Hindawi
Complexity
Volume 2021, Article ID 2727218, 14 pages
https://doi.org/10.1155/2021/2727218

WILEY | Hindawi

Research Article

Decomposition-Based Multistep Sea Wind Speed Forecasting Using Stacked Gated Recurrent Unit Improved by Residual Connections

Jupeng Xie, Huajun Zhang, Linfan Liu, Mengchuan Li, and Yixin Su

School of Automation, Wuhan University of Technology, Wuhan 430000, China

Correspondence should be addressed to Huajun Zhang; zhanghj@whut.edu.cn

Received 17 May 2021; Revised 3 October 2021; Accepted 29 October 2021; Published 15 November 2021

Academic Editor: Nishant Malik

Sea wind speed forecast is important for meteorological navigation system to keep ships in safe areas. The high volatility and uncertainty of wind make it difficult to accurately forecast multistep wind speed. This paper proposes a new decomposition-based model to forecast hourly sea wind speeds. Because mode mixing affects the accuracy of the empirical mode decomposition-(EMD-) based models, this model uses the variational mode decomposition (VMD) to alleviate this problem. To improve the accuracy of predicting subseries with high nonlinearity, this model uses stacked gate recurrent units (GRU) networks. To alleviate the degradation effect of stacked GRU, this model modifies them by adding residual connections to the deep layers. This model decomposes the nonlinear wind speed data into four subseries with different frequencies adaptively. Each stacked GRU predictor has four layers and the residual connections are added to the last two layers. The predictors have 24 inputs and 3 outputs, and the forecast is an ensemble of five predictors' outputs. The proposed model can predict wind speed in the next 3 hours according to the past 24 hours' wind speed data. The experiment results on three different sea areas show that the performance of this model surpasses those of a state-of-the-art model, several benchmarks, and decomposition-based models.

1. Introduction

Sea wind always threatens the safe navigation of ships. According to the Marine Casualties and Incidents Reports published by the International Maritime Organization (IMO), there were 1561 well-documented ship accidents in the first 10 years of the 21st century, of which 755 were caused by strong winds and large waves caused by strong winds, and accidents caused by strong winds accounted for 48.3% of total accidents. In addition, when the wind wave and swell appear at the same time, the danger of navigation will be greatly increased [1]. Therefore, the accurate wind speed forecasting is of great significance for routes optimization and navigation risk management.

Wind speed forecasts are divided into 4 categories, super-short-term [2, 3], short-term [4–7], medium-term [8, 9], and long-term [10], ranging from a few seconds to 30 minutes, from 30 minutes to 6 hours, from 6 hours to 24 hours, and from 24 hours to a week or more, respectively [11]. According to the principle of the wind speed forecasts models, they are classified into physical models and statistical models. The second class includes time series models and machine learning models [11, 12]. The physical methods, like the numerical weather prediction (NWP), construct differential equations about physical factors such as wind speed, wind direction, air temperature, and pressure. Solving meteorological equations requires a large number of computing resources and time. The result belongs to long-term forecasting of a large area. The time series methods model the relationship between current wind speeds and historical wind speeds, which are suitable for short-term and medium-term forecast. Most time series methods, such as the autoregressive integrated moving average (ARIMA) [2, 13] and autoregressive moving average with exogenous variables (ARMAX) [14], assume that there is a linear relationship between current data and past data or errors. The

construction, order identification of these models is easy to understand, but their linear assumptions lead to poor forecast performance on nonlinear data. Machine learning methods are suitable for short-term or super-short-term forecasting. They take each time point of past series as an input feature and that of predicted series as an output feature and construct nonlinear relationship. Many complex machine learning models, such as the long short-term memory (LSTM) [15] and gated recurrent unit (GRU) [4, 16], are able to learn temporal correlation and often outperform time series models. Among them, GRU not only alleviates the risk of gradient explosion and vanishing but also is faster than LSTM.

Decomposition-based methods have attracted much attention recently. These methods decompose the original wind speed into several subseries and use a group of same or different individual prediction models to learn each subseries [17]. Usually, time series or machine learning models are selected as predictor. The decomposition-based methods reduce the complexity of original data and make the predictors easier to learn. In addition, multimodel ensemble decreases the risk of getting stuck in local optima in the training process [18]. Therefore, decomposition-based forecasting is more accurate than direct forecasting via individual model. In the field of wind speed forecasting, wavelet transform (WT) methods and empirical mode decomposition (EMD) methods are the most used algorithms [19]. Usually, the repeated WT and ARIMA are used to predict super-short-term wind speed of a 10 min scale [20]. Because LSTM is more effective than ARIMA in nonlinear system, the WT and LSTM are combined to predict hourly wind speed, and feature selection based on mutual information is executed between decomposition and predictors [6]. The characteristics of linear and nonlinear are different; then the ARIMA and multilayer perceptron (MLP) are used to predict linear and nonlinear subseries which are classified based on the EMD [21]. There are also some decomposition-based methods that are used to deal with nonlinearity. Moving average (MA) filter [22] and ARIMA filter [23] are used to separate linear components, and MLP is used to predict the nonlinear parts [24]. Besides the linear and nonlinear predictors, a predictor is used to predict the residual of EMD, since it includes some information [25]. For short-term forecast, a permutation entropy (PE) method is used to predict a 3-step hourly forecast. The subseries is reorganized into several new series according to their PE values. Because it is difficult to capture the nonlinear features, this method uses MLP to predict each component [26].

According to above references, the decomposition-based methods have many advantages. However, they have some problems that have not been widely solved. Firstly, the decomposition algorithms have some defects. Although the WT and EMD are used in wind speed forecast, there are some defects that decrease the forecast accuracy. It is difficult to use WT to analyse local low-frequency changes [27] and the decomposition behaviour depends on wavelet basis functions [15]; different wavelet basis functions bring different decomposition results. The EMD decomposes a time series into subseries with different frequency domain bandwidths and the frequency bands have no overlap ideally. When there is a frequency band overlap in the subseries, multiple modes are mixed, and it is not suitable for further processing [28]. In [29], the ensemble empirical mode decomposition was used to predict short-term wind speed. In [30], the complementary ensemble empirical mode decomposition was used to alleviate mode mixing. These two methods add multiple white noises to the original data and then integrate the results of multiple EMDs. Variational mode decomposition (VMD) is proposed to solve the mode mixing and does not depend on fixed basis. Different from the EMD-based algorithms, it avoids mode mixing as much as possible by solving specific intrinsic mode functions (IMFs) [31]. In recent studies, the VMD, MLP, and autoregressive moving average (ARMA) are used to predict wind speed with 10-minute interval [32] and 30-minute interval [19].

Secondly, the subseries' predictors can be improved by stacking. Although the decomposed subseries are simplified in frequency, they still have relatively high nonlinearity. Many prediction studies used the support vector machine regression (SVR) and MLP as nonlinear predictors [21, 33]. The neural networks are good at nonlinear modelling, so complex neural networks, such as LSTM [6, 34] and GRU [4], are helpful to improve the accuracy of forecast. A hybrid predictor that includes the VMD and a single-layer GRU is used to predict the wind power interval [3]. Ideally, stacking more models will significantly improve the ability of nonlinear modelling. However, the actual performance of a stacked network often becomes worse when there are more layers. It is difficult to train deeper layers to fit an identity mapping and it leads to the degradation of stacked models. The residual connections solve this degradation phenomenon by building linear paths between deep layers [35]. The stacked LSTM with residual connections shows superior accuracy in machine translation and sentiment intensity prediction [36, 37], but this improvement has not been applied to wind speed forecasting. In [36], two 8-layer LSTMs are added with residual connections every 2 layers. In [37], an 8-layer LSTM is added with residual connections every 1 and 2 layers, respectively, and two types have their own advantages. In wind speed forecasting field, these two types remain to be verified by experiments.

This paper proposes a VMD-Stacked GRU model with residual connections to forecast the short-term global sea wind speed with multiple steps. The decomposition and predictor are designed based on analysis and experiments. Original wind speed data is complex and the VMD is used to decompose the wind speed data; it makes an adaptive decomposition that overcomes the defect caused by mode mixing in EMD-based models. A modified GRU is used as subseries predictor to improve its nonlinear modelling ability. The performances of the stacked GRU are improved by adding the residual connections between the last two layers. This improvement by adding residual connection is very novel in the field of wind speed forecasting. In addition, a lot of experiments are carried out on the European Reanalysis (ERA5) dataset. It has been proved to surpass the

performances of several benchmark and baseline models. Different from most studies based on wind farm observation data, it supports the study of sea wind speeds. The experiment results show that the performance of the proposed model is better than those of some benchmark and baseline models.

