Contents lists available at ScienceDirect

Expert Systems With Applications

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

Accurate water quality prediction with attention-based bidirectional LSTM and encoder–decoder[☆]

Jing Bi[a], Zexian Chen[a], Haitao Yuan[b,*], Jia Zhang[c]

[a] *School of Software Engineering in Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China*
[b] *School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China*
[c] *Department of Computer Science in the Lyle School of Engineering at Southern Methodist University, Dallas, TX 75205, USA*

ARTICLE INFO

Keywords:
Water quality prediction
Variational mode decomposition
BiLSTM
Attention mechanisms
Encoder–decoder

ABSTRACT

Accurate prediction of water quality indicators can effectively predict sudden water pollution events and reveal them to water users for reducing the impact of water quality pollution. Neural networks, *e.g.*, Long Short-Term Memory (LSTM) and encoder–decoder, have been widely used to predict time series data. However, as the water quality data increases, it becomes unstable and highly nonlinear, and therefore, its accurate prediction becomes a big challenge. To solve it, this work proposes a hybrid prediction method called VBAED to predict the water quality time series. VBAED combines Variational mode decomposition (VMD), a Bidirectional input Attention mechanism, an Encoder with bidirectional LSTM (BiLSTM), and a Decoder with a bidirectional temporal attention mechanism and BiLSTM. The definition of VBAED is an Encoder–Decoder model that uses VMD as mode decomposition, combining BiLSTM with a bidirectional attention mechanism. Specifically, VBAED first adopts VMD to decompose historical data of a predicted factor, and its decomposed results are adopted as the input along with other features. Then, a bidirectional input attention mechanism is adopted to add weights to input features from both directions. VBAED adopts BiLSTM as an encoder to extract hidden features from input features. Finally, the predicted result is obtained by a BiLSTM decoder with a bidirectional temporal attention mechanism. Real-life data-based experiments demonstrate that VBAED obtains the best prediction results compared with other widely used methods.

1. Introduction

As a precious resource, water is closely related to human production and lives. With the emergence of Internet of Things (IoTs) and big data (Fortino et al., 2021; Imran et al., 2021), a large number of high-frequency multivariate time series data have been accumulated in a water environment through large-scale deployment of water quality monitoring sensors in rivers and lakes (Bi et al., 2020). An accurate and real-time water quality prediction method can help predicting sudden water pollution, and provide decision support for water quality detection and warning (Dong et al., 2019). Basically, water quality prediction is a time series prediction problem. Traditional statistical time series prediction methods extract linear relations of data by exponential smoothing, an auto-regressive moving average model, and an Auto-Regressive Integrated Moving Average model (ARIMA) (Guo et al., 2019). For example, Najah et al. (2011) propose a hybrid method based on ARIMA that utilizes the advantages of linear and non-linear machine learning models to predict the water level of Red River.

However, due to the development of monitoring technologies, water quality data has become non-linear and unstable, and it is affected by many factors. Traditional statistical methods do not well perceive subtle water quality changes and capture non-linear characteristics of large-scale water quality series. Then, some researchers turn their attentions to models suitable for handling non-linear data. Zhang et al. (2020) adopt Support Vector Regression (SVR) to predict maximum rainfall in annual and non-monsoon sessions. However, SVR consumes a lot of resources when processing a large amount of data. In addition, most ordinary neural networks cannot capture long-term dependence, and they have problems of gradient disappearance or gradient explosion.

More and more data-driven models based on deep learning are used to realize water quality time series prediction (Baigang et al., 2021; Gao et al., 2023; Guo et al., 2022). As a typical example, LSTM (Principi et al., 2019) can capture long-term dependence and effectively avoid the gradient disappearance problem in traditional recurrent neural networks. Liu et al. (2019) adopt LSTM to establish a set of water quality

[☆] This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 62073005 and 62173013, the Beijing Natural Science Foundation, China under Grant 4232049 and the Fundamental Research Funds for the Central Universities, China under Grant YWF-23-03-QB-015.

[*] Corresponding author.
E-mail addresses: bijing@bjut.edu.cn (J. Bi), chenzexian@emails.bjut.edu.cn (Z. Chen), yuan@buaa.edu.cn (H. Yuan), jiazhang@smu.edu (J. Zhang).

https://doi.org/10.1016/j.eswa.2023.121807
Received 19 August 2022; Received in revised form 21 September 2023; Accepted 22 September 2023
Available online 4 October 2023

monitoring systems based on IoT sensors to predict the water quality of Guazhou water sources in Yangzhou, China. Although LSTM has been widely used in time series prediction, it can only encode from front to back and cannot capture the information from back to front. At the same time, due to increasing features of time series, some noise features may have negative impact. Qin et al. propose Dual-Stage Attention-Based Recurrent Neural Network (DA-RNN) (Qin et al., 2017), which adopts an attention mechanism to add attention weights to the input features. However, it cannot capture the information from back to front. In addition, water quality data may contain both important information and noise, which cannot be separated by above methods.

To solve the above problems, this work proposes a hybrid method called VBAED. VBAED integrates Variational mode decomposition (VMD), a Bidirectional input Attention mechanism, an Encoder with bidirectional LSTM (BiLSTM), and a Decoder with a bidirectional temporal attention mechanism and BiLSTM. Main contributions of this work are summarized as follows.

(1) BiLSTM is adopted as an encoder to capture features from two directions. BiLSTM is improved with a bidirectional input attention mechanism to add attention weights to the input independently from two directions.
(2) BiLSTM is adopted as a decoder, and it is combined with a bidirectional temporal attention mechanism to capture the long-term dependence, thereby adaptively selecting important hidden states of the encoder across all time steps, and decoding them from two directions.

For clarity, we note major differences between the current work and our prior one (Bi et al., 2022) as follows.

1. Different from Bi et al. (2022), this work adopts a bidirectional LSTM as the decoder to better capture the information of the hidden states.
2. Different from Bi et al. (2022), this work adopts a bidirectional temporal attention mechanism in the decoder to capture important hidden states and ignore ones with negative effects.
3. The work in Bi et al. (2022) adopts a single type of water quality data collected from an automatic water quality station in a river in the Beijing–Tianjin–Hebei (BTH) region from September 2018 to December 2021. Different from it, this work further adopts another type of water quality data of a section of the Alabama River from May 2017 to August 2019 to demonstrate the robustness and prediction accuracy of our proposed VBAED.

The main structure of this work is as follows. First, we describe the related work in Section 2. We introduce details of the proposed method in Section 3, and present experimental results and discussion in Section 4. Finally, Section 5 draws the conclusion.

