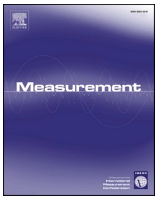
Measurement 242 (2025) 115947

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement



A novel Kalman smoothing (Ks) – Long Short-Term Memory (LSTM) hybrid model for filling in short- and long-term missing values in significant wave height

Yulian Wang [a,b], Taili Du [a,b,c,*], Yuanye Guo [a,b], Fangyang Dong [a,b], Jicang Si [a,b,*], Minyi Xu [a,b,*]

[a] Dalian Key Lab of Marine Micro/Nano Energy and Self-Powered Systems, Marine Engineering College, Dalian Maritime University, Dalian 116026, China
[b] State Key Laboratory of Maritime Technology and Safety, Dalian 116026, China
[c] Collaborative Innovation Research Institute of Autonomous Ship, Dalian Maritime University, Dalian 116026, China

ARTICLE INFO

Keywords:
Significant wave height
Kalman smoothing
LSTM
Missing values
Time series

ABSTRACT

Continuous recordings of significant wave height (SWH) are of significant importance to facilitate the development and application of wave energy assessments. However, SWH recording is often interrupted by a variety of factors, including severe weather, apparatus failure, etc. The presence of short-term intermittent and long-term continuous missing values is recognized as a significant challenge for the accurate analysis and energy assessment of wave data. To effectively fill in the short- and long-term missing values simultaneously, a hybrid Kalman smoothing (KS) – Long Short-Term Memory (LSTM) model, which is applied to fill in SWH for the first time, is proposed. Initially, the short-term intermittent missing values, with missing ratios of 10%-50%, are filled by using the KS method, achieving a root mean square error (RMSE) as low as 0.082. Based on short-term missing values addressed by KS, the LSTM model is introduced to effectively predict the long-term continuous missing values. The maximum RMSE reduction rate of KS-LSTM is reduced by 49.6%, 59.4%, and 57.8% compared to KS-ARIMA, LSTM, and ARIMA, respectively, within the short-term intermittent missing ratios of 10%-50%. The maximum reduction in mean square error (MSE) is observed to be 71.4%, 83.6%, and 82.1%. Similarly, the maximum reduction in mean absolute error (MAE) is 74.6%, 61.1%, and 55.5%. The lowest prediction RMSE of long-term missing values by KS-LSTM is only 0.091, which demonstrates that the effectiveness of filling in both short-term and long-term missing values simultaneously by the KS-LSTM.

1. Introduction

Ocean is recognized to possess substantial potential for generating renewable energy sources, such as wave energy, tidal energy, and thermal energy of the ocean [1,2]. With the increasing demand for reliable energy production in human societies and the expansion of renewable energy, infrastructure, researchers have conducted several useful explorations of ocean energy applications [3]. Among these, wave energy, as a form of renewable energy technology (RET) [4], has been particularly noted for its high energy density [5], better predictability [6], and notable environmental friendliness [7]. The research and development of wave energy are closely related to marine meteorological parameters such as Significant Wave Height (SWH), Mean Wave Direction (MWD), Dominant Wave Period (DPD), etc. It is recognized that SWH can intuitively reflect the characteristics and kinematic properties of waves and plays an important role in the comprehensive assessment of wave energy resources and power generation [8,9]. However, SWH time series observation records are often interrupted or missing due to unfavorable climatic conditions and equipment failures [10,11], posing a serious challenge to the reliability of wave energy studies. Therefore, the effective recovery of missing values in SWH time series records is regarded as important for advancing wave energy research and resource assessment.

According to related studies, missing values of SWH are normally identified as having two distinct forms: Short-term missing values are generally within 10, such as 4 to 5 small items of magnitude [12]. Long-

* Corresponding authors at: Dalian Key Lab of Marine Micro/Nano Energy and Self-Powered Systems, Marine Engineering College, Dalian Maritime University, Dalian 116026, China.

E-mail addresses: dutaili@dlmu.edu.cn (T. Du), sjc@dlmu.edu.cn (J. Si), xuminyi@dlmu.edu.cn (M. Xu).

https://doi.org/10.1016/j.measurement.2024.115947
Received 14 June 2024; Received in revised form 4 October 2024; Accepted 9 October 2024
Available online 11 October 2024

term missing values usually represent a continuous period of 1 month (about 30 h) or more [13]. The prevailing challenge lies in addressing the coexistence of short- and long-term missing values within time series data. Short-term intermittent missing values are often omitted directly, which may lead to the loss of critical information [14,15]. To address this issue, researchers have utilized the interpolation method, which captures temporal information in the data, to fill in the missing values. Linear, quadratic, and cubic interpolation techniques have been employed [16] to fill the SWH intermittent missing values. By comparison, it was found that the linear interpolation methods were able to provide a better fit to the data. Quinteros et al [17] attempted to reconstruct air quality datasets with data loss greater than 20 % using multiple interpolation techniques. Cubic spline interpolation was used by Abdullah et al [18] to fill approximately 20 % of the missing values in 3,000 wave height data points at the Karawang station. Furthermore, to compensate for the shortcomings of the interpolation method, which struggles to capture the dynamic change pattern of intermittent missing values, effective dynamic system state estimation has been carried out by researchers utilizing Kalman smoothing (KS) [19]. By reducing system noise and integrating past and current observations, KS can dynamically fill varying missing values in a time series [20]. For example, KS was utilized by Umar and Gray [21] to fill the water level data with a missing ratio of 5 % at the Kainji water station on the Niger River, achieving an average root mean square error (RMSE) of 13.61. Terms of up to more than 40 h in snow depth sensor data were filled using KS by Avanzi et al [22]. The research results demonstrate that KS has been widely used to evaluate missing data in snow depth sensors, water level data, and other fields with favorable outcomes. However, the effectiveness of KS in filling SWH missing values still needs further verification.

In the research of long-term continuous missing values filling, the essence of this filling process is identified as the forecasting of time series data. Long-term missing values are typically predicted using physical and data-driven models by researchers [23]. In terms of physical models, the SWAN-SWASH model was utilized by Umesh and Behera [24] to predict SWH off the east coast of India using a 1 m grid resolution. The wave generation in the Black Sea was simulated by Soran et al [25] using the third-generation wave model WAVEWATCH III. Although physical models that rely on energy balance equations provide accurate results [26], they typically require substantial computational resources and reliable wind field data as inputs, which limits their applicability for rapid forecasting. Therefore, fast and reliable methods for predicting long-term continuous missing values are provided by machine learning-based data-driven models [27]. Among these, machine learning methods (ML), such as Artificial Neural Networks (ANN) [28], Support Vector Machines (SVM) [29] and Random Forests (RF) [30] have been widely employed in the field of coastal and marine engineering. ML models, including the Bidirectional Gated Recurrent Unit (BiGRU) and Cressman analysis, were used by Wang et al [31] to recover 450 missing wave height data points from a buoy station. A Genetic Programming (GP)-based model [32] was employed to predict a missing consecutive month of data in the Gulf of Mexico. The ML methods described above are capable of filling in both short-term and long-term missing data based on temporal and spatial correlations. However, when the amount of available data in the time series is insufficient, the accuracy of methods based on temporal correlation in recovering missing values could be drastically reduced.

To address above issues, studies on SWH missing values recovery using classical time series methods, such as the autoregressive (AR) model [33] and moving average (MA) model [34], have been conducted by researchers as another data-driven domain approach. Long-term missing values in time series were predicted using an autoregressive moving average (ARMA) model by Ferreiro [35]. However, due to the complex nonlinear characteristics of SWH time series data, the predictive performance of models may be degraded by relying on linear expressions and data smoothing assumptions [36]. The effectiveness of the Long Short-Term Memory (LSTM) model in capturing nonlinear features in time-series data and its excellent time-series autoregressive ability [37] have been demonstrated across various fields, such as speech recognition, natural language processing, and image recognition [38]. In research on SWH, SWH data for 12 consecutive hours were predicted by Yao and Wu [39] using an extended LSTM with a multistep training set, resulting in a reduction in prediction error by 52.83 %. The wave height of the southwest Atlantic Ocean was predicted with an accuracy of nearly 87 % using LSTM [40]. Although some progress has been made in SWH prediction based on LSTM, the model's effectiveness in filling simultaneous short-term and long-term missing values of SWH still necessitates further in-depth study.

In summary, considerable explorations have been accomplished by researchers in recovering short-term intermittent and long-term continuous missing values in time series recordings. However, the research on simultaneously filling both types of missing values mentioned above in SWH is still in its early stages. To address this, a novel hybrid KS-LSTM model is proposed in this work, which firstly employs KS to efficiently fill intermittent short-term missing values. Subsequently, on the basis of the time series after the filling of short-term intermittent missing values, LSTM is utilized to realize the efficient filling of long-term continuous missing values. To evaluate the generalizability of the proposed hybrid model, datasets with five different missing ratios, including 10 %-50 %, are randomly constructed from five different SWH datasets from the actual public buoy stations. Higher applicability and prediction accuracy in filling short- and long-term concurrent missing values in variable environments and different data characteristics are demonstrated by the KS-LSTM model, as compared and analyzed with other models. The overall structure of the paper is organized as follows: Section 2 outlines the research methodology and the evaluation indicators to fill in missing values. In Section 3, the study area, materials, application and the creation of datasets with different ratios of missing SWH are described for the five public stations. Section 4 is dedicated to the discussion and analysis of the prediction results. Finally, the content and contributions of this paper are summarized in Section 5.

2. Research methods and evaluation indicators for filling in missing values

2.1. Kalman smoothing

The theoretical framework of Kalman smoothing (KS) is founded on the recursive estimation of the system state and applies to both linear and nonlinear systems [41,42]. In the process of filling in missing data values in the SWH time series, the filling state is estimated by KS using current and past observations. Through successive applications of KS prediction, updating, and smoothing, a coherent and smooth time series is generated, effectively filling the short terms of missing data points in SWH. This method is realized by performing a series of steps in the Kalman filtering (KF) and KS process [43–45], which are specified below.

Step 1 Forecasting

State prediction:

$$\widehat{x_{k|k-1}} = F_k \widehat{x_{k-1|k-1}} + B_k u_k \tag{1}$$

Covariance prediction:

$$P_{k|k-1} = F_k P_{k-1|k-1} F_k^T + Q_k \tag{2}$$

Where, $\widehat{x_{k|k-1}}$ represents the state at the time k, predicted using the information available up to time $k-1$. The transfer matrix is denoted by F_x, and the control input matrix is denoted by B_k. The control vector is represented by u_k, and the prediction covariance matrix is denoted by $P_{k|k-1}$. Finally, the noise covariance matrix of the process is denoted by

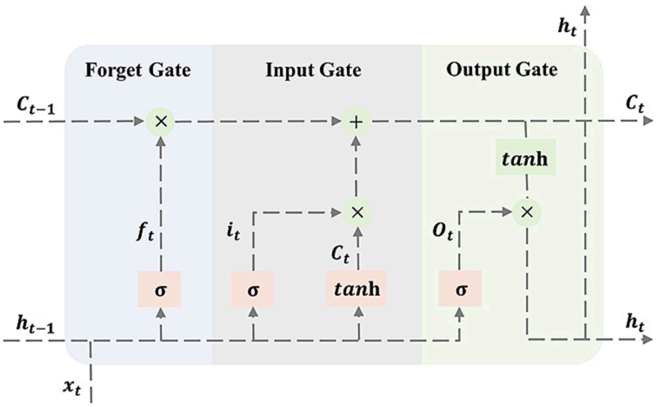


Fig. 1. Cell structure of LSTM.