This paper is organized as follows. Section 2 describes three methods involved in the proposed model. Section 3 details the proposed model's architecture and evaluation criteria. Section 4 details the experiments and analysis that are used to obtain the best forecast performances. Section 5 provides discussion and Section 6 summarises the conclusions. An acronyms list is shown in Table 1.

2. Methodology

2.1. Variational Mode Decomposition. Wind speed series are nonlinear nonstationary signals which contain a variety of period characteristics. For example, the Fourier transform for hourly wind speed shows that it is not a 24-hour period, but there are many significant periods. It means that multiperiod wind speed cannot be represented by an instantaneous frequency. When forecasting wind speed directly, the complex periodicity will be disadvantageous to model learning. To understand the signals with complex periodic patterns, an effective strategy is to use IMFs, which are ideal functions with fixed instantaneous frequency. Since there is no complex periodicity, it is relatively easy to predict the IMFs.

To extract IMFs from the original series, the EMD adopts a completely different iteration method to deal with the original data adaptively [27]. But, in practice, there are some imperfections such as overshoots, undershoots, asymmetric wave forms, and ends swing in the results of EMD, which make them not the ideal IMFs [38]. In order to alleviate the above problems, the VMD is proposed to calculate the IMFs more accurately. By constructing and solving a constrained variational problem, the VMD obtains all modal components nonrecursively and improves the decomposition robustness to noise. Under the constraint that the summation over all modes is equal to the original signal, the sum of the all estimated bandwidths of modes is minimized, and the following optimization problem is constructed [31]:

$$\begin{cases} \min_{\{u_k\},\{w_k\}} & \sum_{k=1}^{K} D_t\left(u_k, w_k\right) s.t. \sum_{k=1}^{K} u_k = f, \end{cases} \tag{1}$$

where u_k and w_k are the k-th modal component and the center frequency after decomposition, respectively. $u_k = A_k(t)\cos(\varphi_k(t))$, $w_k(t) = \varphi_k'(t)$, $k \in \{0, 1, 2, \ldots, K\}$. f is the original time series, and K is its mode decomposition number. D_t is the estimated bandwidth of each modal component:

$$D_t = \partial_t \left[\left(\delta(t) + \frac{j}{\pi t}\right) * u_k(t)\right] \exp\left(-jw_k t\right)_2^2, \tag{2}$$

where $\frac{2}{2}$ is the squared L^2 -norm of the gradient and $*$ represents the convolution operation. ∂_t is the partial derivative operation, and $\delta(t)$ is the Dirac distribution.

By using quadratic penalty factor α and Lagrangian multipliers λ, the lowest point of this variational constraint problem is transformed into saddle point of augmented Lagrange equation defined as follows. The augmented Lagrange equation is shown as follows [31]. The equation can be iteratively calculated by the Alternating Direction Multiplier Algorithm.

$$\begin{aligned} L\left(\{u_k\}, \{w_k\}, \lambda(t)\right) = \alpha \sum_k \left\| \partial_t \left[\left(\delta(t) + j/\pi t\right) * u_k(t) \right] \exp\left(-jw_k t\right) \right\|_2^2 + \left\| f(t) - \sum_k u_k(t) \right\|_2^2 + \\ \left\langle \lambda(t), f(t) - \textstyle\sum_k u_k(t) \right\rangle. \end{aligned} \tag{3}$$

2.2. Stacked Gate Recurrent Unit. The conventional machine learning methods deal with time series problem; each moment of a sample is regarded as a different independent random variable, and it is given into the regression model or neural network for training. However, these models assume that the data at different moments are independent of each other, and their sequence in time is not considered. The recurrent neural network (RNN) is proposed to capture this temporal correlation by using the machine learning. The GRU is a modified RNN based on the LSTM. When error signals propagate backwards through time in the conventional RNN, the signals tend to vanish or blow up, and both of the cases lead to the failure of the network to learn from data [28]. The GRU not only retains the ability to prevent the previously mentioned problems but also reduces the complexity of the structure without losing the efficient learning ability [39].

The structure of the GRU at each step is the GRU cell, which is shown in Figure 1. In this figure, the reset gate and the update gate are fully connected layers with sigmoid activation, which are used to control the memory. The previous hidden state preserves the past memory, the reset gate controls how to combine the input with the past memory to become a candidate hidden state, and the update gate controls how to add the candidate hidden state into the hidden state [39]. Finally, the candidate hidden state, previous hidden state, and output of the update gate constitute the current hidden state and output. The GRU cell can be expressed as follows:

Table 1: Acronyms used in the article.

Acronyms	
ARIMA	Autoregressive integrated moving average model
ARMA	Autoregressive moving average
ARMAX	Autoregressive moving average with exogenous variables
EMD	Empirical mode decomposition
ERA5	The European Reanalysis dataset
GRU	Gated recurrent unit
IMFs	Intrinsic mode functions
LSTM	Long short-term memory
MA	Moving average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MLP	Multilayer perceptron
NWP	Numerical weather prediction
PE	Permutation entropy
RMSE	Root Mean Square Error
RNN	Recurrent neural network
SVR	Support vector machine regression
VMD	Variational mode decomposition
WT	Wavelet transform

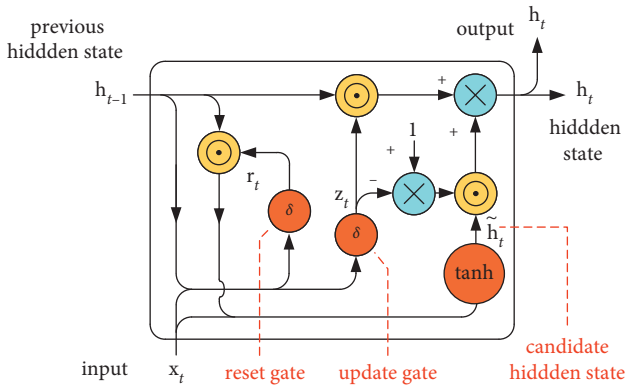


Figure 1: Cell of gated recurrent unit.

$$\begin{aligned}
\mathbf{z}_t &= \sigma\left(\mathbf{W}^z\mathbf{x}_t + \mathbf{U}^z\mathbf{h}_{t-1} + \mathbf{b}^z\right),\\
\mathbf{r}_t &= \sigma\left(\mathbf{W}^r\mathbf{x}_t + \mathbf{U}^r\mathbf{h}_{t-1} + \mathbf{b}^r\right),\\
\tilde{\mathbf{h}}_t &= \tanh\left(\mathbf{W}\mathbf{x}_t + \mathbf{U}\left(\mathbf{r}_t \odot \mathbf{h}_{t-1}\right) + \mathbf{b}\right),\\
\mathbf{h}_t &= \left(1 - \mathbf{z}_t\right) \odot \tilde{\mathbf{h}}_t + \mathbf{z}_t \odot \mathbf{h}_{t-1},
\end{aligned} \tag{4}$$

where $\mathbf{h}_{t-1}$ is the hidden state at $t-1$ and $\mathbf{x}_t$, $\mathbf{z}_t$, $\mathbf{r}_t$, $\tilde{\mathbf{h}}_t$, and $\mathbf{h}_t$ are the input of the GRU cell, output of the update gate, output of the reset gate, candidate hidden state, and hidden state at t, respectively. $\mathbf{W}$ and $\mathbf{U}$ are the weight matrices of the fully connected layer, and $\mathbf{b}$ is the bias vector. σ and tanh are sigmoid and tanh activation function, respectively. $\odot$ represents the element-wise product between two matrices of the same size. The GRU cell is shown in Figure 1.

To make the GRU work, its current hidden state is connected to the next hidden state input. In order to improve the learning ability, multiple GRU cells can be stacked along the input-output direction, and the output of the GRU cell at each step can be used as the input of the next GRU cell at corresponding step. Compared with single-layer GRU, stacked GRU has multiple hidden layers, which can improve the ability to learn time series. The structure of stacked GRU along the time axis is shown in Figure 2.

2.3. Residual Connections. With the appearance of normalization and dropout, the vanishing and exploding gradients problem of the stacked neural network is greatly alleviated, which makes the training of deep network no longer difficult. In theory, the learning ability of the stacked neural network increases with the number of layers, and its error also decreases until it remains unchanged. But actually, when the number of layers increases, the network's performance will degrade rapidly. At present, stacked RNN, LSTM, or GRU generally has no more than four recurrent layers [36].

The latest research pointed out that overfitting is not the cause of stack network degradation. The assumption that the performance of a deep network is not lower than that of a shallow network is based on the ability of the deep part of the network identity mapping its input, in other words, the ability of the deep part of the network fitting $f(x) = x$ [35]. However, artificial neural network has been proved to be difficult to apply in learning linear relationship [33]. In order to give the network layer this ability, the residual connections as shown in Figure 3 are proposed [35].