2. Related work

Accurate and real-time water quality prediction helps water environment practitioners to deal with unexpected water pollution events in time and protect the river ecological environment. Advanced sensors are widely deployed to detect, transmit and measure more complex and nonlinear water quality data (Wu et al., 2020). The strong non-linearity of the water quality data brings a challenge for accurate water quality prediction (Chang et al., 2015). Currently, its prediction methods are generally divided into classical prediction ones and deep learning-based ones.

2.1. Classical prediction methods

Traditional linear prediction methods, such as models of Auto-Regressive (AR) (Yule, 1927) and ARIMA (Box & Pierce, 1970), are widely used in the time series prediction (Sharma et al., 2021). Moeeni et al. (2017) adopt ARIMA to predict the temperature and the flow of rivers with four river datasets, proving that ARIMA works well in the water quality prediction. Guo et al. (2019) propose a prediction model based on ARIMA, which predicts the future vehicle speed and road slope with appropriate accuracy. Ding et al. (2019) propose a hybrid model based on ARIMA and generalized autoregressive conditional heteroskedasticity to predict the subway short-term ridership. However, these above methods all adopt the approximately linear fitting of time series, and therefore, non-linear features in the data cannot be effectively captured.

To capture non-linear characteristics in the time series data, researchers have turned their attentions to advanced methods suitable for non-linear data. Among them, the Support Vector Machine (SVM) is a widely used typical method (Eseye & Lehtonen, 2020). Bae et al. (2017) propose an hourly solar irradiance prediction scheme based on SVM. Their experimental results show that SVM significantly improves the prediction accuracy of solar irradiance. Yang et al. (2015) introduce an SVM-enhanced Markov model to obtain better prediction accuracy of short-term wind power. However, although above methods solve the problems existing in traditional methods, they need a lot of memory resources when processing large-scale data. Artificial Neural Networks (ANNs) have been widely adopted in the time series prediction because of its prediction and generalization abilities. Buhan and Çadırcı (2015) propose a real-time wind-electric power generation forecasting method based on the combination of ANN and SVM. Experimental results show that the proposed model outperforms traditional methods in terms of short-term prediction accuracy. However, because of its simplicity, ANN still has lower accuracy than deep learning methods in the time series prediction.

Different from these studies, this work proposes a hybrid deep learning method named VBAED, which effectively captures complex features and long-term correlations in the time series and improves the prediction accuracy.

2.2. Deep learning-based methods

Deep learning, with its powerful ability to automatically extract features and process large-scale data, has become a hot topic in current studies (Bandara et al., 2021; Yang et al., 2022; Zhou et al., 2023). Recurrent Neural Networks (RNNs) capture long-term dependencies in the time series, and they are widely used in different areas, *e.g.*, traffic flow forecasting (Wang et al., 2021), renewable energy generation forecasting (Xia et al., 2021), or wind power prediction (Zhou et al., 2019). Among many variants of RNNs, LSTM is the most popular method to solve the problems of gradient explosion and disappearance in the training of RNNs (Hochreiter & Schmidhuber, 2019). Zhang et al. (2019) propose a model based on LSTM that adopts multivariable inputs to predict groundwater depth in agricultural areas. Considering multivariable characteristics of electricity consumption, (Kong et al., 2019) propose a short-term residential load forecasting method based on LSTM. Hou et al. (2021) propose a hybrid deep neural network including graph convolutional network and LSTM to investigate graph-structured interactions among stocks and the fluctuation of stock prices. Although LSTM captures long-term dependence, it cannot distinguish the importance of different features for multi-feature prediction tasks.

To solve the above problems, recent studies combine the attention mechanism with LSTM (Zheng et al., 2021) and obtain excellent performance in some sequence modeling tasks. The attention mechanism can be regarded as a feature contribution evaluation mechanism to improve the efficiency of neural networks by selecting important features. Recent studies have shown that the combination of the attention mechanism and LSTM can be applied to realize the time series prediction with good performance. Hsu et al. (2022) propose a temporal convolution-based LSTM network with the attention mechanism to predict the remaining useful life of the equipment. Experiments show that it obtains lower prediction errors than temporal convolutional networks and LSTM. Xie et al. (2022) propose a time-aware attention