Q_k.

Step 2 Update

Kalman gain:

$$K_k = P_{k|k-1}H_k^T\left(H_kP_{k|k-1}H_k^T + R_k\right)^{-1} \tag{3}$$

State update:

$$\widehat{x_{k|k}} = \widehat{x_{k|k-1}} + K_k\left(z_k - H_k\widehat{x_{k|k-1}}\right) \tag{4}$$

Covariance update:

$$P_{k|k} = (I - K_kH_k)P_{k|k-1} \tag{5}$$

Where, K_k represents the Kalman gain; H_k denotes the observation model matrix; z_k is the actual measurement at the time k; R_k is the observation noise covariance matrix; and I stands for the identity matrix.

Step 3 Backward Correction

Smoothing state estimation:

$$\widehat{x_{k|N}} = \widehat{x_{k|k}} + J_k\left(\widehat{x_{k+1|N}} - F_k\widehat{x_{k|k}}\right) \tag{6}$$

Smoothing covariance estimation:

$$P_{k|N} = P_{k|k} + J_k\left(P_{k+1|N} - P_{k|k-1}\right)J_k^T \tag{7}$$

Where, N represents the final time step; $\widehat{x_{k|N}}$ and $P_{k|N}$ are the smoothed state estimate and covariance estimate, respectively. $J_k = P_{k|k}F_k^TP_{k|k-1}^{-1}$ denotes the smoothing gain.

2.2. ARIMA

ARIMA (Autoregressive Integrated Moving Average Model) is a classical statistical model for time series forecasting [46] that has been demonstrated to be highly efficient in short-term forecasting [47]. Due to its effectiveness in revealing seasonal variations and long-term trends, ARIMA will be used to address the issue of long-term continuous missing values in SWH time series records. Comprising three main components: the autoregressive (AR) term, the differencing (I) term, and the moving average (MA) term, the ARIMA model is typically as denoted ARIMA (*p, d, q*). Its mathematical expression is outlined in references [48–50].

$$\left(1 - \sum_{i=1}^{p}\varnothing_iL^i\right)(1 - L)^d y_t = \left(1 + \sum_{i=1}^{q}\theta_tL^i\right)\varepsilon_t \tag{8}$$

where, L is the lag operator; d is the order of differencing; p is the order of the autoregressive part; q is the order of the moving average part; $\varnothing_i$ are the coefficients of the autoregressive terms; θ_i are the coefficients of the moving average terms; ε_t is the error term at time t.

2.3. LSTM

LSTM is a deep learning network model that is primarily employed to solve time series prediction problems. Due to its ability to capture complex non-linear patterns and the long-term dependencies inherent in such data, LSTM has significant advantages in non-linear time series prediction. The challenge of gradient vanishing is overcome in LSTM by integrating of forgetting, input, and output gates, along with unit states. These gates allow the retention of newly acquired information while

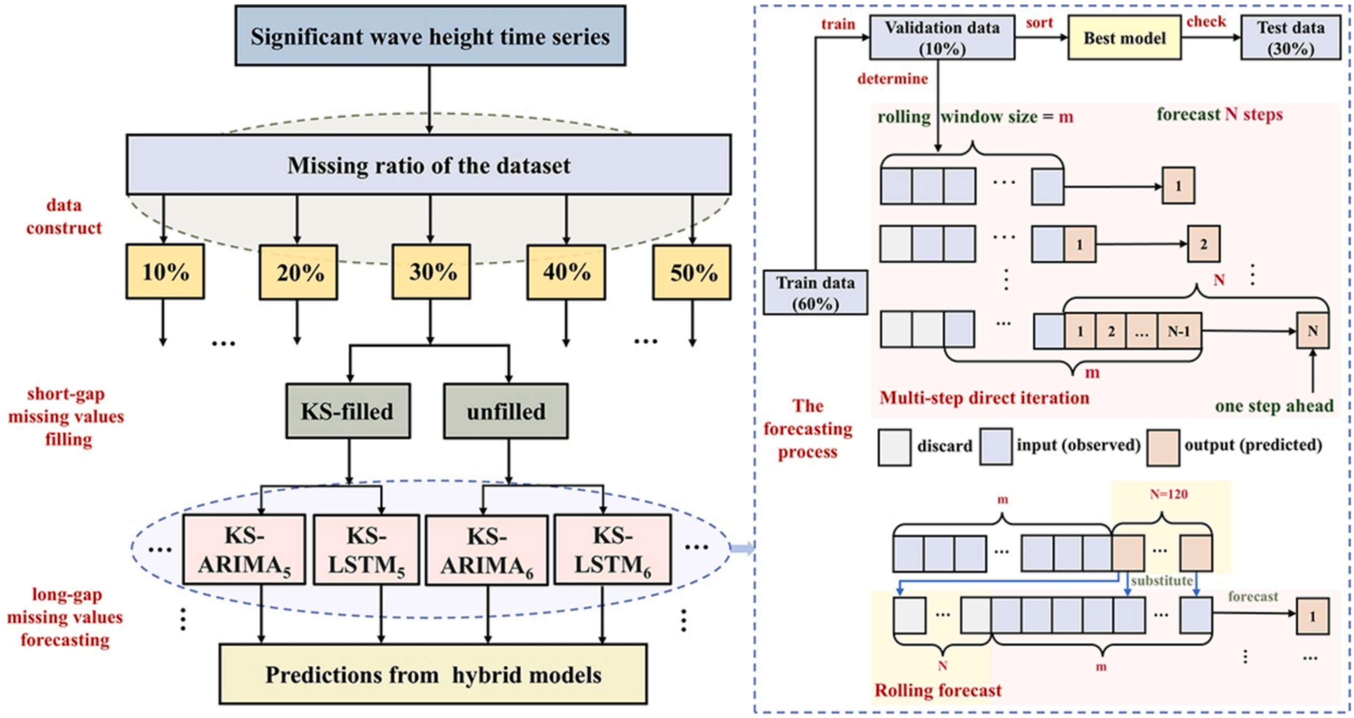


Fig. 2. KS-LSTM flow chart to fill in the missing values of SWH.

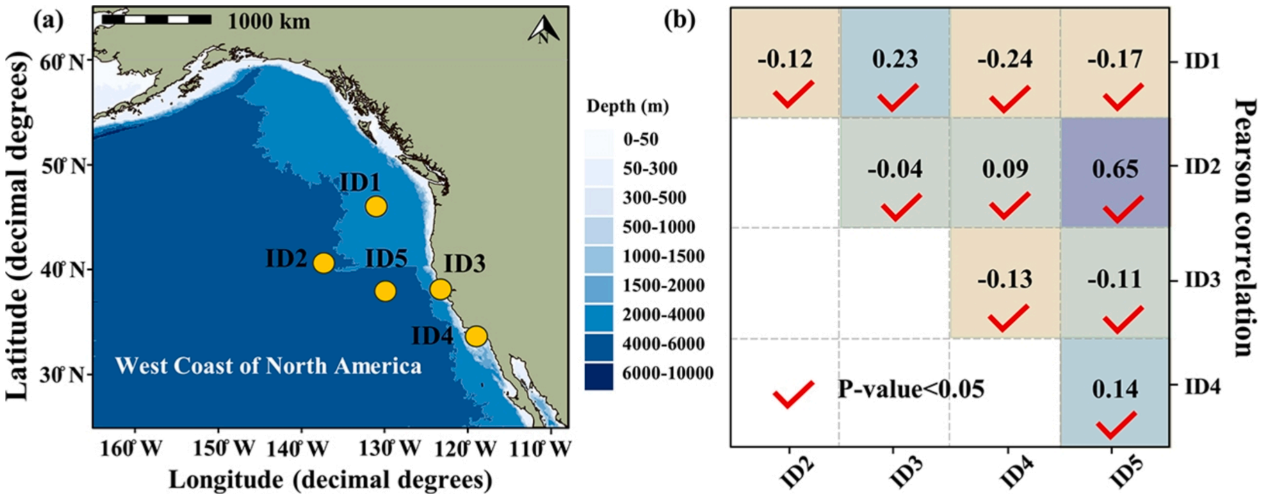


Fig. 3. (a) Geographic location of the selected buoys. (b) Pearson coefficient matrix between buoys.

preventing critical data loss [51,52]. The structure of LSTM is illustrated in Fig. 1, with the formulation provided below.

$$\begin{cases} f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \\ i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ Z_t = \tanh(W_z \cdot [h_{t-1}, x_t] + b_z) \\ C_t = f_t * C_{t-1} + i_t * Z_t \\ o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t = o_t * \tanh(C_t) \end{cases} \tag{9}$$

As shown in the above equations, the weight matrix $\left(W_f, W_i, W_z, W_o\right)$, the bias vector $\left(b_f, b_i, b_z, b_o\right)$, and the activation function tanh are used to represent the forgetting gate f_t, input gate i_t, output gate o_t, and cell state C_t in the LSTM.

Depending on research needs, an LSTM network model implemented within the TensorFlow deep learning framework is used in this work. An Adam optimizer with a learning rate of 0.001 is used, and the model is configured with 20 hidden layers, each comprising 50 neurons. To mitigate model overfitting and enhance generalization capability, a culling layer with a 20 % probability of deactivation for each neuron is added. The training process is completed after 2000 epochs, achieving a mean squared error (MSE) of 0.0254.

2.4. The proposed model (KS-LSTM)

This work aims to construct a novel KS-LSTM model that fills the short- and long-term missing values in the SWH time series simultaneously and efficiently. As depicted in Fig. 2, the process of the proposed KS-LSTM model is mainly divided into three stages, which are described as follows.

Step 1 KS-based short-term missing values filling: To assess the impact of datasets with varying missing ratios on the model's filling performance, random missing ratios ranging from 10 % to 50 % are introduced in this work, forming several datasets characterized by short-term intermittent missing values at different ratios. Two different treatments are adopted for these datasets, as one involves filling the datasets for each missing level using the KS method, and the other involves ignoring the missing ratio and directly deleting all missing values.

Step 2 LSTM/ARIMA-based prediction of long-term continuous missing values: After the intermittent missing values are filled using the KS method, the second phase involves a study on predicting long-term continuous missing values using LSTM and ARIMA. And then, the long-term missing values filling performance by using the LSTM/ARIMA without KS will be conducted to assess the effectiveness of the KS filling technique. The results reveal the impact of filling in short-term missing values on the improvement of accuracy in predicting continuous long-term missing values.

Step 3 Comparison of model inputs and results: The third stage involves a comprehensive assessment of the KS-LSTM performance in filling datasets with different ratios of missing values at different buoy stations. The generalizability and optimal application scenarios of the KS-LSTM are determined by comparing its performance in addressing both short- and long-term missing values and its performance with traditional methods such as LSTM or ARIMA independently.