The red part in Figure 3 is the added residual connections, also known as skip connections or shortcut connections. After adding residual connections, the input of the network layer is directly superimposed with the output, and the layer is transformed from fitting $f(x) = x$ to fitting $f(x) = h(x) - x$. $h(x)$ is the approximate identity mapping of x; the network layer is changed to learning the nonlinear residual of identity mapping. It is much easier for neural network to learn a group of nonlinear data close to zero compared to linear data.

The residual connections shown in Figure 3 were first used to solve the degradation problem of the deep convolutional neural network in image recognition, but they can also be applied to any stacked network. Figure 4 shows the stacked GRU structure with residual connections, which is the same as Google's stacked LSTM in its machine translation model [36]. Different from Figure 3, the residual connections in Figure 4 skip one GRU layer instead of two, and Add is set before the activation function. The GRU network layer inputs the second GRU layer after the element-wise addition of the output and input at each step. Each GRU layer with residual connections constitutes a residual block, which can be defined as follows:

$$\begin{cases}
\mathbf{h}_t^i = \mathrm{GRU}^i\left(\mathbf{x}_t^{i-1}, \mathbf{h}_{t-1}^i; \mathbf{W}^i, \mathbf{U}^i\right)\\
\mathbf{x}_t^i = \mathbf{h}_t^i + \mathbf{x}_t^{i-1}
\end{cases}, \tag{5}$$

where the function is composed of equations (6)–(8), representing the i -th GRU layer.

The residual connections can significantly improve the flow of gradients between network layers. In theory, the network with any number of layers can be trained after

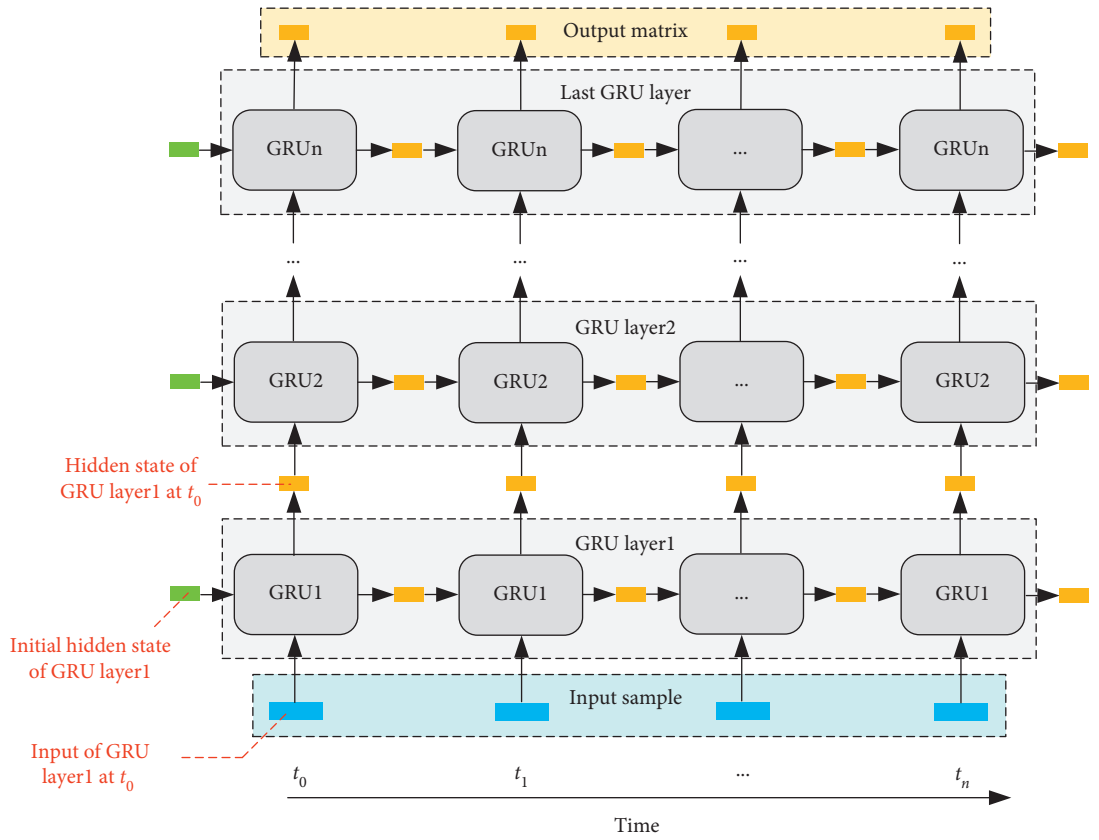


Figure 2: Structure of stacked GRU along the time axis.

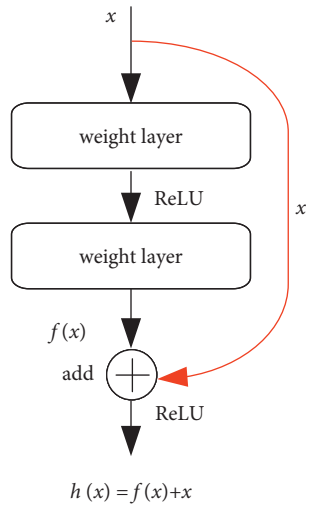


Figure 3: Structure of the residual connections.

stacking residual blocks. But, in practical works, the sum of LSTM layers with and without residual network is no more than 8 [36].

3. The Proposed Model

3.1. Model Architecture. The proposed model architecture is shown in Figure 5. It contains three parts: data split, data decomposition, and components prediction. The process of the proposed model is summarized as follows:

(1) The data split part splits the original wind speed series into three subsets: training set, validation set, and test set. The train-validation-test split percentage is 60%-20%-20%. The test set is considered unknown and does not participate in the training process; and the validation set is used to determinate

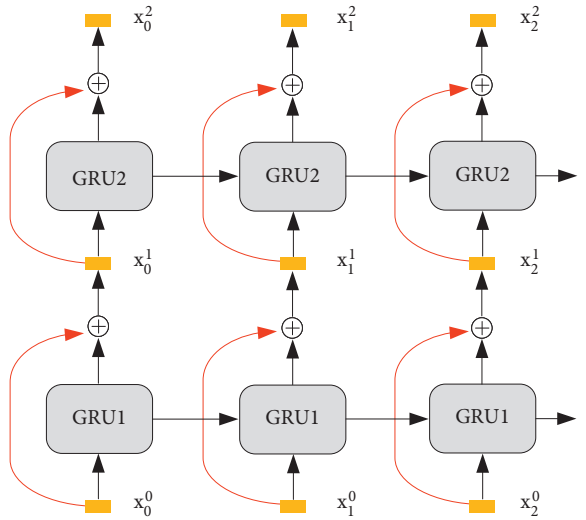


Figure 4: Structures of stacked GRU with residual connections.