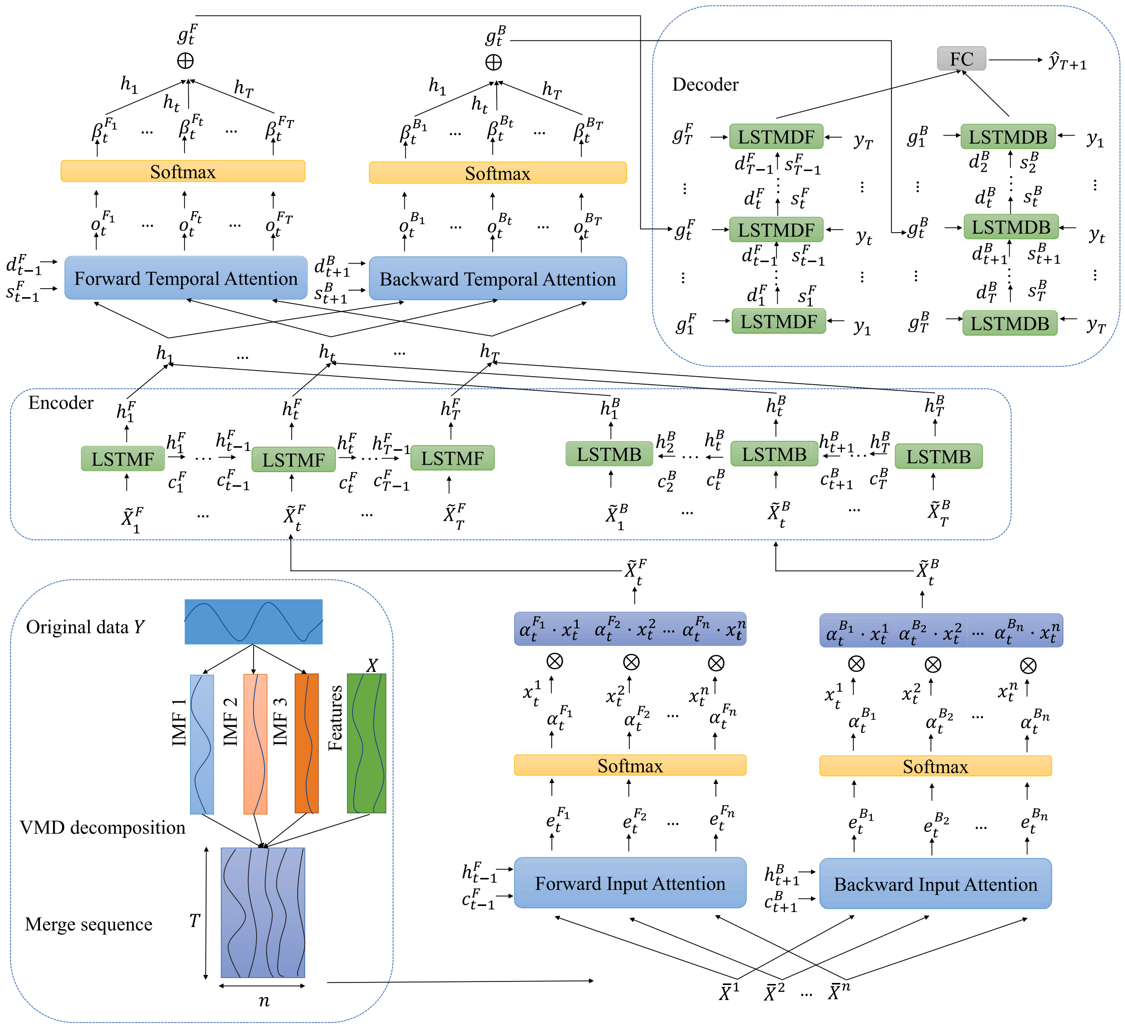

Fig. 1. The proposed VBAED model including VMD decomposition, a bidirectional input attention layer, a BiLSTM encoder, a bidirectional temporal attention layer, a BiLSTM decoder, and a fully connected layer.

module to extract behavior information from consecutive historical transactions with time intervals, which enables the proposed model to capture periodicity and behaviors in historical transaction data.

Different from the above studies, we innovatively combine the attention mechanism with BiLSTM in the input dimension and time one, and adopt VMD to decompose the water quality data and separate important modes from noise ones to further improve the prediction accuracy. Specifically, the historical data of the predicted factor is decomposed into multiple modes with VMD. The modes and other features are then encoded by BiLSTM with the bidirectional input attention mechanism, and they are decoded by BiLSTM with the bidirectional temporal attention mechanism for yielding the final prediction.

3. Proposed methodology

This section introduces VBAED in detail. By decomposing the historical data of a predicted factor in the water quality time series through VMD, VBAED reduces the nonlinearity and volatility of the input data and improves the prediction accuracy. The proposed bidirectional input attention mechanism can adaptively select important features in the input. BiLSTM (Zou et al., 2022) can capture more long-term dependencies. The bidirectional temporal attention mechanism extracts important features in the time dimension, and obtains the predicted value with BiLSTM as the decoder. We integrate these methods to further improve the prediction accuracy. Fig. 1 illustrates the proposed VBAED model.

3.1. Sequence problem statement

Sequence modeling is widely adopted in many fields, such as natural language processing and time series prediction. Let $X = \{X_1, \dots, X_t, \dots, X_T\} \in \mathbb{R}^{\acute{n}\times T}$ denote a series with a time span of T. $\acute{n}$ denotes the number of original features. In this work, our features include Potential of Hydrogen (pH) and Total Phosphorus (TP). $\bar{X} = \{\bar{X}_1, \dots, \bar{X}_t, \dots, \bar{X}_T\} \in \mathbb{R}^{n\times T}$ denotes a series with a time span of T processed by VMD. n denotes the number of features after VMD. $Y = \{y_1, \dots, y_t, \dots, y_T\} \in \mathbb{R}^{1\times T}$ is a series of ground truth values. $\hat{y}_{T+1}$ and y_{T+1} denote the predicted value and its ground truth one at time step T+1, respectively. We adopt the data in past T time steps to predict the ground truth value at time step T+1 by finding nonlinear mapping from the input value to the ground truth one, which minimizes the prediction error. In the BTH dataset, we adopt values of pH, Total Nitrogen (TN) and TP in past T time steps to predict the TN value at time step T+1. In the Alabama dataset, we adopt the values of Dissolved Oxygen (DO) in past T time steps to predict the DO value at time step T+1. The nonlinear function F(·) that we need to learn is expressed as:

$$\hat{y}_{T+1} = \mathrm{F}(X, Y) \tag{1}$$

3.2. Variational mode decomposition

This work decomposes the historical data of the predicted factor through VMD (Dragomiretskiy & Zosso, 2022), which is an adaptive signal processing method. It iteratively searches for the optimal solution of variational modes, constantly updates each modal function and central frequency, and obtains several Intrinsic Mode Functions (IMFs) (Wang & Li, 2018). The variational problem is defined as solving k IMFs to minimize the sum of estimated bandwidth of each mode. Y is decomposed into k modes.

VMD decomposition can reduce the nonlinearity and volatility of time series and avoid the negative impact of mode mixing. Different modal components have different effects on the prediction result. By separating them and combining them with the input attention mechanism, VBAED has the ability to adaptively select important modes, filter out the noisy modes from multiple modes, and focus on the modes containing important information. The modal components can be selectively removed according to the experimental results.

This work decomposes the target value series into three components by VMD, which are used as features. This guides neural networks to more attentively learn more complex features and improve the prediction accuracy. The new input after VMD is $\bar{X} = \{\bar{X}_1, \ldots, \bar{X}_t, \ldots, \bar{X}_T\} \in \mathbb{R}^{n \times T}$ where $n = \acute{n}+3$.

3.3. Encoder with BiLSTM and bidirectional input attention

As a typical variant of RNNs, LSTM (Principi et al., 2019) can avoid the gradient explosion and gradient disappearance existing in conventional RNNs. It can effectively capture the long dependence, and it is often used for encoding in natural language processing and time series prediction. However, LSTM cannot encode information from back to front. In the time series prediction, the information from back to front is hidden, which cannot be obtained with LSTM as the encoder.