2.5. Evaluation indicators of KS-LSTM model for filling in SWH missing values

To comprehensively evaluate the effectiveness of the KS-LSTM model on the missing values filling performance, three evaluation indicators are introduced in this study, including root mean square error (RMSE), mean square error (MSE) and mean absolute error (MAE). The RMSE is employed to quantify the deviation between predicted and actual values. It is calculated as the square root of the mean of the squared errors. A lower RMSE value indicates that the model's prediction accuracy has been improved [53]. The MSE is defined as the average of the squared errors, reflecting the overall prediction performance of the model by emphasizing larger errors. A smaller MSE indicates that the model's accuracy has been enhanced [54]. The MAE is calculated as the average of the absolute differences between the predicted and actual values, reflecting the average magnitude of error. MAE is used to directly assess the degree of deviation between predicted and true values, with smaller values indicating better accuracy [55]. Assuming x_i represents the true values, y_i represents the predicted values, and m is the sample size. The explicit mathematical formulas for these evaluation indicators for filling errors are as follows.

$$\begin{cases} RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i - y_i)^2} \\ MSE = \frac{1}{m}\sum_{i=1}^{m}(x_i - y_i)^2 \\ MAE = \frac{1}{m}\sum_{i=1}^{m}|x_i - y_i| \end{cases} \tag{10}$$

3. Study areas and materials

3.1. Study area and SWH data

The SWH data in this study are obtained from the buoy data provided

Table 1
The information of buoys and the SWH statistical characteristics.

ID	Site code	Location	Min (m)	Max (m)	Mean (m)	Std (m)	Depth (m)	Data amount
1	46,005	(46°8′36″ N 131°5′24″ W)	0.65	11.00	2.69	1.39	2821	13,418
2	46,006	(40°45′52″ N 137°22′37″ W)	0.78	12.50	2.91	1.42	4347	15,322
3	46,013	(38°14′5″ N 123°19′1″ W)	0.49	9.59	2.18	0.98	127	14,555
4	46,025	(33°45′19″ N 119°2′42″ W)	0.39	6.15	1.19	0.53	890	17,615
5	46,059	(38°4′9″ N 129°58′34″ W)	0.82	11.70	2.50	1.20	4640	13,040

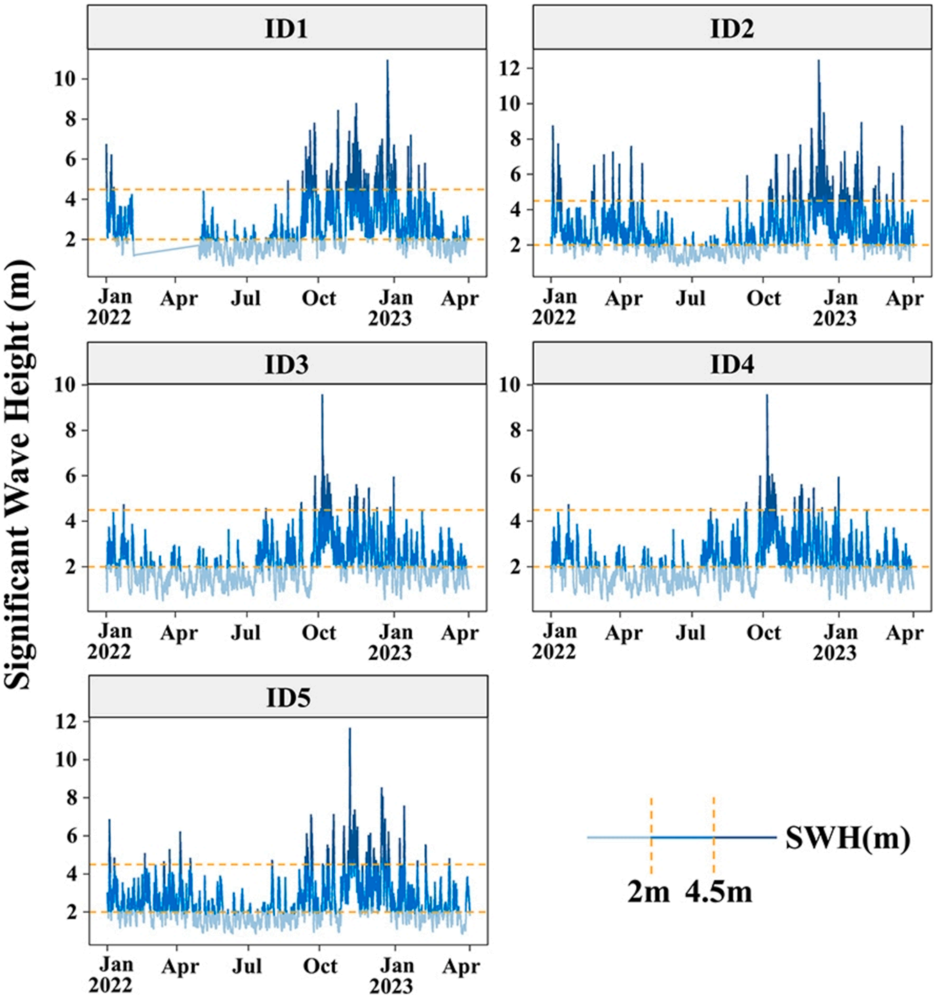


Fig. 4. Time series plot of significant wave height for five buoys.

by the National Data Buoy Center website of the United States. To validate the effectiveness of the hybrid model, buoys 46005, 46006, 46013, 46025, and 46,059 are selected based on spatial correlation, which covers different geographical areas of the west coast of the US, from Washington State to Southern California, so as to better increase the representativeness and comprehensiveness of the methodology in this research The SWH data are accessed on October 16, 2023, and Fig. 3 **(a)** shows the specific geographical location map of the buoys. The ID, coordinates, station code (NDBC code), water depth, and statistical characteristic information (minimum, maximum, mean, variance, data amount) [56] of the buoys are shown in Table 1.

The water depths of the buoys range from approximately 100–1000 m to 2500–5000 m. To verify that different characteristic patterns exhibited by the buoys, the Pearson correlation coefficient is calculated between the buoys using SWH data over a common time period. As shown in Fig. 3**(b)**, the numbers within the squares represent the Pearson coefficients, with more purple hues indicating larger coefficients. The red check marks denote a P-value less than 0.05 (the statistical significance level), indicating that the correlations between the buoy pairs are statistically significant based on the statistical test.

Fig. 4 visualizes the time series of SWH at five buoy sites from January 2022 to April 2023. Each sub-figure shows the change in wave height at the corresponding site during this period, with the orange dashed lines marking the reference wave heights of 2 m and 4.5 m, respectively. The SWH in Buoy ID1 is vacant from 13:00 on February 13, 2022, to 9:00 on May 20, 2022, while the range of missing values at other buoys is small. The SWH of buoys ID1, ID2, ID3, and ID5 are mainly concentrated between 2 m and 4.5 m, with wave height increasing significantly in winter. The SWH of buoy ID4 is generally low, mainly concentrated below 2 m, and the wave height remains relatively stable throughout the year. The frequency histogram in Fig. 5 further shows the overall distribution characteristics of SWH. Different color indicates the intensity of the frequency different SWH appears, from light pink (low distribution frequency) to dark purple (high distribution

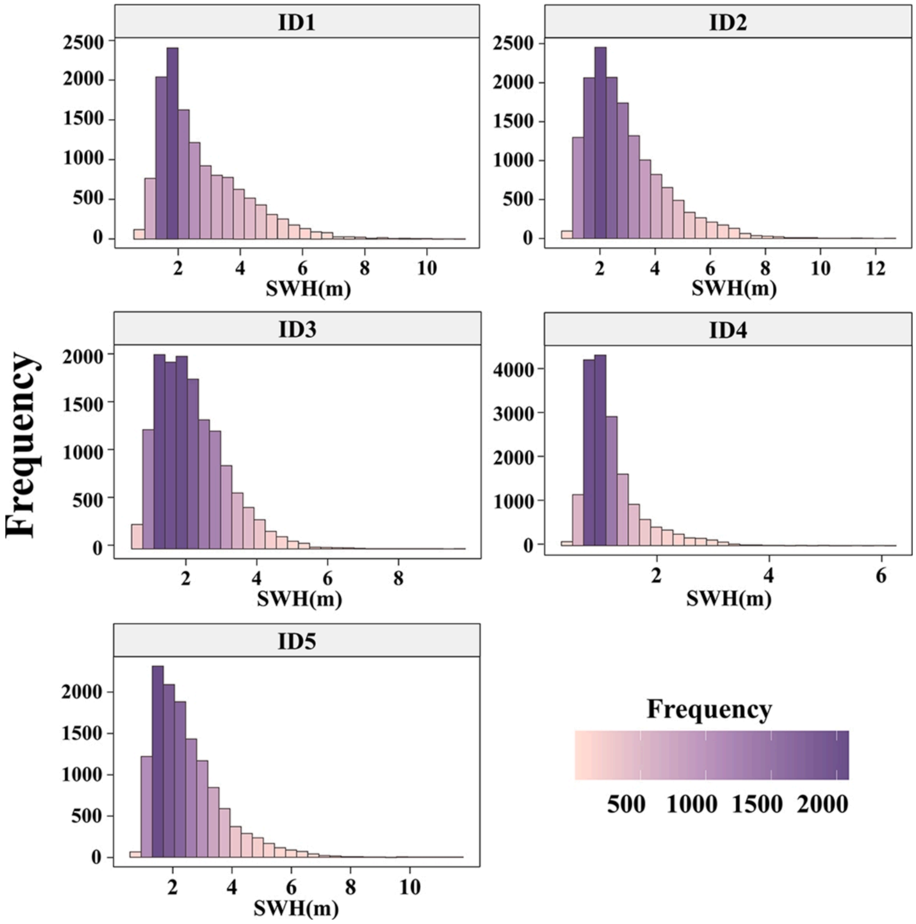


Fig. 5. Histogram of the distribution frequency of different SWH in the five buoys.

frequency). The SWH distribution frequency for buoys ID1, ID2, and ID5 are concentrated between 1 and 4 m, with wave heights above 4 m being relatively rare. Compared to these three buoys, the SWH distribution for buoy ID3 is observed to be slightly wider, with a higher frequency of SWH between 4 and 6 m. The peak SWH distribution frequency, exceeding 4000, is recorded at buoy ID4, indicating that waves of a certain height are very frequently encountered at this location.

3.2. *Simulation of missing datasets*

Since actual values for the missing data are unavailable, directly comparing the filled data with real values becomes infeasible. For this reason, datasets with various missing ratios are artificially constructed to simulate potential missing patterns in real SWH records. Specifically, based on the 16 months SWH sample series of complete records from buoys ID1 to ID5, five different ratios of SWH missing values (NA), with range of 10 %-50 %, are randomly introduced to simulate the actual short-term missing values of SWH.

Short intermittent missing values are defined as situations where the number of continuous missing values in a data set does not exceed 10 h. Long-term missing values are only considered when the missing duration extends to 120 continuous hours in this research. This definition helps in identifying and analyzing the patterns and characteristics of short-term intermittent missing values in the data, and in avoiding confusion between short-term missing values and long-term continuous missing values during processing and prediction. The short-term sizes of randomly generated missing values and their frequency of occurrence under different missing ratio conditions are depicted in Fig. 6 to visualize the data missing patterns. In these graphs, the vertical coordinates are sorted by term size (NAs in a row). The orange bar in the graphs represents the number of times a missing interval of a particular size occurs in the data, while the green bar indicates the number of missing values for each short-term interval Fig. 6**(a)** presents four different sizes of missing intervals in the dataset with 10 % missing ratio. The distribution of different short-term is shown in Fig. 6**(b)-(e)**, where the frequency of 1NA-short-term is the most prominent. Fig. 6**(b) and (e)** further illustrate the short-term containing larger short-term sizes, such as 10 missing values respectively. However, these larger short-terms represent only a small portion of the dataset, 1NA-short-term is the most prevalent among the missing value types.