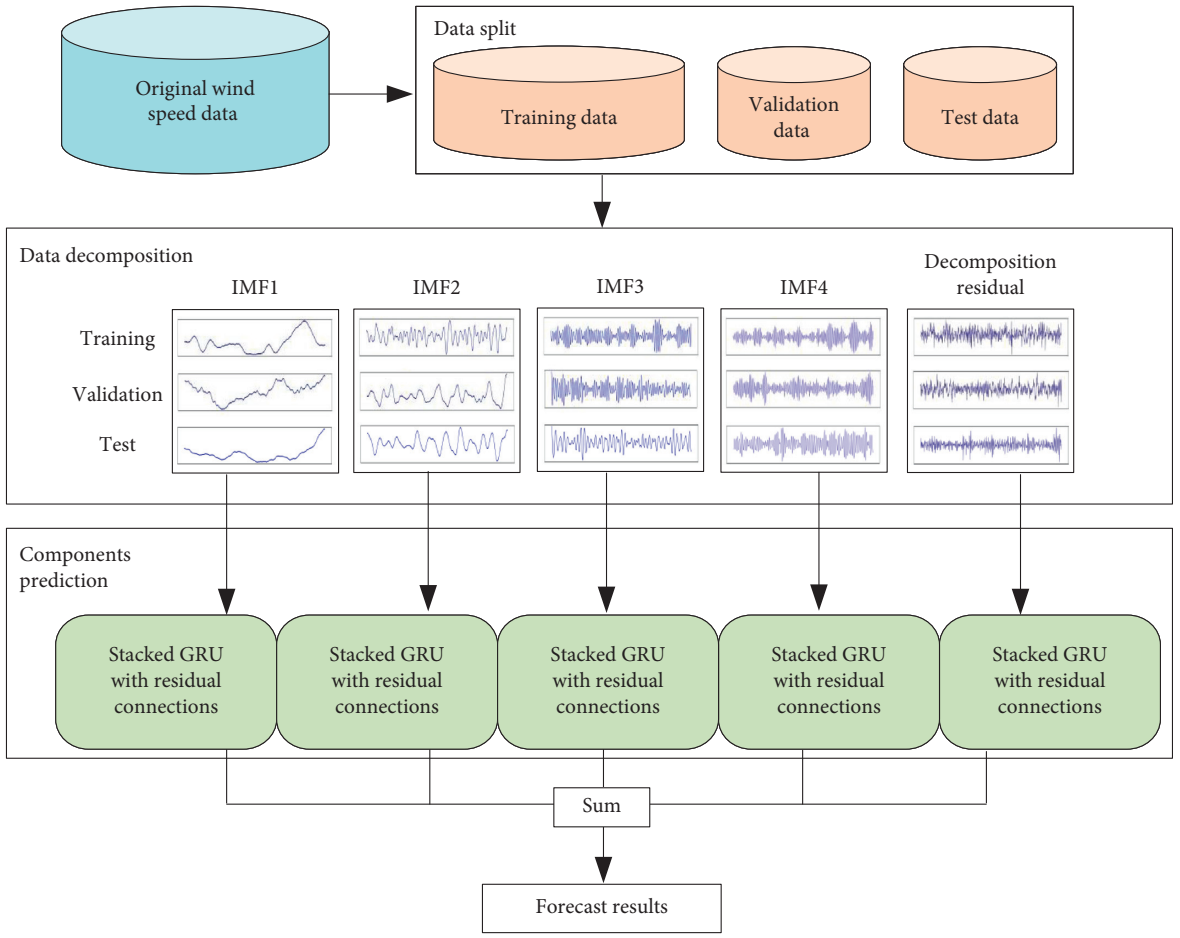


Figure 5: Architecture of the proposed model.

hyperparameters. In order to speed up model training, we also normalize the three subsets to eliminate the range differences and accelerate the gradient descent. The maximum and minimum values of the training set are obtained to scale the validation set and test set.

(2) The data decomposition part uses VMD to solve the constrained variational problem and then

reconstruct the component series and calculates the decomposition residual. Correlated information remains in the decomposition residual, so it is necessary to set up a predictor for the decomposition residual. The number of subseries determines the number of predictors and total training time. In order to make a trade-off between the training time and forecasting accuracy, the wind speed series are decomposed into four subseries in this part under termination conditions = 1e-7.

(3) In the components prediction part, since the subseries have different frequency characteristics, five stacked GRU models with residual connections are used to predict the subseries and a decomposition residual, respectively. The final forecast values are obtained by integrating the forecast outputs of all predictors. Since this paper is a short-term hourly forecasting; the data from the past 24 hours are highly related to the forecast values. Therefore, the data from 24 hours are used to make a 3-hours-ahead wind speed forecasting.

According to the above section, the residual connections should be set in the deep layer of the network, so we design a stacked GRU with four layers, and GRU layer 1 and GRU layer 2 are independent, while the residual connections are set at the input of GRU layer 3 and the output of GRU layer 4. The output of GRU layer 2 will be fed to the output of GRU layer 3 and added to it, and then the sum of them will be fed to the output of GRU layer 4 and added to it. In order to match the output of the stacked GRU with the desired output step size, we use a flatten layer to reshape the output into a one-dimensional vector, and then a dense layer with linear activation function is used for linear conversion. A detailed parameters determination is described in the *Parametric Study* section.

3.2. Evaluation Criteria. Time series forecasting can be converted into a supervised regression problem, so we use three regression metrics to evaluate the forecasting performance. These regression metrics are the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). They are defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - f\left(x_i\right)\right)^2}, \tag{6}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - f\left(x_i\right)\right|, \tag{7}$$

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - f\left(x_i\right)}{y_i}\right|, \tag{8}$$

where x_i represents the model inputs, $f(x_i)$ is the forecast value, y_i is the corresponding actual value, and n is the number of actual values.

RMSE and MAE have the same physical dimensions as the original data and range from 0 to ∞. A lower RMSE or MAE means the model has a higher forecasting accuracy. MAE uses the absolute value to describe the gap between two curves, and the error of each prediction point has the same weight in the final error. Therefore, the MAE is less than the RMSE on the same data. Since the square term of the RMSE magnifies the error between two points, the large gap between the RMSE and MAE can indicate that some prediction points contribute significantly to the final error of the prediction curve. The MAPE is a dimensionless metric ranging from 0 to ∞, and a lower MAPE means the model has a higher forecasting accuracy. We use this metric because it considers the proportion of error in the total data and is able to evaluate performance of different models on the same dataset. In addition, 5-fold cross-validation strategy is used in the Result of Multistep Wind Speed Forecasting section.

4. Case Study

4.1. Datasets. Marine meteorological datasets are collected, sorted, and released by scientific research institutions in various countries. The use of the datasets varies greatly depending on the observation method, observation area, observation period, and observation elements. Selecting a high-quality, long-term, and high-resolution marine meteorological dataset is the premise of modelling. Reanalysis datasets are produced from the buoy and satellite observation data by determining the optimal estimation of the system state. The reanalysis datasets can be regarded as the real global ocean data and are currently used as the data source for the NWP.

Therefore, the ERA5 is selected in the case study. The ERA5 is the latest global meteorological dataset released by the European Centre for Medium-Range Weather Forecasts (ECMWF). The ERA5 provides hourly wind speed data in 137 levels from the surface up to a height of 80 km, covering the global land and ocean with 30 km grids [40]. The wind speeds are decomposed into u-component and v-component. The positive u-component of wind is eastward wind speed, and the negative counterpart is westward wind speed. The positive v-component of wind is northward wind speed, and the negative counterpart is southward wind speed.

In order to verify the applicability of the proposed model in global ocean, we use hourly wind speed u-component from 1 January 2016, 00 : 00, to 31 December 2017, 23 : 00, which includes 17544 hours to make forecast experiments. The forecast areas are located in the Pacific, Indian, and Atlantic Oceans, respectively. Figure 6 shows the coordinates of three forecast areas and their surrounding areas on the map, as well as the heat map of wind speed in January 2016. The map in Figure 6 is drawn based on the National Oceanic and Atmospheric Administration (NOAA) Panoply software. The statistical indices of three areas in two years are shown in Table 2.

To further analyse the characteristics of the dataset, Figure 7 shows the original time series, as well as [0, 1] normalized trends and seasonality of the first month with 24

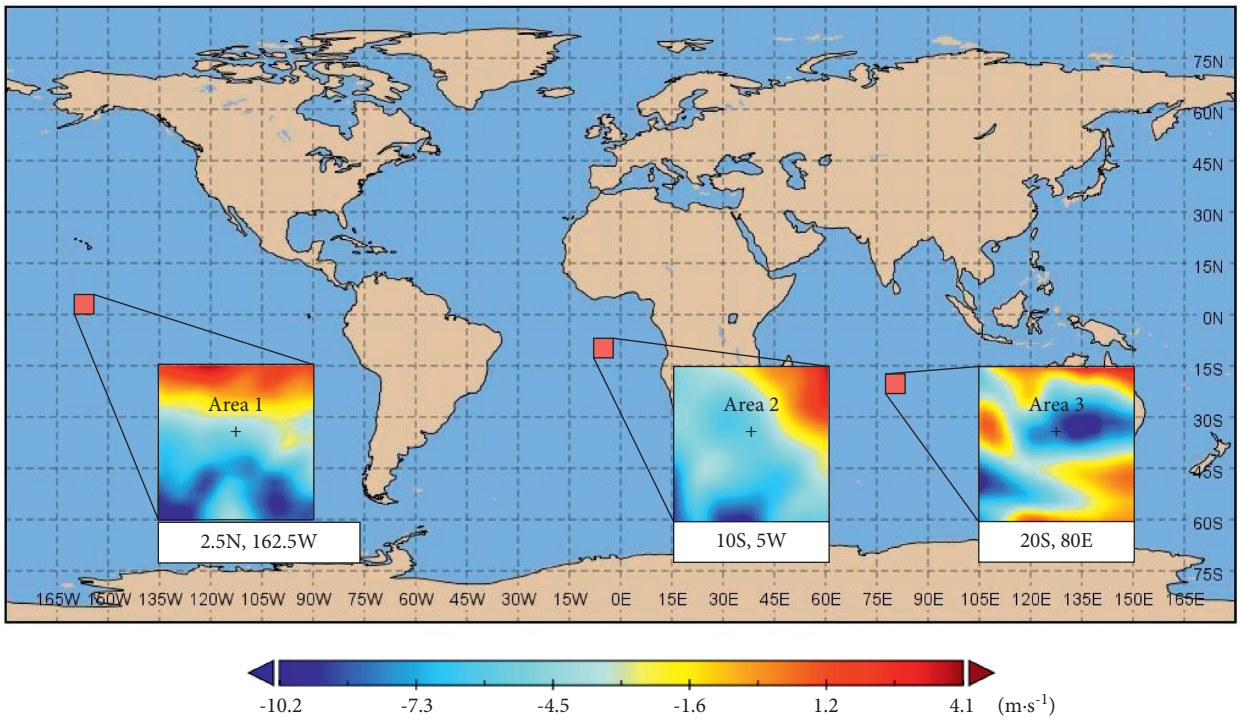


Figure 6: Coordinates of three areas and their surroundings on the map.