Therefore, this work adopts BiLSTM as the encoder, which solves the disadvantage that LSTM cannot obtain the information from back to front. BiLSTM consists of two independent LSTM units. The first LSTM unit is called LSTMF, which encodes information from front to back. The second LSTM unit is called LSTMB, which encodes information from back to front. Then, the information from the two directions is combined to obtain the hidden state h_t of the encoder at time step t. Especially, at time step t, LSTMF computes its hidden state h_t^F based on the previous hidden state h_{t-1}^F at time step t−1, the cell state c_{t-1}^F at time step t−1, and the input $\bar{X}_t$. LSTMB computes its hidden state h_t^B based on the hidden state h_{t+1}^B, the cell state c_{t+1}^B and the input $\bar{X}_t$. Then, the forward hidden state h_t^F and the backward hidden state h_t^B are combined into the hidden state of BiLSTM. LSTMF and LSTMB are two independent LSTM units, and they do not share parameters. h_t^F, h_t^B and h_t are given as:

$$h_t^F = \mathrm{LF}(h_{t-1}^F, c_{t-1}^F, \bar{X}_t) \tag{2}$$

$$h_t^B = \mathrm{LB}(h_{t+1}^B, c_{t+1}^B, \bar{X}_t) \tag{3}$$

$$h_t = [h_t^F; h_t^B] \tag{4}$$

where LF is an LSTMF unit and LB is an LSTMB unit. m denotes the hidden state size of each BiLSTM unit, and $h_t^F \in \mathbb{R}^m$, $h_t^B \in \mathbb{R}^m$, $h_t \in \mathbb{R}^{2m}$.

To better capture important features, this work designs an input attention mechanism for BiLSTM. The attention mechanism can adaptively select important features from a large number of features and focus on them. An attention weight represents the importance of information. Since LSTMF and LSTMB are two LSTM units with independent parameters, we add an input attention mechanism layer to them, respectively. The forward input attention layer is for LSTMF and the backward one is for LSTMB. LSTMF and LSTMB encode from different directions, and the input attention mechanism focuses on different features. The advantage of designing an individual input attention mechanism for each of them independently is that they can adaptively extract important features in two directions, which can improve the robustness of the model and the prediction accuracy.

Then, the details of the input attention mechanism in BiLSTM are given here. For the lth feature $\bar{X}^l = (x_1^l, x_2^l, \ldots, x_T^l) \in \mathbb{R}^T$ in $\bar{X}$, we refer to h_{t-1}^F and c_{t-1}^F in the LSTMF unit to construct the forward input attention mechanism. We adopt h_{t+1}^B and c_{t+1}^B in the LSTMB unit to construct the backward input attention mechanism.

$\alpha_t^{F_l}$ and $\alpha_t^{B_l}$ denote the forward attention weight and the backward one of the lth input feature (x_t^l) at time step t. They are obtained as:

$$\alpha_t^{F_l} = \frac{\exp(e_t^{F_l})}{\sum_{i=1}^n \exp(e_t^{F_i})} \tag{5}$$

$$\alpha_t^{B_l} = \frac{\exp(e_t^{B_l})}{\sum_{i=1}^n \exp(e_t^{B_i})} \tag{6}$$

In (5) and (6), energy scores $e_t^{F_l}$ and $e_t^{B_l}$ are transformed by the softmax function to ensure that $\alpha_t^{F_l}$ and $\alpha_t^{B_l}$ are in the range of (0,1), respectively.

$$e_t^{F_l} = \mathbf{v}_e^{F\top} \tanh(\mathbf{W}_e^F [h_{t-1}^F; c_{t-1}^F] + \mathbf{U}_e^F \bar{X}^l) \tag{7}$$

$$e_t^{B_l} = \mathbf{v}_e^{B\top} \tanh(\mathbf{W}_e^B [h_{t+1}^F; c_{t+1}^B] + \mathbf{U}_e^B \bar{X}^l) \tag{8}$$

where $\mathbf{v}_e^F \in \mathbb{R}^T$, $\mathbf{W}_e^F \in \mathbb{R}^{T \times 2m}$, $\mathbf{U}_e^F \in \mathbb{R}^{T \times T}$, $\mathbf{v}_e^B \in \mathbb{R}^T$, $\mathbf{W}_e^B \in \mathbb{R}^{T \times 2m}$ and $\mathbf{U}_e^B \in \mathbb{R}^{T \times T}$ are the parameters that can be learned. The parameters with superscript F belong to LSTMF and those with superscript B belong to LSTMB.

Then, we obtain two new inputs at time step t and input them into LSTMF and LSTMB, respectively. Specifically, $\tilde{X}_t^F$ and $\tilde{X}_t^B$ denote the inputs of LSTMF and LSTMB at time step t, respectively, which are given as:

$$\tilde{X}_t^F = (\alpha_t^{F_1} x_t^1, \alpha_t^{F_2} x_t^2, \ldots, \alpha_t^{F_n} x_t^n) \tag{9}$$

$$\tilde{X}_t^B = (\alpha_t^{B_1} x_t^1, \alpha_t^{B_2} x_t^2, \ldots, \alpha_t^{B_n} x_t^n) \tag{10}$$

Then, (4) is given as:

$$h_t = [\mathrm{LF}(h_{t-1}^F, c_{t-1}^F, \tilde{X}_t^F); \mathrm{LB}(h_{t+1}^B, c_{t+1}^B, \tilde{X}_t^B)] \tag{11}$$

The forward and backward attention mechanisms are independent and learnable. VBAED can learn the attention weights in two directions, respectively and extract important features adaptively. This method enhances the robustness of the model and improves the prediction accuracy of the water quality.

3.4. Decoder with BiLSTM and bidirectional temporal attention

In the decoder, VBAED adopts BiLSTM with the bidirectional temporal attention mechanism. The bidirectional temporal attention mechanism naturally and accurately captures the key information of hidden states. Similar to the encoder, the BiLSTM decoder consists of two independent LSTM units including LSTMDF and LSTMDB. The former decodes information from front to back, and the latter decodes information from back to front. m denotes the hidden state size of each BiLSTM unit. h_i denotes the ith hidden state of the encoder.

Details of the bidirectional temporal attention mechanism in BiLSTM are given here. The forward attention weight and the backward one of the ith hidden state (h_i) at time step t are denoted by $\beta_t^{F_i}$ ($\beta_t^{F_i} \in (0,1)$) and $\beta_t^{B_i}$ ($\beta_{B_t}^i \in (0,1)$), respectively. $\beta_t^{F_i}$ is calculated based on the hidden state d_{t-1}^F ($d_{t-1}^F \in \mathbb{R}^p$), and the cell state s_{t-1}^F ($s_{t-1}^F \in \mathbb{R}^p$) of LSTMDF at time step t−1. Besides, $\beta_t^{B_i}$ is calculated based on the hidden state d_{t+1}^B ($d_{t+1}^B \in \mathbb{R}^p$), and the cell state s_{t+1}^B ($s_{t+1}^B \in \mathbb{R}^p$) of LSTMDB at time step t+1. $o_t^{F_i}$ and $o_t^{B_i}$ denote the energy score of h_i at time step t, which are obtained as:

$$o_t^{F_i} = \mathbf{v}_o^{F\top} \tanh(\mathbf{W}_o^F [d_{t-1}^F; s_{t-1}^F] + \mathbf{U}_o^F h_i), 1 \le i \le T \tag{12}$$

$$o_t^{B_i} = \mathbf{v}_o^{B\top} \tanh(\mathbf{W}_o^B [d_{t+1}^B; s_{t+1}^B] + \mathbf{U}_o^B h_i), 1 \le i \le T \tag{13}$$

$$\beta_t^{F_i} = \frac{\exp(o_t^{F_i})}{\sum_{q=1}^{T} \exp(o_t^{F_q})} \tag{14}$$

$$\beta_t^{B_i} = \frac{\exp(o_t^{B_i})}{\sum_{q=1}^{T} \exp(o_t^{B_q})} \tag{15}$$

where $[d_{t-1}^F; s_{t-1}^F] \in \mathbb{R}^{2p}$ is the concatenation of d_{t-1}^F and s_{t-1}^F, and $[d_{t+1}^B; s_{t+1}^B] \in \mathbb{R}^{2p}$ is the concatenation of d_{t+1}^B and s_{t+1}^B. $\mathbf{v}_o^{F\top} \in \mathbb{R}^{2m}$, $\mathbf{W}_o^F \in \mathbb{R}^{2m\times 2p}$, $\mathbf{U}_o^F \in \mathbb{R}^{2m\times 2m}$, $\mathbf{v}_o^{B\top} \in \mathbb{R}^{2m}$, $\mathbf{W}_o^B \in \mathbb{R}^{2m\times 2p}$ and $\mathbf{U}_o^B \in \mathbb{R}^{2m\times 2m}$ are learning parameters. The forward and backward context vectors at time step t are denoted by g_t^F and g_t^B, which are the weighted sums of all hidden states of the encoder, respectively.

$$g_t^F = \sum_{i=1}^{T} \beta_t^{F_i} h_i \tag{16}$$

$$g_t^B = \sum_{i=1}^{T} \beta_t^{B_i} h_i \tag{17}$$

We combine g_t^F and g_t^B with the historical ground truth value y_t to obtain new inputs $\tilde{y}_t^F$ and $\tilde{y}_t^B$ for LSTMDF and LSTMDB at time step t, respectively.

$$\tilde{y}_t^F = \mathbf{w}^{F\top}[y_t; g_t^F]+b^F \tag{18}$$

$$\tilde{y}_t^B = \mathbf{w}^{B\top}[y_t; g_t^B]+b^B \tag{19}$$

where $[y_t; g_t^F] \in \mathbb{R}^{2m+1}$ is the concatenation of y_t and g_t^F, and $[y_t; g_t^B] \in \mathbb{R}^{2m+1}$ is the concatenation of y_t and g_t^B. $\mathbf{w}^{F\top} \in \mathbb{R}^{2m+1}$, $b^F \in \mathbb{R}$, $\mathbf{w}^{B\top} \in \mathbb{R}^{2m+1}$ and $b^B \in \mathbb{R}$ are learning parameters. $[y_t; g_t^F]$ and $[y_t; g_t^B]$ are mapped to $\tilde{y}_t^F$ and $\tilde{y}_t^B$, and their dimension equals the input size of the decoder.

$\tilde{y}_t^F$ and $\tilde{y}_t^B$ are used to update the hidden states d_t^F and d_t^B of LSTMDF and LSTMDB at time step t, respectively, *i.e.*,

$$d_t^F = \mathrm{LDF}(d_{t-1}^F, s_{t-1}^F, \tilde{y}_t^F), 1 \le t \le T \tag{20}$$

$$d_t^B = \mathrm{LDB}(d_{t+1}^B, s_{t+1}^B, \tilde{y}_t^B), 1 \le t \le T \tag{21}$$

where LDF is an LSTMDF unit and LDB is an LSTMDB unit.

Finally, the predicted value $\hat{y}_{T+1}$ is obtained as:

$$\hat{y}_{T+1} = \mathbf{v}_y^{\top}(\mathbf{W}_y[d_T^F; g_T^F; d_1^B; g_1^B]+b_w)+b_v \tag{22}$$

where $\mathbf{W}_y \in \mathbb{R}^{p\times(2p+4m)}$ and $b_w \in \mathbb{R}^p$ map $[d_T^F; g_T^F; d_1^B; g_1^B]$ to a vector, the dimension of which equals the hidden state size of the decoder, and $\mathbf{v}_y \in \mathbb{R}^p$ and $b_v \in \mathbb{R}$ are used to yield the final predicted result $\hat{y}_{T+1}$.

3.5. Training procedure

This work adopts the metric of Mean Squared Error (MSE) as a loss function to minimize the difference between the ground truth value of y_{T+1} and the predicted one of $\hat{y}_{T+1}$. The loss function, denoted by Δ, is defined as:

$$\Delta = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_{T+1}^i - y_{T+1}^i)^2 \tag{23}$$

where N denotes the number of training samples.

The reason is that VBAED is smooth and differentiable, and MSE can be used as the loss function to learn the parameters.

4. Experimental evaluation

This section presents our experiments and discusses the results. VBAED is implemented on a server with GTX1080 GPU, 16 GB memory and an Intel (R) Xeon (R) CPU E5-2683. We adopt Adaptive Moment Estimation (Adam) (Kingma & Ba, 2015) as the optimizer to optimize our loss function. The learning rate starts from 0.001 and decreases by 10% every 20 iterations.

Table 1
Statistics of two datasets.

Dataset	Feature dimension	Training set size	Validation set size	Test set size
BTH	3	5000	1000	1200
Alabama	1	15 889	1986	1987

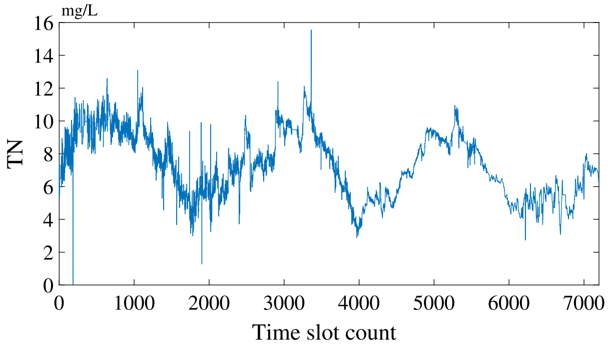

Fig. 2. TN time series data of the BTH dataset.