4. Results and discussion

4.1. *Performance of short-term missing values filling*

To explore the effectiveness of KS in filling short-term intermittent missing values in SWH time series records, the data with five different missing ratios created previously are filled and analyzed compared to the cubic spline interpolation (CSPI) method. The evaluation indicators of the KS model and the CSPI method, after filling in the short-term missing values, are presented in Table 2.

As shown in Table 2, the maximum RMSE values for the KS and CSPI are recorded at 0.201 and 0.276, respectively. The MSE range for the KS is maintained between 0.007 and 0.041, while for the CSPI ranges between 0.012 and 0.076. The maximum MAE of KS and CSPI are noted at 0.137 and 0.189, respectively. It is observed that KS consistently maintains higher filling accuracy under different missing ratios and at different buoy stations compared to the CSPI method. Additionally, the error metrics for the KS and CSPI methods exhibit a tendency to remain relatively stable as the missing ratio increases. This stability can be

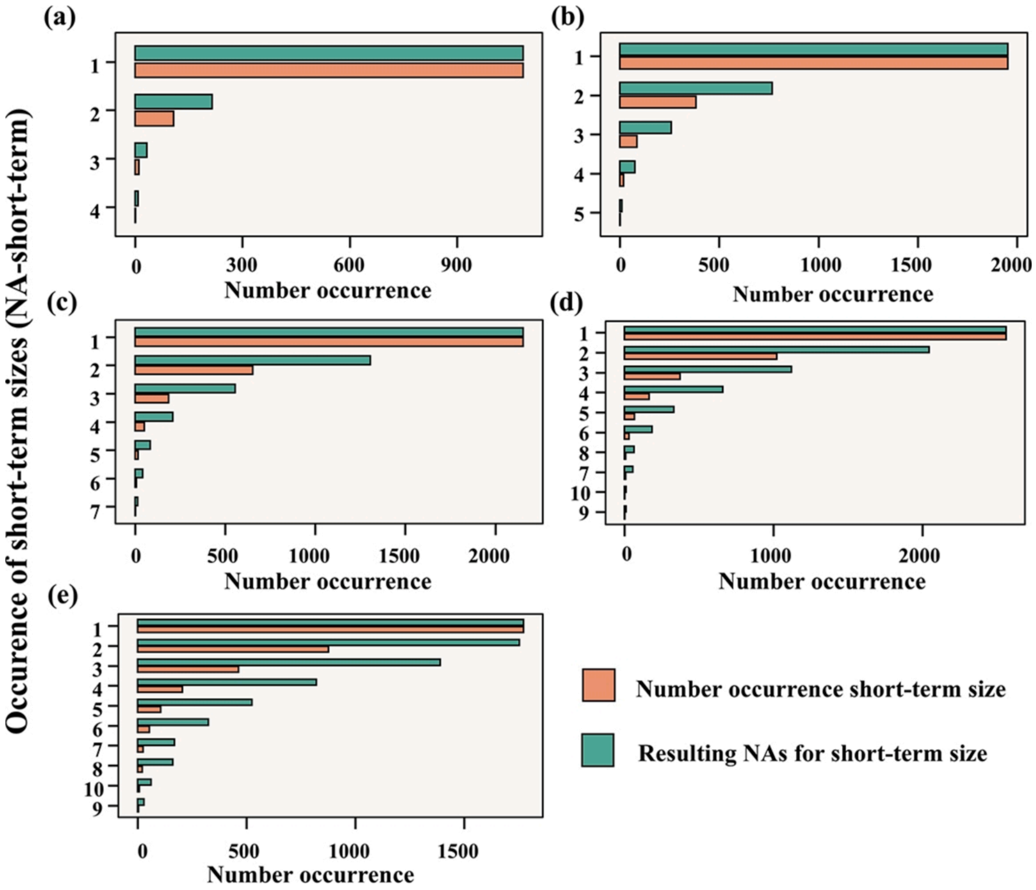


Fig. 6. SWH missing values distribution map: The data distribution with (a) 10% missing ratio at buoy ID1, (b) 20% missing ratio at buoy ID2, (c) 30% missing ratio at buoy ID3, (d) 40% missing ratio at buoy ID4, (e) 50% missing ratio at buoy ID5.

Table 2
The evaluation indicators for spline-filled and KS-filled errors.

Buoy code	Missing ratio	CSPI-filled values			KS-filled values			Filling amount
		RMSE	MSE	MAE	RMSE	MSE	MAE	
ID1	10 %	0.243	0.059	0.167	0.179	0.032	0.124	1342
ID2	20 %	0.276	0.076	0.187	0.201	0.041	0.137	3064
ID3	30 %	0.199	0.040	0.139	0.144	0.021	0.102	4366
ID4	40 %	0.109	0.012	0.075	0.082	0.007	0.057	7046
ID5	50 %	0.240	0.057	0.161	0.180	0.033	0.124	6520

attributed to the favorable data distribution characteristics and the inherent robustness of the KS, which enables it to maintain low error levels even at high missing ratios. These findings indicate that the KS method is most effective when the data amount is substantial and the missing ratio is moderate, as observed at buoy ID4.

The error distribution diagrams in Fig. 7 are presented in the form of box plots in a unified manner. Each color in the pair group represents a specific combination of data pair. For instance, “10 %NA. CSPI” or “30 % NA. KS” indicate datasets with 10 % missing ratio filled by using CSPI, or 30 % missing ratio filled by using KS, respectively. As the missing ratio increases from 10 % to 50 %, the error distribution is gradually expanded, indicating that with a higher missing ratio, greater errors and increased data dispersion are observed. When the same missing ratio is considered, it is observed that the box plot for the KS method is shorter than that for CSPI, and the median error and overall distribution are found to be lower than those for CSPI. This suggests that the error distribution of data points after being filled by the KS method is more concentrated, while the data points filled by CSPI are found to contain more outliers. The effectiveness of the KS method in filling short-term missing values is thereby further demonstrated.

4.2. Performance of long-term missing values filling

Based on the filling of datasets with different short-term missing ratios mentioned above, research on the prediction ability of SWH long-term continuous missing values is conducted in this section. It should be noted that, given the similarity in applicability and validity of the SWH data obtained at buoys ID1-ID5, and for the sake of clarity in the discussion process, only buoy ID1, which contains 10 % missing ratio, and buoy ID5, which contains 50 % missing ratio, are used as examples to display the detailed analysis results. Finally, the predictive performance of different forecasting methods is compared and analyzed at three buoys under conditions where the SWH missing ratios range from 10 % to 50 %. To ensure the effectiveness and generalization ability of the model, the datasets are divided into three parts: 60 % of the data is used as a training set, 30 % of the data is used as a test set, and the remaining 10 % is used as a validation set. This division helps us comprehensively evaluate the performance of the model on different data subsets and

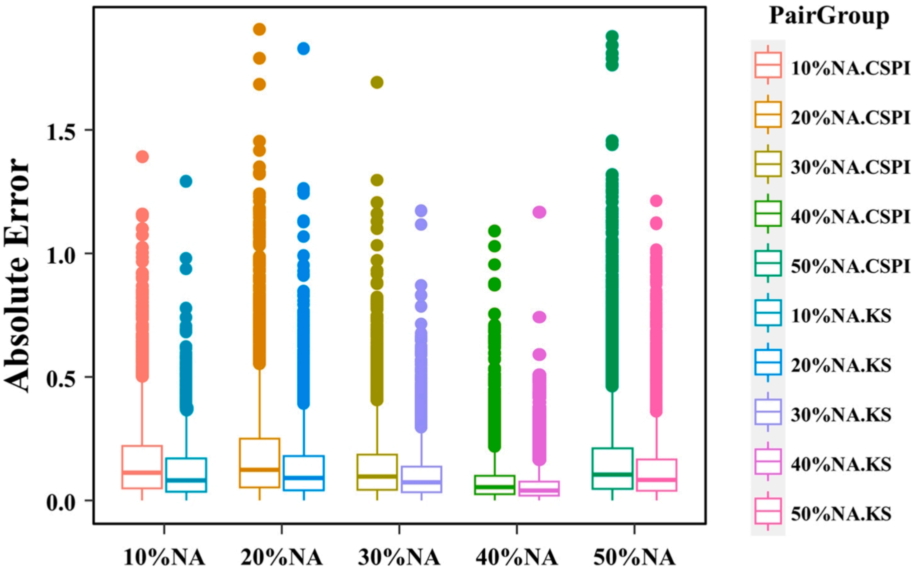


Fig. 7. Boxplot of absolute error distributions for missing ratios (10%-50%) for the five buoys SWH data calculated using KS and CSPI.

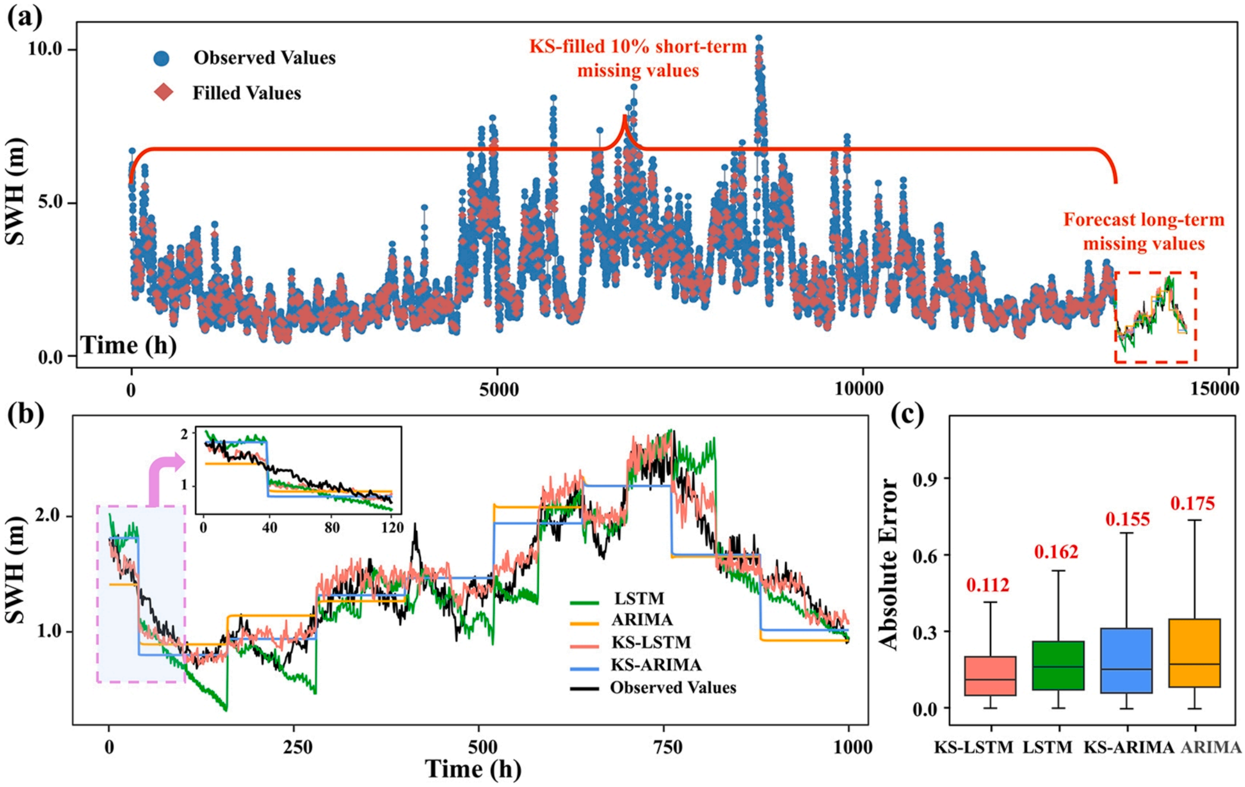


Fig. 8. Filling plot for short- and long-term missing values: (a) Complete time series plot of SWH missing values filled by ARIMA, KS-ARIMA, LSTM, and KS-LSTM with 10% missing ratio at buoy ID1, (b) Zoomed-in plots of prediction of long-term missing values, and (c) Box plots of absolute prediction errors for each model.

ensure the stability and reliability of the model in practical applications.