Table 2: Statistical indices of three sea areas.

Wind speed ($m \cdot s^{-1}$)	Area 1	Area 2	Area 3
Maximum eastward speed	13.5472	8.9360	1.0977
Maximum westward speed	11.3178	24.1463	9.3708
Minimum absolute speed	0.0093	0.0269	0.1613
Average speed	−5.7063	−7.4246	−5.1017
Standard deviation	1.9026	3.0145	1.3896
Average speed of trendy series	−5.6236	−7.1146	−4.8573
Standard deviation of trendy series	1.6162	2.8849	1.2616
Maximum speed of seasonal series	9.3964	12.7637	7.7910
Standard deviation of seasonal series	0.4388	0.4328	0.3293

hours as period. The trendy series present great changes in a month. The seasonal series do not show repetitive patterns, which indicates that the wind speed data contain multiple seasonalities. Therefore, it is necessary to use decomposition methods that support multiple frequencies.

In Table 2, the positive average wind speed means eastward wind speed, and the negative means westward wind speed. It can be seen from Table 2 that the maximum westward wind speeds in areas 2 and 3 are significantly higher than the maximum eastward wind speed, while they are similar in area 1. Area 2 has the highest average wind speed and the largest standard deviation. In addition, although the maximum wind speed in area 3 is just $9.3708\,m \cdot s^{-1}$, its minimum value is $0.1613\,m \cdot s^{-1}$, much higher than other areas. According to the statistical indices of trends and seasonality, area 1 and area 2 have similar seasonal standard deviations, and the higher volatility of area 2 is attributed to its trendy part. Area 3 has the lowest trendy and seasonal standard deviations among the three areas.

4.2. Comparison between Decomposition-Based Models. In order to prove that the VMD is superior to other decomposition methods and that the performance of the stacked GRU is improved by the residual connections as a component predictor, the following experiments are carried out. All of data are normalized when passed to the model for training.

First, a group of experiments are carried out in area 1 to prove that the proposed combination of the VMD and the stacked GRU is superior to the combinations of the other decomposition and prediction models. In the experiment, three decomposition methods and four prediction models are cross-combined. The WT, EMD, and VMD are selected. Among them, the decomposition level of the WT is 4, which means that the wind speed sequence will be decomposed into an approximate component, four detail components, and a decomposition residual sequence. The EMD processes the sequence adaptively, so the wind data in area 1 is decomposed into 9 to 11 subseries. Therefore, the highest frequent subseries will be discarded until the number of all subseries does not exceed 9. The VMD decomposes wind data into four subseries under termination conditions = 1e-7.

The prediction methods include the LSTM, GRU, stacked LSTM, and stacked GRU. The LSTM and GRU are designed as a single-layer structure with 512 neural units. The stacked LSTM and stacked GRU are designed as four layers with 512, 32, 32, and 32 neural units in each layer, respectively. The batch size is 25 and Adam optimizer's learning rate is 0.001. The results are shown in Table 3.

The three metrics of the EMD are slightly lower than those of WT. There are some exceptions in Figure 3. For

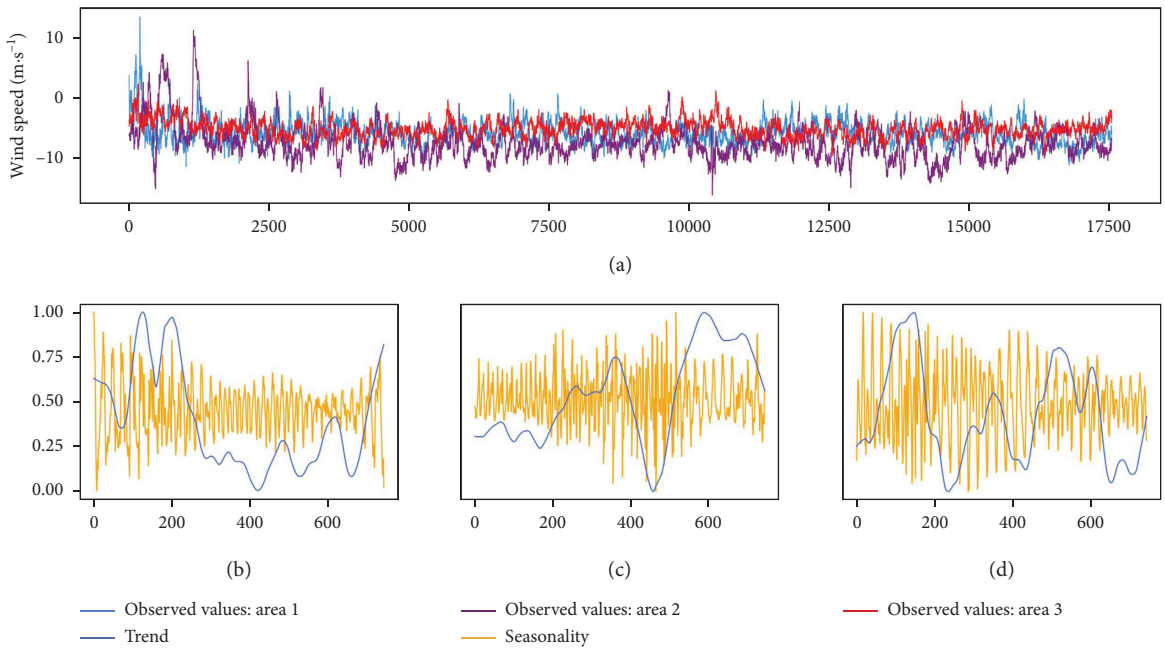


Figure 7: Visualization of datasets used in experiments. (a) Original time series. (b–d) Trends and seasonality in areas 1–3.

Table 3: Result of different combination of decomposition and prediction methods.

	Step	LSTM			GRU			Stacked LSTM			Stacked GRU		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
WT	1	0.2745	0.2108	4.2170	0.2574	0.1927	3.8381	0.2982	0.2272	4.5937	0.2882	0.2182	4.3407
	2	0.4507	0.3495	7.0160	0.4574	0.3533	6.9554	0.4782	0.3698	7.4155	0.4746	0.3696	7.2957
	3	0.5372	0.4191	8.3136	0.5463	0.4257	8.3779	0.5645	0.4378	8.8016	0.5681	0.4412	8.7430
EMD	1	0.2304	0.1892	3.2575	0.2709	0.2103	3.8818	0.4361	0.3522	6.0360	0.2691	0.2131	3.6862
	2	0.2724	0.2183	3.7819	0.4051	0.3124	5.7697	0.5206	0.4182	7.2722	0.4167	0.3290	5.8089
	3	0.3316	0.2632	4.5792	0.5425	0.4185	7.7118	0.5835	0.4681	8.3242	0.4485	0.3486	6.3692
VMD	1	0.1598	0.1251	2.4619	0.1385	0.1058	2.0586	0.1682	0.1298	2.4828	0.1436	0.1080	2.0625
	2	0.1896	0.1495	2.9270	0.1643	0.1290	2.5175	0.2087	0.1645	3.1505	0.1733	0.1361	2.6172
	3	0.2048	0.1615	3.1113	0.1779	0.1393	2.7237	0.2257	0.1784	3.3679	0.1889	0.1485	2.7730

example, when the EMD is combined with the GRU and stacked LSTM, the RMSE is 0.2709 and 0.4361, while when the WT is combined with them, the RMSE is 0.2574 and 0.2982. It can be seen from the table that the VMD is better than WT and EMD in all combinations.

The RMSE of GRU is lower than LSTM under the three decomposition methods, and the RMSE of stacked GRU is lower than the stacked LSTM. This shows that GRU shows better performance than the LSTM in both single layer and multiple layers. However, the metrics of stacked GRU are higher than GRU, and the metrics of stacked LSTM are higher than LSTM. For example, the 1-step RMSE of VMD-Stacked GRU is 0.1436, and the 1-step RMSE of VMD-GRU is 0.1385. This shows that the stacked GRU and stacked LSTM are degraded when combined with the VMD. This degradation is found in the models based on all three decomposition methods.