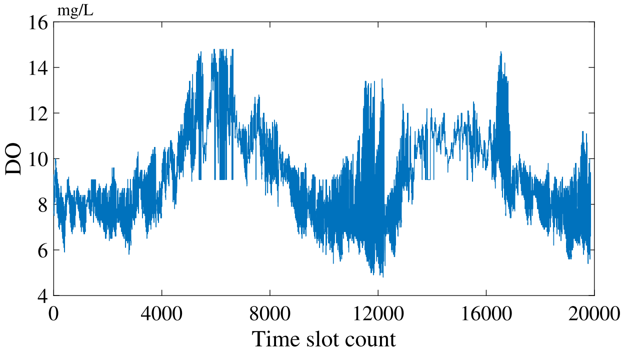

Fig. 3. DO time series data of the Alabama dataset.

4.1. Dataset

To evaluate the performance of different time series prediction methods, we adopt two different real-life datasets including multi-feature and single-feature ones, which are shown in Table 1. The BTH dataset is collected from an automatic water quality station in a river in the Beijing–Tianjin–Hebei region from September 2018 to December 2021. The collection interval is once every 4 h, involving pH, TN and TP. In the experiments, TN is used as the ground truth, and pH and TP are used as features. For a small number of missing values, we adopt linear interpolation to complement it. In total, we have 7200 data samples. We take the first 5000 data samples as the training set, the next 1000 data samples as the validation set, and the remaining 1200 data samples as the test set. The TN time series is shown in Fig. 2.

The Alabama dataset is the water quality data of a section of Alabama River in the United States from May 2017 to August 2019. The data collection interval is one hour. Different from the BTH dataset, the Alabama dataset has only one feature of DO, which is the target value in the Alabama dataset. For a small number of missing values in the Alabama dataset, the linear interpolation method is adopted to complement them. In total, we have 19,862 data samples in the Alabama dataset. In this work, we take the first 15,889 data samples as the training set, the following 1986 data samples as the validation set, and the last 1987 data samples as the test set. The DO time series in the Alabama dataset is shown in Fig. 3.

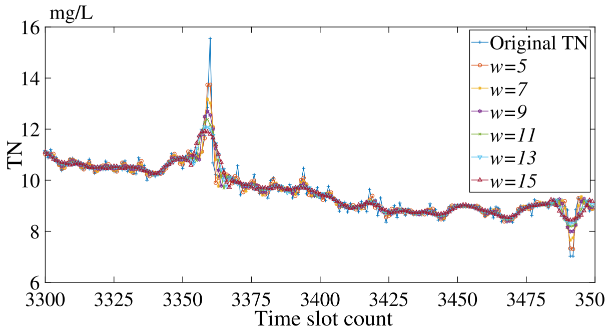

Fig. 4. Filtered TN time series with the SG filter given different w.

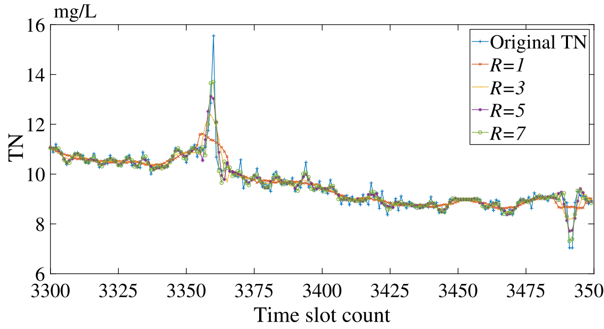

Fig. 5. Filtered TN time series with the SG filter given different R.

4.2. Data preprocessing

Fig. 2 shows that the target value of the BTH dataset has much noise with several peaks, which might be caused by machine failures. The prediction results are severely affected if the model is trained without dealing with noise. Here, the Savitzky Golay (SG) filter (Savitzky & Golay, 1964) is adopted to smooth the time series data of TN, TP, and pH in the BTH dataset to reduce the interference of noise and the influence of local outliers on the overall trend. The SG filter can reduce the interference of noise while maintaining the shape of the original data. Typically, the window size (w) and the highest order term (R) are two important parameters of the SG filter. We first compare different w under the fixed R. Fig. 4 shows the filtered TN time series with the SG filter given different w. It is shown that larger w yields better smoothing result of the data, but too large w changes the trend of the original data.

Then, Fig. 5 shows the filtered TN time series with the SG filter given different R. It is observed that as R increases, the effect of data smoothing is worse, and too low R also changes the trend of data.

Based on Figs. 4 and 5, this work selects the SG filter with w = 11 and R = 5 to smooth the original data of the BTH dataset. Fig. 6 shows the yielded time series smoothed by the SG filter, which is adopted as our experimental data of the BTH dataset. For the Alabama dataset, different from the BTH dataset, we directly predict the target DO value without any preprocessing.

4.3. Benchmark methods

We compare VBAED with widely used benchmark methods: ARIMA (Guo et al., 2019), SVR (Zhang et al., 2020), Extreme Gradient Boosting (XGBoost) (Liu et al., 2021), Back Propagation (BP) (Lu et al., 2016), LSTM (Principi et al., 2019), BiLSTM (Zou et al., 2022) and DA-RNN (Qin et al., 2017). We also add VMD to each baseline method to decompose data, thus resulting in VMD-LSTM (Sun et al., 2019), VMD-BiLSTM, and VMD-DA-RNN.

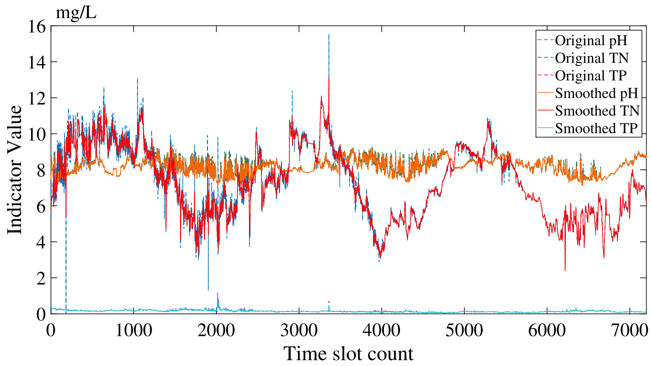

Fig. 6. Smoothed time series of the BTH dataset.

4.4. Evaluation metrics

To verify the performance of VBAED, we adopt three evaluation metrics to compare the prediction accuracy, *i.e.*, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) (Bi et al., 2021).