4.2.1. *Prediction results with 10 % missing ratio at buoy ID1*

Fig. 8 illustrates the distribution of observed values and filled values for both short- and long-term across the entire time series in cases of a 10 % missing ratio. In Fig. 8, the blue dots on the left part of the time series are the actual observed values. The red dots are used to represent short-term intermittent missing values filled by KS. The distribution of these dots on the vertical axis demonstrates the non-linear characteristics of the SWH and the volatility of the filled values. The prediction range of long-term missing values is labeled on the right side of Fig. 8**(a)**. Fig. 8**(b)** displays the predicted SWH long-term continuous missing values using four models, along with comparison among the results of these models and the actual observed values. The inset in the upper left corner of the figure shows a magnification of 120 consecutive hours of long-term missing values to facilitate the observation of details. It follows that although ARIMA can capture the overall trend prediction of SWH, its ability to capture extreme values is limited and prone to delays and prediction errors when faced with rapid and large changes. By incorporating KS, the overall trend tracking of KS-ARIMA is made

Table 3
The evaluation error indicators for four models predicting long-term missing values in ID1.

Buoy Code	Missing ratio	Method	*RMSE*	*MSE*	*MAE*
ID1	10 %	ARIMA	0.308	0.095	0.238
		LSTM	0.241	0.058	0.187
		KS-ARIMA	0.273	0.074	0.206
		KS-LSTM	0.168	0.028	0.134
	20 %	ARIMA	0.330	0.109	0.248
		LSTM	0.310	0.096	0.227
		KS-ARIMA	0.304	0.092	0.219
		KS-LSTM	0.172	0.029	0.136
	30 %	ARIMA	0.350	0.123	0.264
		LSTM	0.333	0.111	0.276
		KS-ARIMA	0.319	0.102	0.228
		KS-LSTM	0.174	0.030	0.139
	40 %	ARIMA	0.418	0.175	0.319
		LSTM	0.345	0.119	0.267
		KS-ARIMA	0.347	0.121	0.260
		KS-LSTM	0.230	0.053	0.193
	50 %	ARIMA	0.443	0.196	0.335
		LSTM	0.461	0.213	0.383
		KS-ARIMA	0.371	0.138	0.280
		KS-LSTM	0.187	0.035	0.149

smoother, particularly in areas with significant data fluctuations. However, due to the linear limitations of ARIMA, KS-ARIMA cannot fully capture rapid changing peaks and troughs when processing complex data. In contrast, excellent trend-following capabilities and stability are exhibited by KS-LSTM, especially in continuous fluctuating data segments. While LSTM generally simulates wave peaks and troughs effectively, a delay in its response is still present in extreme or rapidly changing situations. When combining the long-term dependency processing capabilities of LSTM with the error correction advantages of KS, the KS-LSTM model is rendered the most outstanding in terms of dynamic response capabilities. It not only tracks the overall trend of the data accurately but also efficiently captures the peaks and troughs.

The box plot in Fig. 8**(c)** further demonstrates the performance comparison of the four models in the predicting absolute error distribution. KS-LSTM is shown to have the lowest error, recorded at 0.112, followed by KS-ARIMA and LSTM, and the ARIMA exhibiting the highest median error and the widest error range. The research reveals that significant improvements in the overall prediction performance of the models are achieved when KS is combined (KS-ARIMA and KS-LSTM), particularly in capturing the nonlinear characteristics of the SWH. Compared to KS-ARIMA, KS-LSTM displays a more significant advantage in handling dynamic changes of peaks and troughs. It can be verified that the prediction stability and accuracy of the model for long-term continuous missing values can be significantly enhanced by incorporating KS.

As observed from Table 3, the lowest prediction error is achieved by KS-LSTM when the missing ratio is 10 %, with RMSE, MAE, and MSE recorded as 0.168, 0.134, and 0.028, respectively. In contrast, RMSE, MAE, and MSE of KS-ARIMA are found to be 0.273, 0.206, and 0.074, respectively. The RMSE values when only LSTM and ARIMA are used are noted to be 0.241 and 0.304, respectively. As the missing ratio increases, the prediction error is also observed to increase. When the missing ratio reaches 50 %, the RMSE of KS-LSTM is 0.187, while that of KS-ARIMA rises to 0.371. As illustrated in Fig. 9, reductions in RMSE, MAE, and MSE of KS-LSTM predictions, compared with using only LSTM, are observed to reach up to 59.4 %, 61.1 %, and 83.6 %, respectively. Compared to use only ARIMA, reductions in RMSE, MAE, and MSE of KS-LSTM predictions are recorded as up to 57.8 %, 55.5 %, and 82.1 %, respectively. Furthermore, reductions in RMSE, MAE, and MSE of KS-LSTM predictions, compared with KS-ARIMA, are noted to be up to 49.6 %, 74.6 %, and 46.8 %, respectively. The KS-LSTM model is demonstrated to perform significantly better than other models in handling different missing ratio conditions, particularly in high missing

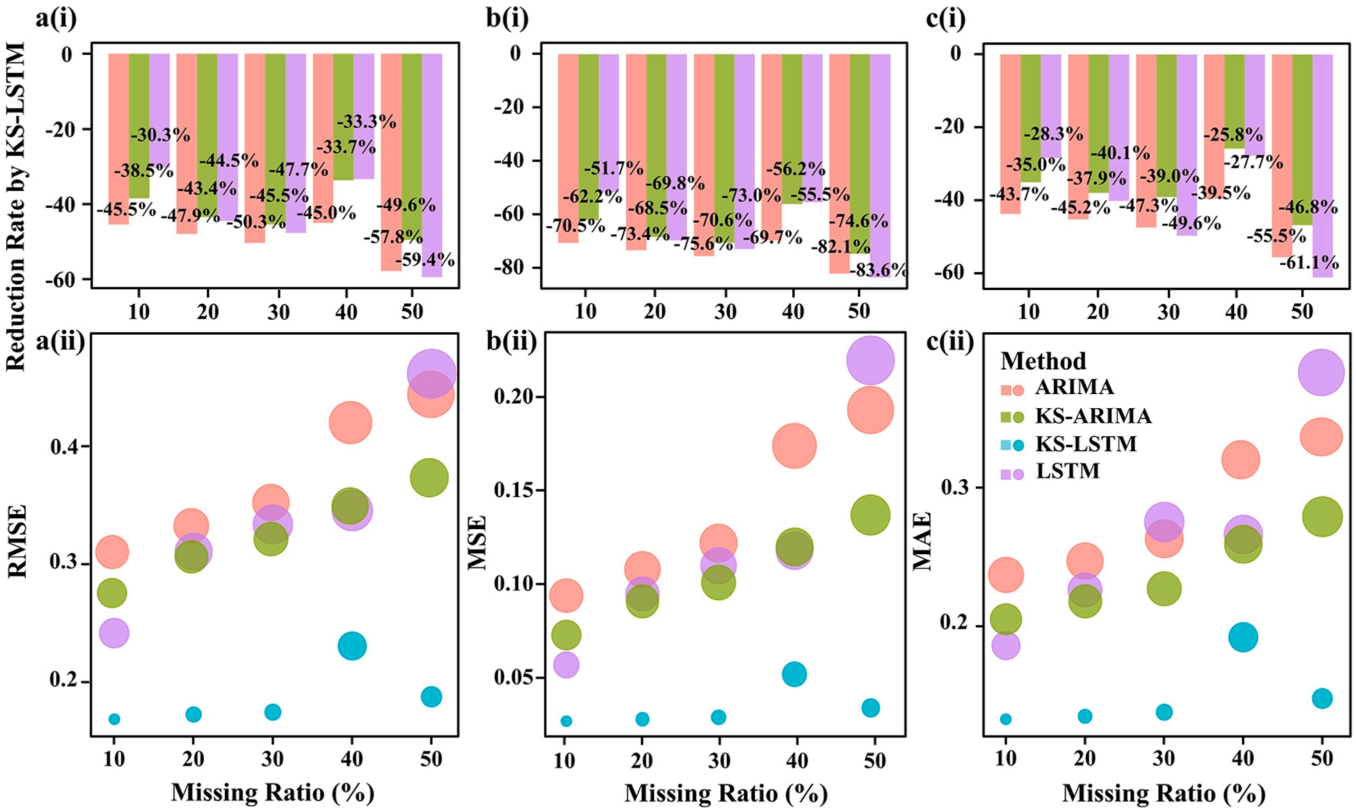


Fig. 9. The bubble plots of the distribution of prediction errors for buoy ID1with 10%-50% missing ratios of a(i) RMSE, b(i) MSE, c(i) MAE; the error reduction rate (%) by KS-LSTM compared to other models of a(ii) RMSE, b(ii) MSE, c(ii) MAE.

