate our model's high generalizability and robustness. Further, our model is compared with six advanced prediction models. In these comparisons, our model shows superior prediction accuracy and stability, especially in long-term forecasts, where its performance is particularly notable. These results strongly affirm the advanced nature and potential applications of our model.

While the proposed method demonstrates superior forecasting accuracy across three wave datasets, two key limitations should be noted:

(1) The time-consuming data preprocessing requirements for optimized 3D dataset construction currently increase computational overhead. (2) The model architecture generates a large number of parameters, which poses a challenge for implementation in edge computing. Future efforts will prioritize computational optimization through lightweight decomposition algorithms and attention pruning techniques.

CRediT authorship contribution statement

Lin Wu: Conceptualization, Methodology, Program, Visualization, Writing – original draft. **Yi An:** Resources, Writing – review & editing. **Pan Qin:** Supervision, Resources, Formal analysis. **Huo-Sheng Hu:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62173055, in part by the Natural Science Foundation Program of Liaoning Province, China under Grant 2023-MS-093, in part by the Major Science and Technology Special Project of Xinjiang Uygur Autonomous Region, China under Grant 2023A01005-1, and in part by the Science and Technology Major Project of Shanxi Province, China under Grant 20191101014.

Data availability

I have included the link to the Git repository for the code and data in the manuscript.

References

Agrawal, J., & Deo, M. (2002). On-line wave prediction. *Marine Structures*, *15*(1), 57–74.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. (pp. 1–4).

Berbić, J., Ocvirk, E., Carević, D., & Lončar, G. (2017). Application of neural networks and support vector machine for significant wave height prediction. *Oceanologia*, *59*(3), 331–349.

Chicco D, J. G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*.

Chowdary, M., Premsai, P., Dasharna, M., Kavya, I., & T R, S. (2023). Ocean wave height forecasting using deep learning neural networks and optimization techniques. In *2023 innovations in power and advanced computing technologies (i-PACT)* (pp. 1–8).

Clauss, G. F. (2002). Dramas of the sea: episodic waves and their impact on offshore structures. *Applied Ocean Research*, *24*(3), 147–161.

Daniel, K., & Adytia, D. (2023). A significant wave height and peak wave period prediction with transformer and LSTM approach in Cilacap, Indonesia. In *2023 international conference on data science and its applications (iCoDSA)* (pp. 344–349).

Demetriou, D., Michailides, C., Papanastasiou, G., & Onoufriou, T. (2021). Coastal zone significant wave height prediction by supervised machine learning classification algorithms. *Ocean Engineering*, *221*, Article 108592.

Domala, V., & Kim, T.-W. (2022). A univariate and multivariate machine learning approach for prediction of significant wave height. In *OCEANS 2022, hampton roads* (pp. 1–7).

Dragomiretskiy, K., & Zosso, D. (2014). Variational mode decomposition. *IEEE Transactions on Signal Processing*, *62*(3), 531–544.

Duan, W., Han, Y., Huang, L., Zhao, B., & Wang, M. (2016). A hybrid EMD-SVR model for the short-term prediction of significant wave height. *Ocean Engineering*, *124*, 54–73.

Etemad-Shahidi, A., & Mahjoobi, J. (2009). Comparison between M5 model tree and neural networks for prediction of significant wave height in Lake Superior. *Ocean Engineering*, *36*(15), 1175–1181.

Foster, J., Li, N., & Cheung, K. F. (2014). Sea state determination from ship-based geodetic GPS. *Journal of Atmospheric and Oceanic Technology*, *31*, Article 140815105258005.

Guan, X. (2020). Wave height prediction based on CNN-LSTM. In *2020 2nd international conference on machine learning, big data and business intelligence* MLBDBI, (pp. 10–17).

Hashim, R., Roy, C., Motamedi, S., Shamshirband, S., & Petković, D. (2016). Selection of climatic parameters affecting wave height prediction using an enhanced Takagi-Sugeno-based fuzzy methodology. *Renewable and Sustainable Energy Reviews*, *60*, 246–257.

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, *15*(14), 5481–5487.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 7132–7141). http://dx.doi.org/10.1109/CVPR.2018.00745.

Huang, Q., & Cui, Z. (2023). Study on prediction of ocean effective wave height based on hybrid artificial intelligence model. *Ocean Engineering*, *289*, Article 116137.

Katalinić, M., & Parunov, J. (2020). Uncertainties of estimating extreme significant wave height for engineering applications depending on the approach and fitting technique—Adriatic Sea case study. *Journal of Marine Science and Engineering*, *8*(4).

Kim, T., Kim, J., Tae, Y., Park, C., Choi, J.-H., & Choo, J. (2022). Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International conference on learning representations* (pp. 1–25).

Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer: The efficient transformer. arXiv:2001.04451.