4.3. Improvement of VMD-Stacked GRU by Residual Connections. After the above analysis, the VMD-GRU is determined as the best combination of decomposition-prediction methods, and the VMD-Stacked GRU is determined as the second best combination. In order to confirm that residual connections improved VMD-Stacked GRU to make it surpass the VMD-GRU method, a comparative experiment was carried out. In the experiment, two kinds of residual connections were used. Residual connections (a) represent the structure shown in Figure 4 and equation (8), and residual connections (b) change the single-layer skipping to double-layer skipping. The results are shown in Table 4.

The VMD-Stacked GRU with residual connections (a) outperforms that with residual connections (b) in most metrics. The VMD-stacked GRU with residual connections (a) perform slightly worst than that with residual connections (b) only on the 1st and 3rd steps of Area 1 and the 3rd step of Area 3. Therefore, it can be considered that residual connections (a) are more suitable than residual connections (b) for wind speed prediction tasks. The VMD-Stacked GRU with residual connections (a) outperforms VMD-GRU in most metrics. It is illustrated that residual connections (a) solve the degradation of VMD-Stacked GRU and make it surpass VMD-GRU.

Table 4: Comparison of VMD-based models with and without residual connections.

	Step	VMD-GRU			VMD-Stacked GRU without residual connections			VMD-Stacked GRU with residual connections (a)			VMD-Stacked GRU with residual connections (b)		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
Area 1	1	0.1385	0.1058	2.0586	0.1436	0.1080	2.0625	0.1369	0.1039	1.9920	0.1353	0.1027	2.0204
	2	0.1643	0.1290	2.5175	0.1733	0.1361	2.6172	0.1583	0.1221	2.2883	0.1583	0.1238	2.4080
	3	0.1779	0.1393	2.7237	0.1889	0.1485	2.7730	0.1730	0.1332	2.5687	0.1655	0.1301	2.5090
Area 2	1	0.1573	0.1186	1.8829	0.1740	0.1295	2.0598	0.1560	0.1172	1.8288	0.1708	0.1280	1.9927
	2	0.2090	0.1622	2.7126	0.2209	0.1712	2.8313	0.2141	0.1665	2.8083	0.2234	0.1733	2.7884
	3	0.2617	0.2044	3.5049	0.2576	0.2004	3.3341	0.2500	0.1950	3.3832	0.2558	0.1994	3.2428
Area 3	1	0.1031	0.0771	1.5846	0.1036	0.0772	1.5947	0.1039	0.0776	1.6081	0.1056	0.0789	1.6002
	2	0.1253	0.0972	2.0268	0.1255	0.0972	2.0249	0.1252	0.0975	2.0334	0.1255	0.0975	2.0457
	3	0.1367	0.1067	2.2642	0.1361	0.1063	2.2395	0.1358	0.1062	2.2520	0.1347	0.1053	2.2406

4.4. Parametric study. To determine the optimal parameters of the proposed model, a detailed parametric study is carried out. The parameters to be determined are the network parameters and training parameters of the stacked GRU. The network parameters include the number of hidden neural units of each layer. The training parameters include batch size, maximum epochs, and the learning rate of the optimizer.

The training parameters are first determined, followed by the network parameters. When determining the training parameters, the network parameters are set in advance according to the latest research. In the article in [4], the optimal two-layer GRU is determined with 512 units in the first layer and 32 units in the second layer. Since the Adam optimizer outperforms classical optimizers such as RMSProp, SGD, and Adagrad [41], it is adopted in the experiment, and its learning rate is searched in {0.1, 0.001, 0.0001}. Batch size is searched in {8, 16, 25, 32, 64} since a study in [42] showed that a smaller batch helps to model training. We set maximum epochs = 50 and use TensorFlow 2.3.1's callback function [43] to monitor the lowest loss value of the validation set in epochs iteration. The parametric study is carried out on area 1, and results are average of three time steps. The RMSE results are shown in Figure 8.

The above results show that the best configuration of training parameters is batch size = 24 and learning rate = 0.001. The stacked GRU network parameters are, respectively, marked as (a^*, a^*), (a, a^*, a^*), and (a, b, b^*, b^*) according to different layers. For example, (a^*, a^*) represents two-layer GRU, and the units number is a; and * means that there are residual connections in this layer. Units search is firstly conducted in {10, 100, 200, . . ., 600} and then a more accurate search is conducted in the best interval. The RMSE results are shown in Figure 9.

It can be seen that $(500, 50, 50^*, 50^*)$ are the optimal network parameters. Above all, the best parametric configuration set is shown in Table 5.

4.5. Result of Multistep Wind Speed Forecasting. To prove the superiority of the proposed model VMD-Stacked GRU with residual connections, we choose seven time series and machine learning models, as well as a published EMD-based model [26], as baseline models. These models directly learn wind speed series without composition. Through Autocorrelation Function and Partial Autocorrelation Function diagrams, the ARIMA parameters $p = 2$, $d = 1$, and $q = 12$ are determined. The support vector machine regression (SVR) uses RBF kernel and establishes the relationship between past information and each forecast time step. The structure of MLP is a four-layer structure with 400, 32, 32, and 32 units in each layer, respectively. The EMD-PE-ANN reconstructs IMFs into IMF_1, IMF_2, and $\sum_{i=3}^{9} IMF_i$. 5-Fold cross-validation strategy is used to obtain the results in Tables 6–8. It can be seen from Tables 6–8 that, compared with predicting wind speed directly, the proposed model has lower error metrics at three time steps in three areas. Most of the error metrics of GRU are lower than those of LSTM, especially in area 2. Most of the error metrics of stacked GRU are lower than those of stacked LSTM, and the RMSE is lower than stacked LSTM only on the 3rd step of area 1 and the first step of area 3. When directly predicting wind speed, the stacked model also has a less obvious degradation effect. For example, compared with the GRU, the RMSE of the stacked GRU shows degradation in the 3 steps of area 1 and the 1st step of area 2 and area 3. In addition, although not surpassing the proposed model, the two classic methods, ARIMA and SVR, have relatively good metrics which are close to GRU.

To further illustrate the superiority of the proposed model, the following figures show the comparison curves of models in area 1. It can be seen from Figure 10 that the fitting effect of the prediction curve (red line) of the proposed model is significantly higher than those of the other models. The values at the last input time step (Persistence) and the overall mean values of inputs (Average) are also added in the figure as benchmarks, and the experiment results show that the proposed model surpasses the benchmarks.

5. Discussion

The case study concludes that, compared with other direct prediction or decomposition-based prediction models, the proposed VMD-Stacked GRU model with residual connections is more accurate in multistep forecasting. The proposed model performs well on the ERA5 sea surface wind

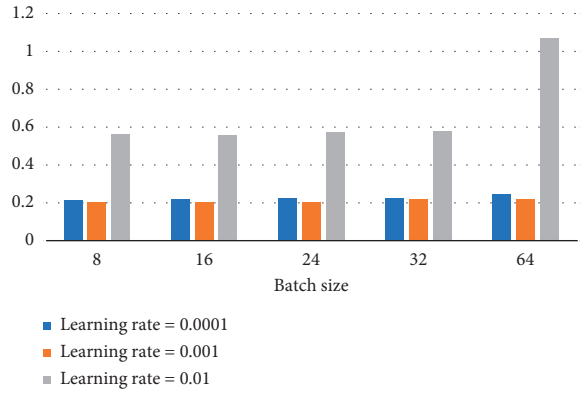


Figure 8: The RMSE results of determining batch size and learning rate.

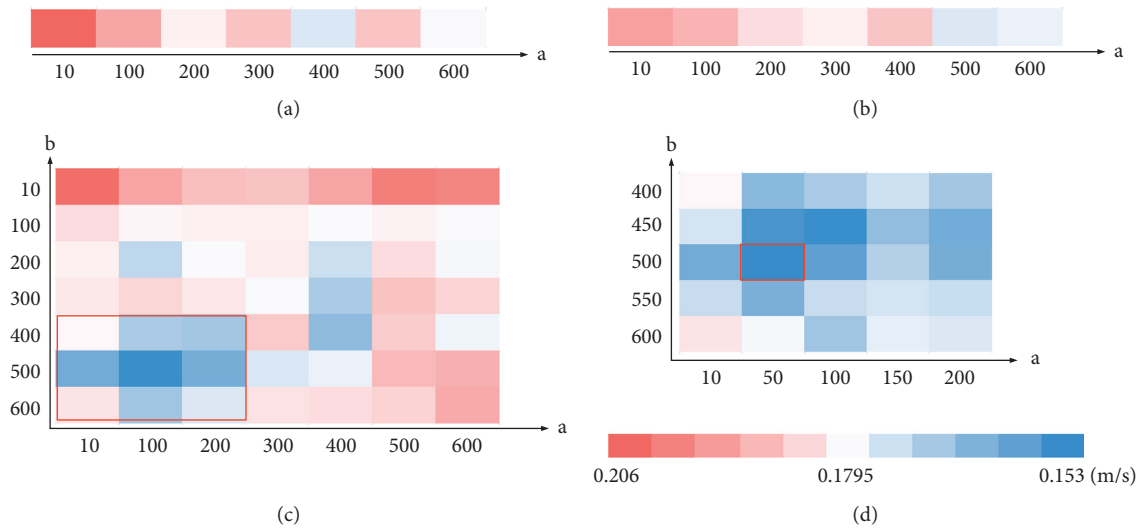


Figure 9: The RMSE results of determining network parameters. (a) Search for 2-layer network. (b) Search for 3-layer network. (c) First search for 4-layer network. (d) Second search for 4-layer network.