4.5. Parameter tuning

There are a number of hyperparameters in VBAED that have significant impacts on the performance of prediction. They include the number of time steps in the window (T), optimizer, the encoder hidden state size (m) and the decoder hidden state size (p). T is one of the most important parameters of the model. Fig. 7 shows the change of RMSE in the BTH dataset when T increases. It is shown that when T = 30, RMSE reaches the lowest value. When T increases from 60 to 90, RMSE shows an increasing trend. Therefore, in the BTH dataset, T = 30. Fig. 8 shows the trend of RMSE with the increase of T in the Alabama dataset. It is shown that RMSE achieves the lowest value when T = 30. When T is higher than 30, RMSE shows an increasing trend with the increase of T. Therefore, in the Alabama dataset, the T value of the model is set to 30.

Regarding the choice of optimizer, we compare four kinds of candidate optimizers including Stochastic Gradient Descent (SGD), Adaptive delta (Adadelta), Adaptive gradient algorithm (Adagrad) and Adam. For the BTH dataset, their comparison results are shown in Fig. 9. It is observed that compared with other optimizers, Adam achieves the fastest convergence speed and the lowest loss. Therefore, in the BTH dataset, we choose Adam as the optimizer. For the Alabama dataset, the comparison result of different optimizers is shown in Fig. 10. It is observed that Adam also achieves the fastest convergence speed and the lowest loss in the Alabama dataset. Finally, in the Alabama dataset, we choose Adam as the optimizer of VBAED.

The appropriate hidden state size has significant influence on VBAED. According to Qin et al. (2017), we set $m = p$, and vary m from a set of {16, 32, 64, 128, 256, 512}. Then, the best m is selected by comparing the predicted results. Table 2 shows that in the BTH dataset, RMSE and MAE reach their lowest values and R^2 reaches its highest value when m =64. Table 3 shows that in the Alabama dataset, RMSE and MAE reach their lowest values and R^2 reaches its highest value when m = 64. The above experimental results show that the final parameters obtained by two experiments are the same. The final parameter setting of VBAED in both BTH and Alabama datasets is summarized in Table 4.

To achieve the fair comparison with the benchmark methods, the same tuning process is adopted for the benchmark methods. Table 5 shows the hyperparameter settings of the benchmark methods. For ARIMA, p_a denotes the number of autoregressive terms, d_a denotes that of nonseasonal differences needed for stationarity, and q_a denotes

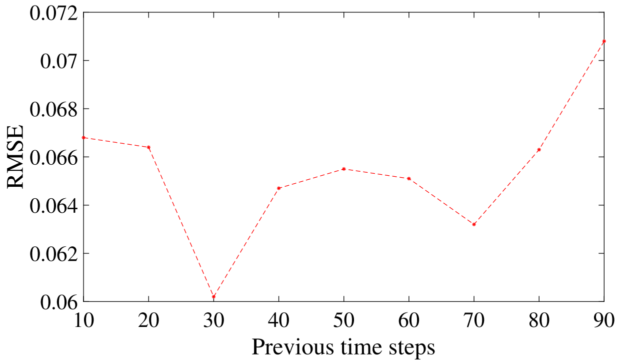

Fig. 7. RMSE of VBAED with varying T in the BTH dataset.

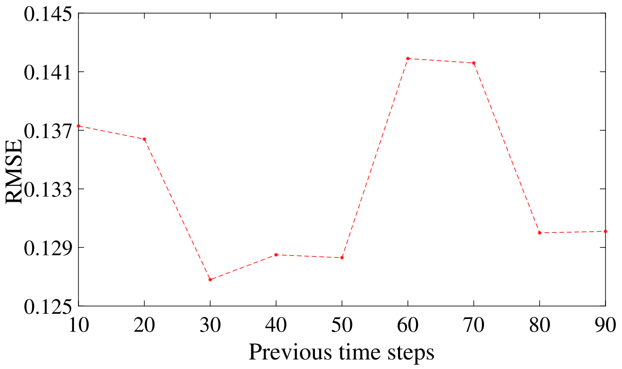

Fig. 8. RMSE of VBAED with varying T in the Alabama dataset.

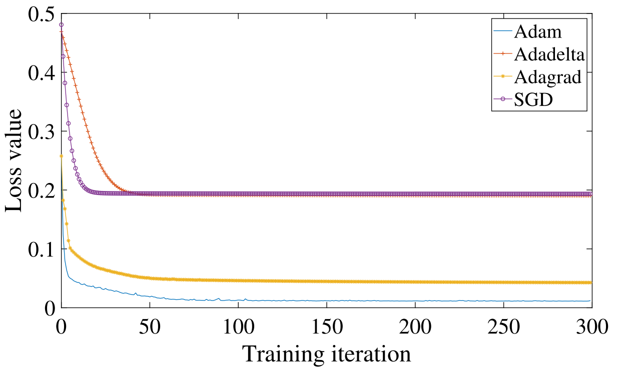

Fig. 9. Loss values for different optimizers in the BTH dataset.

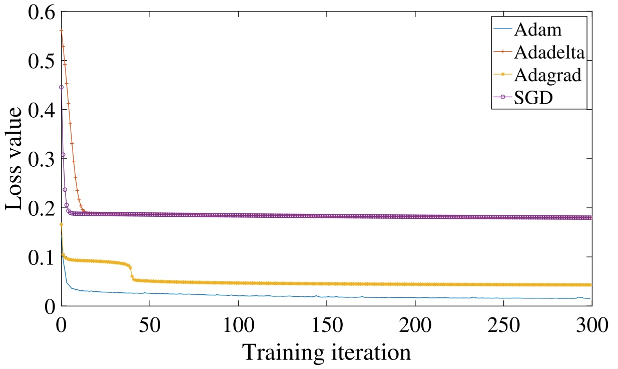

Fig. 10. Loss values for different optimizers in the Alabama dataset.

Table 2
Predicted results with VBAED given different m in the BTH dataset.

m (p)	RMSE	MAE	R^2
16	0.1214	0.0789	0.9849
32	0.0637	0.0425	0.9959
64	**0.0602**	**0.0404**	**0.9963**
128	0.0608	0.0408	0.9962
256	0.0612	0.0409	0.9962
512	0.0623	0.0425	0.9960

Table 3
Predicted results with VBAED given different m in the Alabama dataset.

m (p)	RMSE	MAE	R^2
16	0.1302	0.0918	0.9850
32	0.1284	0.0914	0.9854
64	**0.1268**	**0.0891**	**0.9858**
128	0.1272	0.0901	0.9857
256	0.1744	0.1165	0.9856
512	0.1527	0.1104	0.9794

Table 4
Parameter setting of VBAED in both BTH and Alabama datasets.