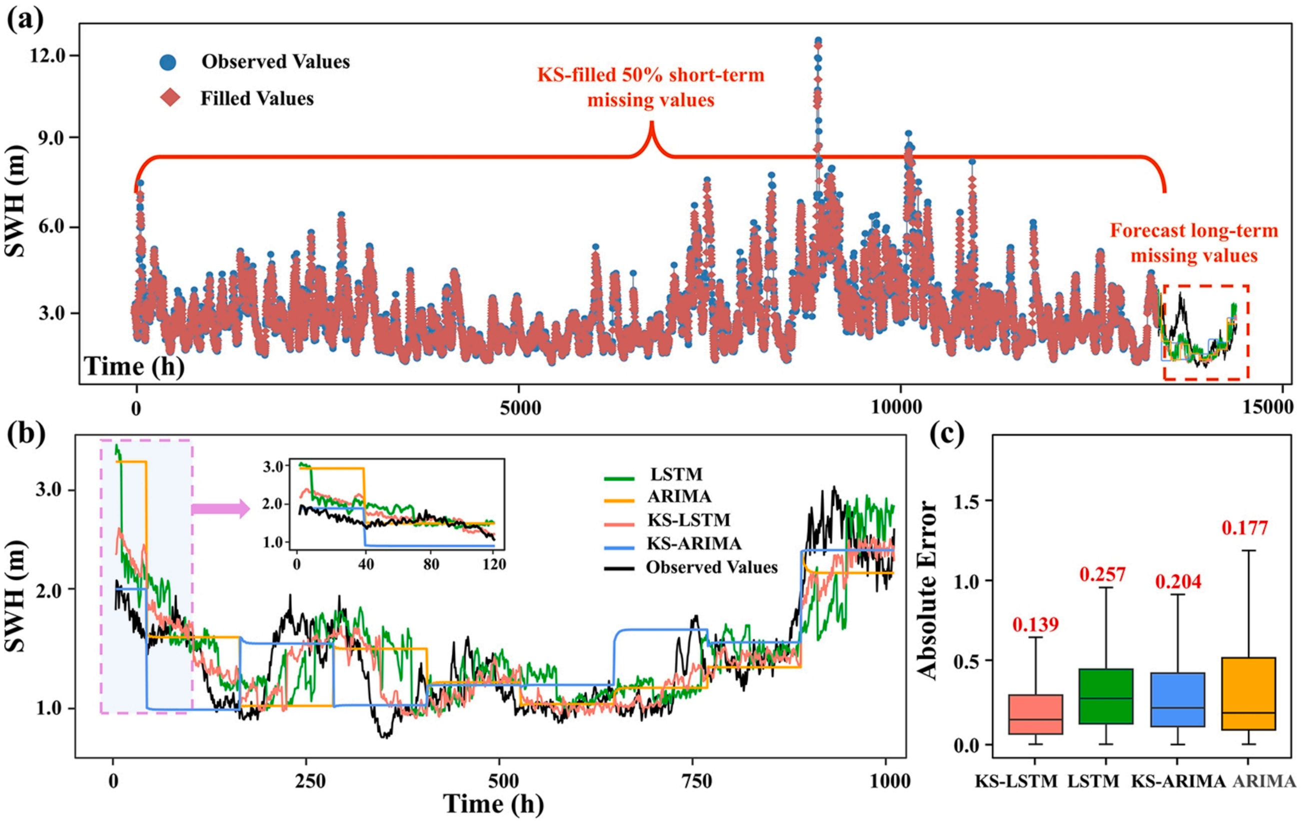


Fig. 10. Filling plot for short- and long-term missing values: (a) Complete time series plot of SWH missing values filled by ARIMA, KS-ARIMA, LSTM, and KS-LSTM with 50% missing ratio at buoy ID5, (b) Zoomed-in plots of prediction of long-term missing values, and (c) Box plots of absolute prediction errors for each model.

Table 4
The evaluation error indicators for four models predicting long-term missing values in ID5.

Buoy Code	Missing ratio	Method	*RMSE*	*MSE*	*MAE*
ID5	10 %	ARIMA	0.381	0.145	0.281
		LSTM	0.330	0.109	0.244
		KS-ARIMA	0.321	0.103	0.229
		KS-LSTM	0.228	0.052	0.170
	20 %	ARIMA	0.399	0.159	0.290
		LSTM	0.345	0.119	0.253
		KS-ARIMA	0.339	0.115	0.256
		KS-LSTM	0.245	0.060	0.183
	30 %	ARIMA	0.381	0.145	0.281
		LSTM	0.403	0.163	0.313
		KS-ARIMA	0.358	0.128	0.273
		KS-LSTM	0.248	0.061	0.181
	40 %	ARIMA	0.420	0.176	0.334
		LSTM	0.399	0.159	0.311
		KS-ARIMA	0.324	0.105	0.230
		KS-LSTM	0.250	0.063	0.184
	50 %	ARIMA	0.439	0.192	0.304
		LSTM	0.389	0.152	0.304
		KS-ARIMA	0.365	0.133	0.280
		KS-LSTM	0.262	0.069	0.193

ratio scenarios, where the reduction of prediction error is especially significant. Therefore, KS can effectively address noise and other uncertain factors in observation data, which significantly improve prediction accuracy, handle data incompleteness, and enhance the model's adaptability to SWH fluctuations.

4.2.2. Prediction results with 50 % missing ratio at buoy ID5

To examine the performance of the KS-LSTM model proposed in this article with higher short-term missing ratio, a detailed analysis is performed on the buoy ID5 dataset with 50 % missing ratio. Fig. 10**(a)** shows the distribution of observed values and filled values for both short- and long-term in the entire time series at 50 % missing ratio. By zooming in to observe Fig. 10**(b)** of the long-term prediction and Fig. 10 **(c)** of the error distribution, it is demonstrated that the prediction accuracy of all models decreases with the increase of missing ratio. The median error of KS-ARIMA is 0.204, but instability is shown when data with high volatility is dealt with. The median error of ARIMA is 0.177, but larger deviations are observed in areas with high volatility. Among all models, the largest error of 0.257 and the highest volatility are exhibited by LSTM, especially in the peak area. KS-LSTM has a relatively low absolute error of 0.139, and better adaptability is demonstrated when peaks and fluctuations in time series are handled. These results indicate that certain risk-resistance capabilities are provided to LSTM and ARIMA after filling short-term missing values by the KS method, which allowing these models to better maintain performance even under high missing ratios.

Table 4 illustrates that the performance of ARIMA and LSTM is significantly inferior to the other two models with 50 % missing ratio. KS-LSTM achieves the lowest RMSE of 0.262, MSE of 0.069, and MAE of 0.193. KS-LSTM demonstrates superior prediction accuracy when the missing ratio ranges from 10 % to 50 %, particularly when predicting extreme values. In brief, KS-LSTM exhibits best performance in processing complex time series with substantial missing data. Fig. 11 display all scatter plots of observed values and predicted long-term continuous missing values at buoy ID5. In all subplots, an ideal line (the red dashed line) represents a perfect prediction, where the predicted values are exactly equal to the observed values. The color of the dot represents the number of predictions. In Fig. 11, it is observed that the data points got by KS-LSTM are closer to the ideal line compared to that by other models, especially in areas with denser data points, indicating that the predicted results align more closely with the actual

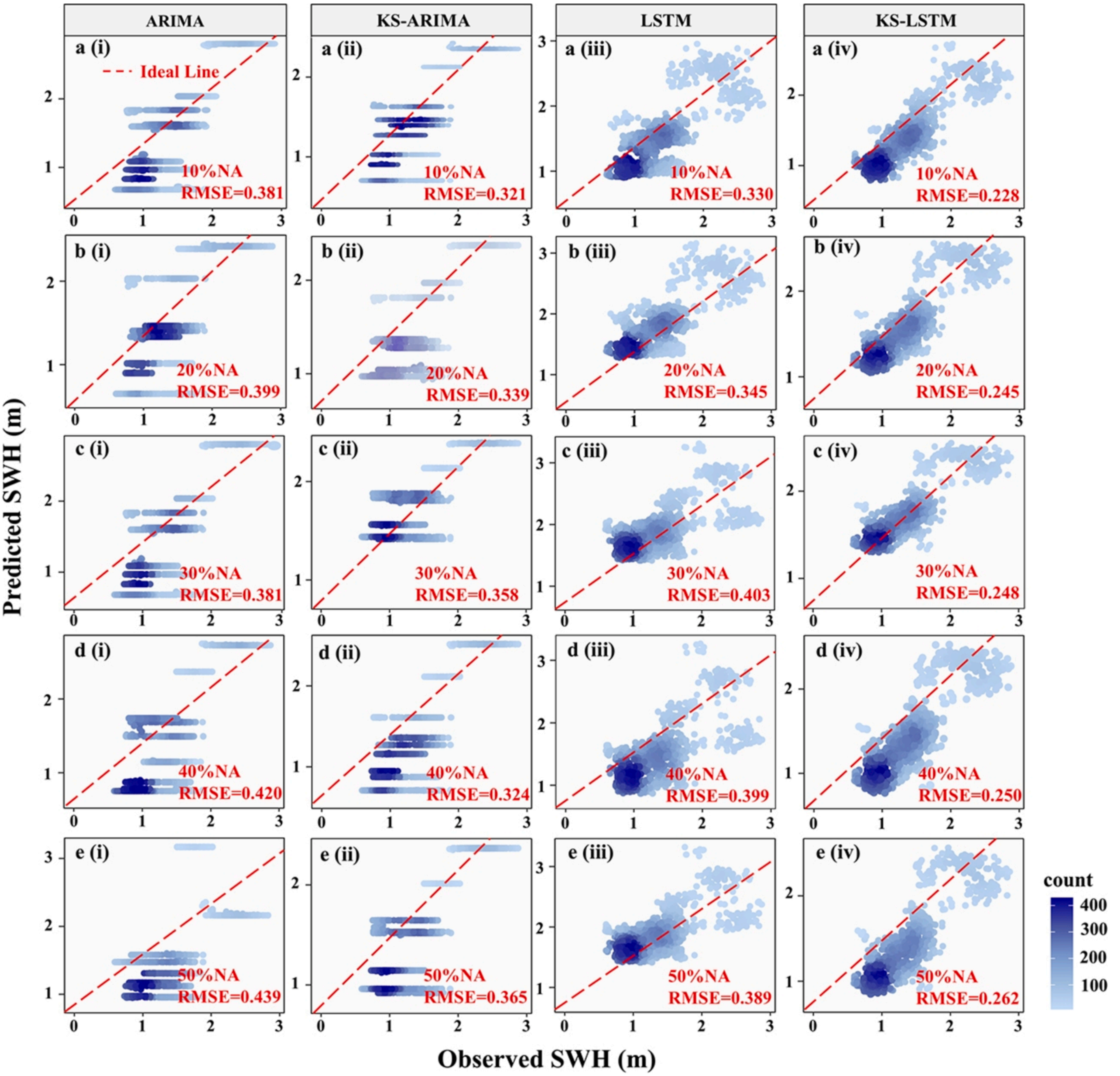


Fig. 11. Scatterplot of observed and predicted SWH (m) at buoy ID5 with 10%-50% missing ratios: Scatter distribution of a(i-v) KS-LSTM, b(i-v) LSTM, c(i-v) KS-ARIMA, d(i-v) ARIMA.

observed values. It illustrates that when the predicted SWH exceeds 3 m, the distribution of data points for LSTM and ARIMA is observed to be relatively scattered. Compared to the traditional LSTM and ARIMA, the KS-LSTM and KS-ARIMA, have more tightly distributed data points, are demonstrated to have superior prediction abilities. Overall, KS-LSTM exhibits higher accuracy than KS-ARIMA when dealing with larger datasets.

4.2.3. *Prediction results using other ratios of missing values*

To further verify the spatial generalization ability of KS-LSTM, the error performance (RMSE, MSE, MAE) of long-term continuous missing values is analyzed by applying different models (ARIMA, LSTM, KS-ARIMA, KS-LSTM) are for prediction at different missing ratios of 10 % – 50 % for different buoys (ID2, ID3, ID4). The evaluation indicators are shown in Table 5, Table 6 and Table 7 respectively. For buoy ID2, when the missing ratio is 10 %, the best performance in all indicators (RMSE: 0.190, MSE: 0.036, MAE: 0.145) is achieved by KS-LSTM, outperforming other models. Although the RMSE (0.224) and MAE (0.174) of LSTM are lower than those of ARIMA, they are still higher than those of KS-LSTM. As the missing ratio increased to 50 %, the best performance is still maintained by KS-LSTM, with an RMSE of 0.259, which is much lower than 0.416 by ARIMA and 0.358 by LSTM. Similar trends are shown by buoys ID3 and ID4. As the missing ratio increases, the prediction errors of each model gradually increase, but extremely high performance is still maintained by the proposed KS-LSTM under various test conditions. Especially under high missing ratio conditions, excellent prediction ability and significantly lower error are demonstrated by KS-LSTM. This fully proves that strong adaptability is possessed by KS-LSTM in time series prediction tasks with missing data, and excellent generalization ability is also shown when dealing with incomplete data.