Kumar, N., Feddersen, F., Uchiyama, Y., McWilliams, J., & O'Reilly, W. (2015). Midshelf to surfzone coupled ROMS–SWAN model data comparison of waves, currents, and temperature: Diagnosis of subtidal forcings and response. *Journal of Physical Oceanography*, *45*(6), 1464–1490.

Li, H., Claremar, B., Wu, L., Hallgren, C., Körnich, H., Ivanell, S., & Sahlée, E. (2021). A sensitivity study of the WRF model in offshore wind modeling over the Baltic Sea. *Geoscience Frontiers*, *12*(6), Article 101229.

Liang, C., Sun, Z., Shu, G., Li, W., Liu, A.-A., Wei, Z., & Yin, B. (2024). Adaptive graph spatial-temporal attention networks for long lead ENSO prediction. *Expert Systems with Applications*, *255*, Article 124492.

Liu, Y., Sun, J., Shang, Y., Zhang, X., Ren, S., & Wang, D. (2023). A novel remaining useful life prediction method for lithium-ion battery based on long short-term memory network optimized by improved sparrow search algorithm. *Journal of Energy Storage*, *61*, Article 106645.

Minuzzi, F. C., & Farina, L. (2023). A deep learning approach to predict significant wave height using long short-term memory. *Ocean Modelling*, *181*, Article 102151.

Obakrim, S., Monbet, V., Raillard, N., & Ailliot, P. (2023). Learning the spatiotemporal relationship between wind and significant wave height using deep learning. *Environmental Data Science*, *2*, Article e5.

Orimolade, A. P., Haver, S., & Gudmestad, O. T. (2016). Estimation of extreme significant wave heights and the associated uncertainties: A case study using NORA10 hindcast data for the Barents Sea. *Marine Structures*, *49*, 1–17.

Pan, X., Jiang, T., Sun, W., Xie, J., Wu, P., Zhang, Z., & Cui, T. (2024). Effective attention model for global sea surface temperature prediction. *Expert Systems with Applications*, *254*, Article 124411.

Raj, N., & Prakash, R. (2024). Assessment and prediction of significant wave height using hybrid CNN-BiLSTM deep learning model for sustainable wave energy in Australia. *Sustainable Horizons*, *11*, Article 100098.

Rekha Sankar, S., & Panchapakesan, M. (2024). Hybrid feature selection model for accurate wind speed forecasting from numerical weather prediction dataset. *Expert Systems with Applications*, *248*, Article 123054.

Robeson, C. J. (2023). Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. *PLoS One*, *18*, 1–8.

Sun, Q., An, Y., Yu, X., Wu, L., & Wei, S. (2024). ATL-Net: Ocean current prediction based on self-attention mechanism and feature enhancement. In *2024 39th youth academic annual conference of Chinese association of automation* YAC, (pp. 1678–1683).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need. ArXiv Preprint arXiv:1706.03762.

Vries, J., Waldron, J., & Cunningham, V. (2003). Field tests of the new datawell DWR-G GPS wave buoy. *Sea Technology*, *44*, 50–55.

Wang, K., Liu, Y., Xing, Q., Qian, Y., Wang, J., & Lv, M. (2024). An integrated system to significant wave height prediction: Combining feature engineering, multi-criteria decision making, and hybrid kernel density estimation. *Expert Systems with Applications*, *241*, Article 122351.

Wei, S., An, Y., Yu, X., Wu, L., & Sun, Q. (2024). Effective wave height prediction based on non-stationary-CNN-transformer for Ocean Waves. *Acta Energiae Solaris Sinica*, *45*(10), 673–682.

Wiebe, D., Park, H., & Cox, D. (2014). Application of the goda pressure formulae for horizontal wave loads on elevated structures. *KSCE Journal of Civil Engineering*, *18*, http://dx.doi.org/10.1007/s12205-014-0175-1.

Wu, H., Liang, Y., Gao, X.-Z., Du, P., & Li, S.-P. (2023). Human-cognition-inspired deep model with its application to ocean wave height forecasting. *Expert Systems with Applications*, *230*, Article 120606.

Wu, Y., Wang, J., Zhang, R., Wang, X., Yang, Y., & Zhang, T. (2024). RIME-CNN-BiLSTM: A novel optimized hybrid enhanced model for significant wave height prediction in the Gulf of Mexico. *Ocean Engineering*, *312*, Article 119224.

Wu, H., Xu, J., Wang, J., & Long, M. (2022). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. arXiv:2106.13008.

Zeng, X., Qi, L., Yi, T., & Liu, T. (2020). A Sequence-to-Sequence model based on attention mechanism for wave spectrum prediction. In *2020 11th international conference on awareness science and technology* ICAST, (pp. 1–5).

Zhang, M., Yuan, Z.-M., Dai, S.-S., Chen, M.-L., & Incecik, A. (2024). LSTM RNN-based excitation force prediction for the real-time control of wave energy converters. *Ocean Engineering*, *306*, Article 118023.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. arXiv:2012.07436.