Table 5: Parameters of the proposed model.

Stacked GRU predictor	Parameters
GRU layer 1	Number of units: 500 Input shape: (24, 1)
GRU layer 2	Number of units: 50
GRU layer 3 (*)	Number of units: 50
GRU layer 4 (*)	Number of units: 50
Flatten layer	None
Dense layer	Number of units: 3 Output shape: (3)
General setting	Batch size: 24 Learning rate: 0.001

speed datasets of three ocean areas around the world. Compared with the classic model WT-LSTM, the proposed model has lower errors metrics. According to the case study, the superior performance of the proposed model is due to the three following reasons:

(1) The VMD is an excellent decomposition method, and its error in combination with LSTM, GRU, stacked LSTM, and stacked GRU is lower than the combination of WT or EMD and these methods.

(2) The GRU is a better forecasting model than LSTM. In the cases of direct prediction, direct prediction after stacking, and decomposition-based prediction after stacking, the GRU's error metrics are lower than LSTM.

(3) The residual connections can alleviate the degradation of stacked GRU and improve its learning ability. The overall error metrics of VMD-Stacked GRU with residual connections are lower than those of VMD-GRU and VMD-Stacked GRU without residual connections.

Table 6: Comparison of the proposed model and other models (a).

	Step	ARIMA			SVR			MLP		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
Area 1	1	0.2501	0.1893	7.1170	0.3073	0.2202	8.4358	0.3191	0.2556	8.3750
	2	0.4589	0.3959	12.3518	0.4920	0.3909	13.5945	0.5023	0.3987	13.1154
	3	0.6008	0.4738	15.7427	0.6381	0.5035	17.3020	0.6510	0.5144	16.7805
Area 2	1	0.3397	0.2433	4.9072	0.4701	0.3495	7.1652	0.4557	0.3415	6.6360
	2	0.5393	0.3925	5.9793	0.6869	0.5114	10.2742	0.6694	0.5009	9.5826
	3	0.8762	0.6341	13.5656	1.0856	0.8040	17.2215	1.0842	0.8062	15.9731
Area 3	1	0.2093	0.1665	3.9011	0.1806	0.1425	3.3526	0.2572	0.2151	4.8661
	2	0.3827	0.2977	6.9902	0.3405	0.2691	6.5123	0.3704	0.3028	7.2131
	3	0.4799	0.3779	7.8140	0.4708	0.3858	9.1068	0.3077	0.2298	4.6195

Table 7: Comparison of the proposed model and other models (b).

	Step	LSTM			GRU			Stacked LSTM		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
Area 1	1	0.2454	0.1901	7.2279	0.2428	0.1851	7.1381	0.2494	0.1915	7.0653
	2	0.4466	0.3547	13.8548	0.4522	0.3570	13.8690	0.4857	0.3636	12.1895
	3	0.6002	0.4756	16.1619	0.5941	0.4696	16.0096	0.5988	0.4721	15.7203
Area 2	1	0.4641	0.3422	7.1769	0.3931	0.2770	5.4830	0.4002	0.2963	5.6956
	2	0.6611	0.4921	9.8987	0.6518	0.4761	9.8307	0.6342	0.4696	9.3063
	3	1.0504	0.7804	16.2258	0.9645	0.7029	14.3546	0.9345	0.7166	15.5719
Area 3	1	0.1811	0.1424	3.3351	0.1755	0.1363	3.1782	0.1672	0.1309	3.0642
	2	0.3248	0.2569	6.0276	0.3248	0.2588	6.1428	0.3174	0.2512	5.8463
	3	0.2385	0.1842	4.0038	0.4536	0.3661	8.6473	0.4445	0.3371	8.6531

Table 8: Comparison of the proposed model and other models (c).

	Step	Stacked GRU			EMD-PE-ANN			VMD-Stacked GRU with residual connections		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
Area 1	1	0.2429	0.1813	6.8944	0.2715	0.2065	7.8447	0.1366	0.1037	4.1950
	2	0.4433	0.3505	11.8503	0.3980	0.3081	10.2622	0.1600	0.1259	5.0163
	3	0.6021	0.4562	15.7775	0.5010	0.3853	13.1785	0.1650	0.1297	4.4211
Area 2	1	0.3973	0.2907	5.5544	0.4111	0.2984	5.6557	0.2427	0.1775	2.3541
	2	0.5506	0.3970	7.5523	0.5109	0.3673	6.6612	0.2837	0.2087	4.1246
	3	0.9100	0.6679	13.2735	0.7215	0.5899	12.2075	0.3917	0.2875	6.1061
Area 3	1	0.1775	0.1386	3.2802	0.1724	0.1349	3.2383	0.1051	0.0803	1.8354
	2	0.3156	0.2500	6.0146	0.2993	0.2565	4.989	0.1264	0.0980	2.3893
	3	0.4443	0.3515	8.5417	0.4009	0.3203	8.0080	0.1320	0.1038	2.5462

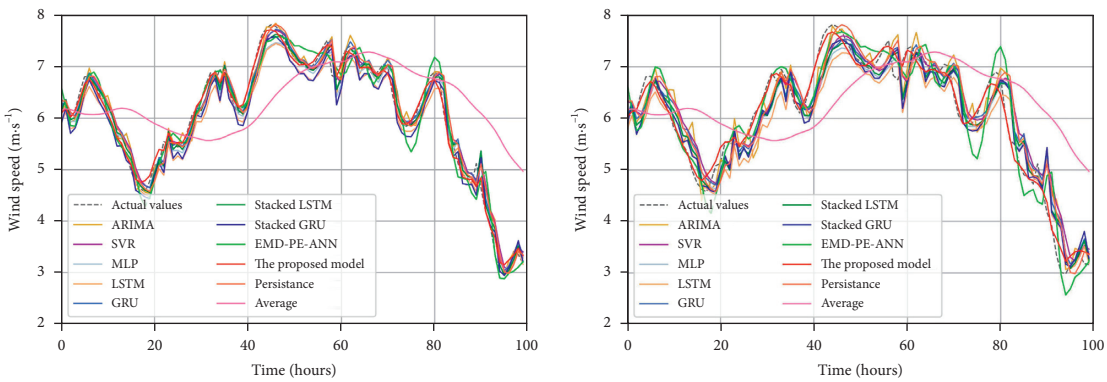
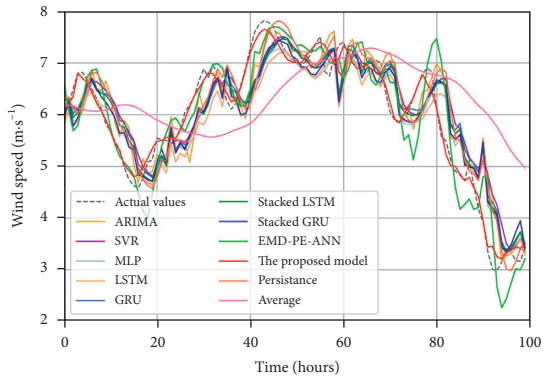


Figure 10: Continued.



(c)

Figure 10: Forecasting results in area 1 of eight models. (a) At the 1st step, (b) at the 2nd step, and (c) at the 3rd step.

6. Conclusions

The sea wind speed forecasting is a key part to guarantee safety of sailing ships. To solve the problems of hourly short-term wind speed forecasting, an ensemble model based on the VMD and stacked GRU is proposed, and the residual connections are used to improve stacked GRU. The model uses VMD to decompose the wind speed series and then uses the stacked GRU model with residual connections to predict each component. In order to prove the performance of the proposed model, three cases from the Pacific, Indian, and Atlantic Oceans are studied. In the experiment, three error metrics, RMSE, MAE, and MAPE, are used to evaluate each time step. Through the case studies, the following conclusions can be illustrated:

(1) Separately predicting the decomposed wind speed sequence and then superimposing it as the final result can improve the prediction effect, and VMD is the most effective one of the various decomposition methods.