Table 5
The evaluation error indicators for three models predicting long-term missing values in ID2.

Buoy code	Missing ratio	Method	RMSE	MSE	MAE
ID2	10 %	ARIMA	0.323	0.104	0.234
		LSTM	0.224	0.050	0.174
		KS-ARIMA	0.268	0.072	0.203
		KS-LSTM	0.190	0.036	0.145
	20 %	ARIMA	0.343	0.117	0.242
		LSTM	0.260	0.068	0.203
		KS-ARIMA	0.275	0.076	0.207
		KS-LSTM	0.183	0.033	0.141
	30 %	ARIMA	0.372	0.138	0.262
		LSTM	0.317	0.100	0.245
		KS-ARIMA	0.311	0.097	0.242
		KS-LSTM	0.226	0.051	0.181
	40 %	ARIMA	0.383	0.147	0.273
		LSTM	0.320	0.120	0.227
		KS-ARIMA	0.347	0.120	0.260
		KS-LSTM	0.202	0.041	0.155
	50 %	ARIMA	0.416	0.173	0.310
		LSTM	0.358	0.128	0.264
		KS-ARIMA	0.359	0.129	0.259
		KS-LSTM	0.259	0.067	0.216

Table 6
The evaluation error indicators for three models predicting long-term missing values in ID3.

Buoy code	Missing ratio	Method	RMSE	MSE	MAE
ID3	10 %	ARIMA	0.427	0.182	0.312
		LSTM	0.341	0.116	0.270
		KS-ARIMA	0.366	0.134	0.275
		KS-LSTM	0.252	0.064	0.195
	20 %	ARIMA	0.440	0.194	0.312
		LSTM	0.367	0.135	0.279
		KS-ARIMA	0.373	0.139	0.295
		KS-LSTM	0.254	0.065	0.193
	30 %	ARIMA	0.457	0.209	0.343
		LSTM	0.376	0.141	0.286
		KS-ARIMA	0.431	0.186	0.334
		KS-LSTM	0.255	0.065	0.198
	40 %	ARIMA	0.488	0.238	0.361
		LSTM	0.400	0.160	0.311
		KS-ARIMA	0.465	0.216	0.348
		KS-LSTM	0.258	0.067	0.200
	50 %	ARIMA	0.556	0.310	0.411
		LSTM	0.418	0.175	0.319
		KS-ARIMA	0.505	0.255	0.366
		KS-LSTM	0.270	0.073	0.211

5. Conclusion

To address the simultaneous existence of short-term intermittent and long-term continuous missing values in SWH time series data, this work proposes an innovative hybrid model that combines KS with LSTM, which is named KS-LSTM, to enhance the filling accuracy. In this approach, short-term intermittent missing values in SWH are effectively filled through real-time recursive calculation and signal smoothing by KS, while the impact of random noise is reduced. Long-term information is remembered and retained, and long-term trends in SWH time series are accurately captured by LSTM using its unique gating structure. Consequently, after the short-term missing values are filled by KS, LSTM can use its modeling ability of complex sequences to more accurately predict SWH long-term continuous missing values. The main findings are concluded as follows.

(1) **Comparison of the filling effect of short-term missing values:** Compared to traditional cubic spline interpolation, the KS method shows a significant advantage in filling intermittent short-term missing values in the SWH time series. The RMSEs for the KS method are found to be 0.179 and 0.180 for 10 % and 50 % missing data, respectively, in contrast to 0.243 and 0.240 for the CSPI method. This indicates that the KS method is more effective in minimizing errors, even when dealing with data with a high missing ratio.

Table 7
The evaluation error indicators for three models predicting long-term missing values in ID4.

Buoy code	Missing ratio	Method	RMSE	MSE	MAE
ID4	10 %	ARIMA	0.140	0.020	0.108
		LSTM	0.093	0.009	0.073
		KS-ARIMA	0.120	0.014	0.098
		KS-LSTM	0.091	0.008	0.072
	20 %	ARIMA	0.132	0.018	0.105
		LSTM	0.102	0.010	0.082
		KS-ARIMA	0.115	0.013	0.092
		KS-LSTM	0.093	0.009	0.074
	30 %	ARIMA	0.140	0.020	0.110
		LSTM	0.115	0.013	0.091
		KS-ARIMA	0.117	0.014	0.091
		KS-LSTM	0.097	0.009	0.076
	40 %	ARIMA	0.146	0.021	0.116
		LSTM	0.122	0.015	0.099
		KS-ARIMA	0.131	0.017	0.107
		KS-LSTM	0.112	0.013	0.088
	50 %	ARIMA	0.150	0.022	0.112
		LSTM	0.138	0.019	0.111
		KS-ARIMA	0.135	0.018	0.105
		KS-LSTM	0.125	0.016	0.100

(2) **Comparison of the prediction effect of long-term missing values:** Compared to KS-ARIMA, LSTM, and ARIMA, KS-LSTM exhibits a maximum RMSE reduction of 49.6 %, 59.4 %, and 57.8 % at short-term intermittent missing ratios of 10 %-50 %, respectively. The maximum reduction in MSE is observed to be 71.4 %, 83.6 %, and 82.1 %. Similarly, the maximum reduction in MAE is 74.6 %, 61.1 %, and 55.5 %. These results indicate that KS-LSTM can more effectively reduce prediction errors and demonstrate strong generalization ability when handling long-term missing values.

(3) **Impact of different short-term missing ratios on model performance:** As the missing ratio increases, KS-LSTM demonstrates stable prediction performance with relatively low errors. Even when the missing ratio of buoy ID4 is as high as 50 %, the RMSE, MSE, and MAE of KS-LSTM are only 0.125, 0.016 and 0.100, respectively. This suggests that KS-LSTM is highly suitable for scenarios requiring high-precision prediction, especially when dealing with complex time series containing a significant amount of missing data.

This research confirms that significant effectiveness in handling missing values in SWH time series data is exhibited by the model combining KS with LSTM. Especially in scenarios where short- and long-term missing values coexist, it is shown that the model not only performs well at multiple measurement points, but also maintains prediction accuracy and robustness even when the data missing ratio is high. Future research can be further explored through the combination of KS with other advanced time-series prediction models (such as other network structures in deep learning) to enhance the model's applicability and prediction accuracy across a wider range of data types.

CRediT authorship contribution statement

Yulian Wang: Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization. **Taili Du:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Yuanye Guo:** Validation, Investigation, Data curation. **Fangyang Dong:** Writing – review & editing, Software, Methodology. **Jicang Si:** Writing – review &

editing, Methodology, Conceptualization. **Minyi Xu:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

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 by the National Key R & D Project from the Minister of Science and Technology (2021YFA1201604), the National Natural Science Foundation of China (Grant Nos. 52101345, 52371345), the China Postdoctoral Science Foundation (Grant No. 2024M750297), the Dalian Outstanding Young Scientific and Technological Talents Project (Grant No. 2021RJ11), the Application Research Program of Liaoning Province (Grant No. 2022JH2/01300219), and the Fundamental Research Funds for the Central Universities (Grant No. 3132024210).

Data availability

Data will be made available on request.

References

[1] M. Melikoglu, Current status and future of ocean energy sources: A global review, Ocean Eng. 148 (2018) 563–573, https://doi.org/10.1016/j.oceaneng.2017.11.045.

[2] T. Wilberforce, Z. El Hassan, A. Durrant, J. Thompson, B. Soudan, A.G. Olabi, Overview of ocean power technology, Energy. 175 (2019) 165–181, https://doi.org/10.1016/j.energy.2019.03.068.

[3] C. Breyer, S. Khalili, D. Bogdanov, M. Ram, A.S. Oyewo, A. Aghahosseini, A. Gulagi, A.A. Solomon, D. Keiner, G. Lopez, P.A. Ostergaard, H. Lund, B. V. Mathiesen, M.Z. Jacobson, M. Victoria, S. Teske, T. Pregger, V. Fthenakis, M. Raugei, H. Holttinen, U. Bardi, A. Hoekstra, B.K. Sovacool, On the history and future of 100% renewable energy systems research, IEEE Access 10 (2022) 78176–78218, https://doi.org/10.1109/access.2022.3193402.

[4] N.L. Panwar, S.C. Kaushik, S. Kothari, Role of renewable energy sources in environmental protection: A review, Renew. Sust. Energ. Rev. 15 (2011) 1513–1524, https://doi.org/10.1016/j.rser.2010.11.037.

[5] M. Lehmann, F. Karimpour, C.A. Goudey, P.T. Jacobson, M.R. Alam, Ocean wave energy in the United States: Current status and future perspectives, Renew. Sust. Energ. Rev. 74 (2017) 1300–1313, https://doi.org/10.1016/j.rser.2016.11.101.

[6] J. Ding, F.Y. Deng, Q. Liu, J.C. Wang, Regional forecasting of significant wave height and mean wave period using EOF-EEMD-SCINet hybrid model, Appl. Ocean Res. 136 (2023) 16, https://doi.org/10.1016/j.apor.2023.103582.

[7] I. López, J. Andreu, S. Ceballos, I.M. de Alegría, I. Kortabarria, Review of wave energy technologies and the necessary power-equipment, Renew. Sust. Energ. Rev. 27 (2013) 413–434, https://doi.org/10.1016/j.rser.2013.07.009.

[8] M.A. Mustapa, O.B. Yaakob, Y.M. Ahmed, C.K. Rheem, K.K. Koh, F.A. Adnan, Wave energy device and breakwater integration: A review, Renew. Sust. Energ. Rev. 77 (2017) 43–58, https://doi.org/10.1016/j.rser.2017.03.110.

[9] J.M. Wu, L.Z. Qin, N. Chen, C. Qian, S.M. Zheng, Investigation on a spring-integrated mechanical power take-off system for wave energy conversion purpose, Energy. 245 (2022) 15, https://doi.org/10.1016/j.energy.2022.123318.

[10] E. Medina-Lopez, D. McMillan, J. Lazic, E. Hart, S. Zen, A. Angeloudis, E. Bannon, J. Browell, S. Dorling, R.M. Dorrell, R. Forster, C. Old, G.S. Payne, G. Porter, A. S. Rabaneda, B. Sellar, E. Tapoglou, N. Trifonova, I.H. Woodhouse, A. Zampollo, Satellite data for the offshore renewable energy sector: synergies and innovation opportunities, Remote Sens. Environ. 264 (2021) 26, https://doi.org/10.1016/j.rse.2021.112588.

[11] M. Abbas, Z.Y. Min, Z.Y. Liu, D.J. Zhang, Unravelling oceanic wave patterns: A comparative study of machine learning approaches for predicting significant wave height, Appl. Ocean Res. 145 (2024) 30, https://doi.org/10.1016/j.apor.2024.103919.

[12] P.C. Ho, J.Z. Yim, A study of the data transferability between two wave-measuring stations, Coast Eng. 52 (2005) 313–329, https://doi.org/10.1016/j.coastaleng.2004.12.003.

[13] O.S. Hidalgo, J. Nieto Borge, C.C. Cunha, C. Guedes Soares, Filling missing observations in time series of significant wave height. ASME, New York, 2 (1995), pp. 9-17.

[14] D.M. Zhu, J.X. Zhang, Q. Wu, Y. Dong, E. Bastidas-Arteaga, Predictive capabilities of data-driven machine learning techniques on wave-bridge interactions, Appl. Ocean Res. 137 (2023) 15, https://doi.org/10.1016/j.apor.2023.103597.

[15] X.X. Guo, X.T. Zhang, X.L. Tian, X. Li, W.Y. Lu, Predicting heave and surge motions of a semi-submersible with neural networks, Appl. Ocean Res. 112 (2021) 12, https://doi.org/10.1016/j.apor.2021.102708.

[16] M.N. Noor, A.S. Yahaya, N.A. Ramli, A.M.M. Al Bakri, Filling missing data using interpolation methods: Study on the effect of fitting distribution, Key Eng. Mater. 594–595 (2013) 889–895, https://doi.org/10.4028/www.scientific.net/KEM.594-595.889.

[17] M.E. Quinteros, S.Y. Lu, C. Blazquez, J.P. Cárdenas, X. Ossa, J.M. Delgado-Saborit, R.M. Harrison, P. Ruiz-Rudolph, Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile, Atmos. Environ. 200 (2019) 40–49, https://doi.org/10.1016/j.atmosenv.2018.11.053.

[18] F.A.R. Abdullah, N.S. Ningsih, T.M. Al-Khan, Significant wave height forecasting using long short-term memory neural network in Indonesian waters, J. Ocean Eng. Mar. Energy. 8 (2022) 183–192, https://doi.org/10.1007/s40722-022-00224-3.

[19] F. Scarpa, G. Milano, Kalman smoothing technique applied to the inverse heat conduction problem, Num. Heat. Tr. B-Fund. 28 (1995) 79–96, https://doi.org/10.1080/10407799508928822.

[20] N. Daouas, M.S. Radhouani, A new approach of the Kalman filter using future temperature measurements for nonlinear inverse heat conduction problems, Num. Heat. Tr. B-Fund. 45 (2004) 565–585, https://doi.org/10.1080/10407790490430598.

[21] N. Umar, A. Gray, Comparing single and multiple imputation approaches for missing values in univariate and multivariate water level data, Water. 15 (2023) 21, https://doi.org/10.3390/w15081519.

[22] F. Avanzi, Z.S. Zheng, A. Coogan, R. Rice, R. Akella, M.H. Conklin, Gap-filling snow-depth time-series with kalman filtering-smoothing and expectation maximization: proof of concept using spatially dense wireless-sensor-network data, Cold Reg. Sci. Tech. 175 (2020) 12, https://doi.org/10.1016/j.coldregions.2020.103066.

[23] T. Yu, J.C. Wang, A spatiotemporal convolutional gated recurrent unit network for mean wave period field forecasting, J. Mar. Sci. Eng. 9 (2021) 17, https://doi.org/10.3390/jmse9040383.

[24] P.A. Umesh, M.R. Behera, On the improvements in nearshore wave height predictions using nested SWAN-SWASH modelling in the eastern coastal waters of India, Ocean Eng. 236 (2021) 18, https://doi.org/10.1016/j.oceaneng.2021.109550.

[25] M.B. Soran, K. Amarouche, A. Akpinar, Spatial calibration of WAVEWATCH III model against satellite observations using different input and dissipation parameterizations in the Black Sea, Ocean Eng. 257 (2022) 19, https://doi.org/10.1016/j.oceaneng.2022.111627.

[26] Z.L. Ti, M.J. Zhang, L.H. Wu, S.Q. Qin, K. Wei, Y.L. Li, Estimation of the significant wave height in the nearshore using prediction equations based on the response surface method, Ocean Eng. 153 (2018) 143–153, https://doi.org/10.1016/j.oceaneng.2018.01.081.

[27] G. Zheng, X.F. Li, R.H. Zhang, B. Liu, Purely satellite data-driven deep learning forecast of complicated tropical instability waves, Sci. Adv. 6 (2020) 9, https://doi.org/10.1126/sciadv.aba1482.

[28] F. Vieira, G. Cavalcante, E. Campos, F. Taveira-Pinto, A methodology for data gap filling in wave records using artificial neural networks, Appl. Ocean Res. 98 (2020) 9, https://doi.org/10.1016/j.apor.2020.102109.

[29] S.T. Chen, Y.W. Wang, Improving coastal ocean wave height forecasting during typhoons by using local meteorological and neighboring wave data in support vector regression models, J. Mar. Sci. Eng. 8 (2020) 15, https://doi.org/10.3390/jmse8030149.

[30] U. Mital, D. Dwivedi, J.B. Brown, B. Faybishenko, S.L. Painter, C.I. Steefel, Sequential imputation of missing spatio-temporal precipitation data using random forests, Front. Water. 2 (2022) 20, https://doi.org/10.3389/frwa.2020.00020.

[31] J.C. Wang, K.H. Wen, F.Y. Deng, Filling gaps in significant wave height time series records using bidirectional gated recurrent unit and cressman analysis, Dyn. Atmos. Oceans. 101 (2023) 12, https://doi.org/10.1016/j.dynatmoce.2022.101339.

[32] K. Ustoorikar, M.C. Deo, Filling up gaps in wave data with genetic programming, Mar. Struct. 21 (2008) 177–195, https://doi.org/10.1016/j.marstruc.2007.12.001.

[33] S. Aziz, M.U. Khan, M. Faraz, G.A. Montes, Intelligent bearing faults diagnosis featuring automated relative energy based empirical mode decomposition and novel cepstral autoregressive features, Measurement. 216 (2023), https://doi.org/10.1016/j.measurement.2023.112871.

[34] C.G. Soares, A.M. Ferreira, Representation of non-stationary time series of significant wave height with autoregressive models, Probab. Eng. Eng. Mech. 11 (1996) 139–148, https://doi.org/10.1016/0266-8920(96)00004-5.

[35] O. Ferreiro, Methodologies for the estimation of missing observations in time series, Stat. Probab. Lett. 5 (1987) 65–69, https://doi.org/10.1016/0167-7152(87)90028-9.

[36] J.Z. Yim, C.-R. Chou, P.-C. Ho, Study on simulating the time series of significant wave heights near the keelung harbor, ISOPE-I-02-278 (2002) 92–96.

[37] S.T. Fan, N.H. Xiao, S. Dong, A novel model to predict significant wave height based on long short-term memory network, Ocean Eng. 205 (2020) 13, https://doi.org/10.1016/j.oceaneng.2020.107298.

[38] A. Torfi, R.A. Shirvani, Y. Keneshloo, N. Tavaf, E.A.J.a.p.a. Fox, Natural language processing advancements by deep learning: a survey. arXiv. (2020) 1-23. https://doi.org/10.48550/arXiv.2003.01200.

[39] J. Yao, W.H. Wu, Wave height forecast method with multi-step training set extension LSTM neural network, Ocean Eng. 263 (2022) 11, https://doi.org/10.1016/j.oceaneng.2022.112432.

[40] C. Jörges, C. Berkenbrink, B. Stumpe, Prediction and reconstruction of ocean wave heights based on bathymetric data using LSTM neural networks, Ocean Eng. 232 (2021) 18, https://doi.org/10.1016/j.oceaneng.2021.109046.

[41] M. Ma, L. Fu, Z. Zhai, R.-B. Sun, Transformer based Kalman Filter with EM algorithm for time series prediction and anomaly detection of complex systems, Measurement. 229 (2024), https://doi.org/10.1016/j.measurement.2024.114378.

[42] A.S. Lee, W. Hilal, S.A. Gadsden, M. Al-Shabi, Combined Kalman and sliding innovation filtering: An adaptive estimation strategy, Measurement. 218 (2023), https://doi.org/10.1016/j.measurement.2023.113228.

[43] H.Z. Fang, M.A. Haile, Y.B. Wang, Robust extended Kalman filtering for systems with measurement outliers, IEEE Trans. Control Syst. Technol. 30 (2022) 795–802, https://doi.org/10.1109/tcst.2021.3077535.

[44] R. Dehghannasiri, X.N. Qian, E.R. Dougherty, A Bayesian robust Kalman smoothing framework for state-space models with uncertain noise statistics, EURASIP J. Adv. Signal Process. (2018) 17, https://doi.org/10.1186/s13634-018-0577-1.

[45] A.Y. Aravkin, J.V. Burke, B.M. Bell, G. Pillonetto, Algorithms for block tridiagonal systems: stability results for generalized Kalman smoothing. IFAC-PapersOnLine. Padova, ITALY, 2021, pp. 821-826. https://doi.org/10.1016/j.ifacol.2021.08.463.

[46] G. Jain, B.J.A.a.S. Mallick, A study of time series models ARIMA and ETS. (2017). Available at SSRN 2898968. http://doi.org/10.2139/ssrn.2898968.

[47] S.H. Bari, M.T. Rahman, M.A.A. Hussain, S.J.C. Ray, Forecasting monthly precipitation in Sylhet city using ARIMA model, Civ. Environ. Res. 7 (2015) 69–77.

[48] A.A. Ariyo, A.O. Adewumi, C.K. Ayo, Stock price prediction using the ARIMA model. UKSim-AMSS 16th Int, Conf. Comput. Model. Simul. (2014) 106–112, https://doi.org/10.1109/UKSim.2014.67.

[49] C.W.J. Granger, Some properties of time series data and their use in econometric model specification, J. Econom. 16 (1981) 121–130, https://doi.org/10.1016/0304-4076(81)90079-8.

[50] A. Meyler, G. Kenny, T. Quinn, Forecasting Irish inflation using ARIMA models, Report 11359 (1998) 1–49.

[51] S. Oh, K. Jang, J. Kim, I. Moon, Online state of charge estimation of lithium-ion battery using surrogate model based on electrochemical model. in: L. Montastruc, S. Negny (Eds.), Comput. Aided Chem. Eng. Elsevier, 2022, pp. 1447-1452. https:/doi.org/10.1016/B978-0-323-95879-0.50242-3.

[52] H.L. Xie, L. Zhang, C.P. Lim, Evolving CNN-LSTM models for time series prediction using enhanced grey wolf optimizer, IEEE Access. 8 (2020) 161519–161541, https://doi.org/10.1109/access.2020.3021527.

[53] C.J. Willmott, K. Matsuura, Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance, J. Clim. Res. 30 (1) (2005) 79–82, https://doi.org/10.3354/cr030079.

[54] R.J. Hyndman, A.B. Koehler, Another look at measures of forecast accuracy, Int. J. Forecast. 22 (4) (2006) 679–688, https://doi.org/10.1016/j.ijforecast.2006.03.001.

[55] S. Oh, K. Jang, J. Kim, I. Moon, Another look at measures of forecast accuracy, Int. J. Forecast. 22 (2006) 679–688, https://doi.org/10.1016/j.ijforecast.2006.03.001.

[56] M. Wang, F.J.A.O.R. Ying, A hybrid model for multistep-ahead significant wave height prediction using an innovative decomposition-reconstruction framework and E-GRU, Appl. Ocean Res. 140 (2023), https://doi.org/10.1016/j.apor.2023.103752.