(2) The forecast error metrics of VMD-Stacked GRU with residual connections are generally lower than those of ARIMA, SVR, MLP, LSTM, GRU, stacked LSTM, and stacked GRU models at the 1st, 2nd, and 3rd steps.

Data Availability

The wind speed data used to support the findings of this study are freely available and supplied by ECMWF. Requests for access to these data should be made through ECMWF website: https://cds.climate.copernicus.eu/cdsapp.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this study.

References

[1] Z. Zhang and X.-M. Li, "Global ship accidents and ocean swell-related sea states," *Natural Hazards and Earth System Sciences*, vol. 17, no. 11, pp. 2041–2051, 2017.

[2] Y. X. Wu, Q. B. Wu, and J. Q. Zhu, "Data-driven wind speed forecasting using deep feature extraction and LSTM," *IET Renewable Power Generation*, vol. 13, no. 12, pp. 2062–2069, 2019.

[3] R. Wang, C. Li, W. Fu, and G. Tang, "Deep learning method based on gated recurrent unit and variational mode decomposition for short-term wind power interval prediction," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 10, pp. 3814–3827, 2019.

[4] Z. Peng, S. Peng, L. Fu et al., "A novel deep learning ensemble model with data denoising for short-term wind speed forecasting," *Energy Conversion and Management*, vol. 207, p. 112524, 2020.

[5] W. Dong, H. Sun, Z. Li, J. Zhang, and H. Yang, "Short-term wind-speed forecasting based on multiscale mathematical morphological decomposition, K-means clustering, and stacked denoising autoencoders," *IEEE Access*, vol. 8, pp. 146901–146914, 2020.

[6] G. Memarzadeh and F. Keynia, "A new short-term wind speed forecasting method based on fine-tuned LSTM neural network and optimal input sets," *Energy Conversion and Management*, vol. 213, Article ID 112824, 2020.

[7] N. Bokde, A. Feijóo, and K. Kulat, "Analysis of differencing and decomposition preprocessing methods for wind speed prediction," *Applied Soft Computing*, vol. 71, pp. 926–938, 2017.

[8] A. Altan, S. Karasu, and E. Zio, "A new hybrid model for wind speed forecasting combining long short-term memory neural network, decomposition methods and grey wolf optimizer," *Applied Soft Computing*, vol. 100, Article ID 106996, 2021.

[9] J. Wang, S. Qin, Q. Zhou, and H. Jiang, "Medium-term wind speeds forecasting utilizing hybrid models for three different sites in Xinjiang, China," *Renewable Energy*, vol. 76, pp. 91–101, 2015.

[10] H. B. Azad, S. Mekhilef, and V. G. Ganapathy, "Long-term wind speed forecasting and general pattern recognition using neural networks," *IEEE Transactions on Sustainable Energy*, vol. 5, no. 2, pp. 546–553, 2014.

[11] S. S. Soman, H. Zareipour, O. Malik, and P. Mandal, "A review of wind power and wind speed forecasting methods with different time horizons," in *Proceedings of the North American Power Symposium 2010*, Arlington, TX, USA, September 2010.

[12] A. Tascikaraoglu and M. Uzunoglu, "A review of combined approaches for prediction of short-term wind speed and power," *Renewable and Sustainable Energy Reviews*, vol. 34, pp. 243–254, 2014.

[13] W. Yao, P. Huang, and Z. Jia, "Multidimensional LSTM networks to predict wind speed," in *Proceedings of the 2018*

37th Chinese Control Conference (CCC), Wuhan, China, July 2018.

[14] J. C. Pelajo, L. E. T. Brandão, L. L. Gomes, and M. C. Klotzle, "Wind farm generation forecast and optimal maintenance schedule model," *Wind Energy*, vol. 22, no. 12, pp. 1872–1890, 2019.

[15] M. M. H. Khan, N. S. Muhammad, and A. El-Shafie, "Wavelet based hybrid ANN-ARIMA models for meteorological drought forecasting," *Journal of Hydrology*, vol. 590, Article ID 125380, 2020.

[16] C. Li, "Short-term wind speed interval prediction based on ensemble GRU model," *IEEE transactions on sustainable energy*, vol. 11, no. 3, pp. 1370–1380, 2019.

[17] N. Bokde, A. Feijóo, D. Villanueva, and K. Kulat, "A review on hybrid empirical mode decomposition models for wind speed and wind power prediction," *Energies*, vol. 12, no. 2, p. 254, 2019.

[18] M. Lei, "A review on the forecasting of wind speed and generated power," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 4, pp. 915–920, 2014.

[19] Z. Qian, Y. Pei, H. Zareipour, and N. Chen, "A review and discussion of decomposition-based hybrid models for wind energy forecasting applications," *Applied Energy*, vol. 235, pp. 939–953, 2019.

[20] S. N. Singh and A. Mohapatra, "Repeated wavelet transform based ARIMA model for very short-term wind speed forecasting," *Renewable Energy*, vol. 136, pp. 758–768, 2019.

[21] W. Ding and F. Meng, "Point and interval forecasting for wind speed based on linear component extraction," *Applied Soft Computing*, vol. 93, Article ID 106350, 2020.

[22] C. N. Babu and B. E. Reddy, "A moving-average filter based hybrid ARIMA-ANN model for forecasting time series data," *Applied Soft Computing*, vol. 23, pp. 27–38, 2014.

[23] A. K. Fard and M.-R. Akbari-Zadeh, "A hybrid method based on wavelet, ANN and ARIMA model for short-term load forecasting," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 26, no. 2, pp. 167–182, 2014.

[24] Ü.Ç. Büyükşahin and Ş. Ertekin, "Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition," *Neurocomputing*, vol. 361, pp. 151–163, 2019.

[25] H.-F. Yang and Y.-P. P. Chen, "Representation learning with extreme learning machines and empirical mode decomposition for wind speed forecasting methods," *Artificial Intelligence*, vol. 277, Article ID 103176, 2019.

[26] J. J. Ruiz-Aguilar, I. Turias, J. González-Enrique, D. Urda, and D. Elizondo, "A permutation entropy-based EMD-ANN forecasting ensemble approach for wind speed prediction," *Neural Computing & Applications*, vol. 33, no. 7, pp. 2369–2391, 2021.

[27] N. E. Huang, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[28] H. Liu, X. Mi, and Y. Li, "Smart multi-step deep learning model for wind speed forecasting based on variational mode decomposition, singular spectrum analysis, LSTM network and ELM," *Energy Conversion and Management*, vol. 159, pp. 54–64, 2018.

[29] A. Kang, "Short-term wind speed prediction using EEMD-LSSVM model," *Advances in Meteorology*, vol. 2017, Article ID 6856139, 22 pages, 2017.

[30] Z. Liu, R. Hara, and H. Kita, "24 h-ahead wind speed forecasting using CEEMD-PE and ACO-GA-based deep learning neural network," *Journal of Renewable and Sustainable Energy*, vol. 13, no. 4, Article ID 046101, 2021.

[31] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 531–544, 2013.

[32] G. Zhang, H. Liu, J. Zhang et al., "Wind power prediction based on variational mode decomposition multi-frequency combinations," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 2, pp. 281–288, 2019.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[34] K. He, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 2016.

[35] Y. Wu, "Google's neural machine translation system: bridging the gap between human and machine translation," 2016, http://arxiv.org/abs/1609.08144.

[36] J. Wang, B. Peng, and X. Zhang, "Using a stacked residual LSTM model for sentiment intensity prediction," *Neurocomputing*, vol. 322, pp. 93–101, 2018.

[37] M. Khashei and M. Bijari, "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting," *Applied Soft Computing*, vol. 11, no. 2, pp. 2664–2675, 2011.

[38] T. Wang, M. Zhang, Q. Yu, and H. Zhang, "Comparing the applications of EMD and EEMD on time-frequency analysis of seismic signal," *Journal of Applied Geophysics*, vol. 83, pp. 29–34, 2012.

[39] J. Chung, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, http://arxiv.org/abs/1412.3555.

[40] H. Hersbach, B. Bell, P. Berrisford et al., "The ERA5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020.

[41] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2014, http://arxiv.org/abs/1412.6980.

[42] D. Masters and C. Luschi, "Revisiting small batch training for deep neural networks," 2018, http://arxiv.org/abs/1804.07612.

[43] M. Abadi, A. Agarwal, P. Barham et al., "Tensorflow: large-scale machine learning on heterogeneous distributed systems," 2016, http://arxiv.org/abs/1603.04467.