Parameter	Value	Description
T	30	Previous time steps
Optimizer	Adam	Optimizer
m	64	Hidden size of the encoder
p	64	Hidden size of the decoder
Batch size	128	Batch size
Number of epochs	300	Number of iterations

Table 5
Parameter settings of benchmark methods.

Methods	Parameter setting for the Alabama dataset	Parameter setting for the BTH dataset
ARIMA	$p_a = 5,\ d_a = 1,\ q_a = 2$	$p_a = 5,\ d_a = 1,\ q_a = 2$
SVR	$T = 30,\ \epsilon = 0.01$	$T = 30,\ \epsilon = 0.01$
XGBoost	$n\_estimators = 300$, $max\_depth = 6$	$n\_estimators = 300$, $max\_depth = 6$
BP	$m = 64,\ T = 30$	$m = 64,\ T = 30$
LSTM	$m = 32,\ T = 30$	$m = 32,\ T = 30$
BiLSTM	$m = 32,\ T = 30$	$m = 64,\ T = 30$
DA-RNN	N/A	$m = p = 128,\ T = 60$
VMD-LSTM	$m = 64,\ T = 30$	$m = 32,\ T = 30$
VMD-BiLSTM	$m = 64,\ T = 30$	$m = 32,\ T = 30$
VMD-DA-RNN	$m = p = 64,\ T = 30$	$m = p = 128,\ T = 60$

that of lagged forecast errors. For SVR, ϵ denotes the acceptable error margin. For XGBoost, *n_estimators* denotes the number of regression tree base learners required by XGBoost, which controls the number of iterations of the boosting process, and *max_depth* denotes the maximum depth of each tree.

4.6. Prediction results

VBAED is trained with the training set and the prediction results are obtained. It is observed in Fig. 11 that for the BTH dataset, the predicted curve and the ground truth one are almost identical, which shows that VBAED is effective in the multi-feature dataset. Fig. 12 shows that for the Alabama dataset, VBAED also works well in the single-feature dataset.

To further verify the robustness and effectiveness of VBAED, we adopt RMSE, MAE, and R^2 to compare it with its other 10 peers as shown in Tables 6 and 7. Hyperparameter tuning and selection for all models are realized by using Taguchi's experimental design method (Gao et al., 2019). It identifies a subset of possible parameter combinations rather than all combinations, thereby reducing the number of

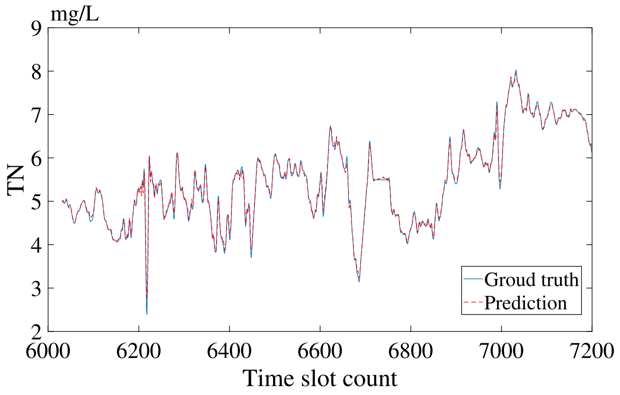

Fig. 11. Prediction results of the water quality time series in the BTH dataset with VBAED.

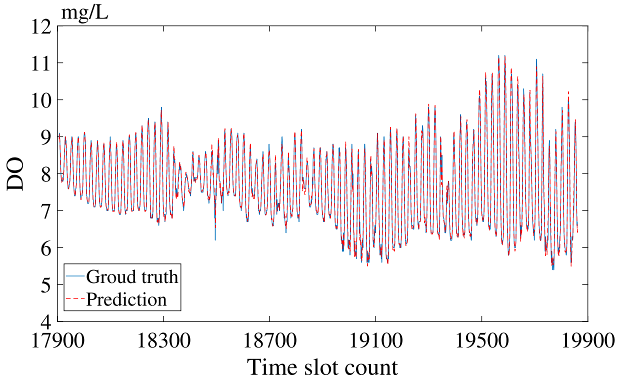

Fig. 12. Prediction results of the water quality time series in the Alabama dataset with VBAED.

Table 6
Performance comparison of different methods in the BTH dataset.

Methods	Evaluation metrics		
	RMSE	MAE	R^2
ARIMA	0.2335	0.1621	0.9373
SVR	0.2293	0.1402	0.9464
XGBoost	0.2684	0.1803	0.9267
BP	0.2487	0.1578	0.9370
LSTM	0.2093	0.1552	0.9552
BiLSTM	0.1657	0.1202	0.9719
DA-RNN	0.1295	0.0868	0.9831
VMD-LSTM	0.1688	0.1363	0.9708
VMD-BiLSTM	0.1475	0.1132	0.9777
VMD-DA-RNN	0.1156	0.0853	0.9840
VBAED	**0.0602**	**0.0404**	**0.9963**

Table 7
Performance comparison of different methods in the Alabama dataset.

Methods	Evaluation metrics		
	RMSE	MAE	R^2
ARIMA	0.2301	0.1683	0.9311
SVR	0.2287	0.1579	0.9411
XGBoost	0.2216	0.1563	0.9491
BP	0.2140	0.1558	0.9512
LSTM	0.1957	0.1414	0.9662
BiLSTM	0.1866	0.1371	0.9692
VMD-LSTM	0.1902	0.1401	0.9671
VMD-BiLSTM	0.1724	0.1232	0.9721
VMD-DA-RNN	0.1555	0.1085	0.9786
VBAED	**0.1268**	**0.0891**	**0.9858**

Table 8
RMSE results of cross validation in the BTH dataset.

Methods	Round				
	1st Round	2nd Round	3rd Round	4th Round	Average
LSTM	0.2995	0.2054	0.1835	0.2093	0.2244
BiLSTM	0.2561	0.1691	0.1436	0.1657	0.1836
DA-RNN	0.2162	0.1211	0.1041	0.1295	0.1427
VMD-LSTM	0.2571	0.1654	0.1483	0.1688	0.1849
VMD-BiLSTM	0.2317	0.1401	0.1259	0.1475	0.1613
VMD-DA-RNN	0.2027	0.1037	0.0939	0.1156	0.1289
VBAED	**0.1523**	**0.0546**	**0.0412**	**0.0602**	**0.0770**

Table 9
RMSE results of cross validation in the Alabama dataset.

Methods	Round				
	1st Round	2nd Round	3rd Round	4th Round	Average
LSTM	0.2643	0.5324	0.2483	0.2316	0.3191
BiLSTM	0.2458	0.4255	0.2277	0.2056	0.27615
VMD-LSTM	0.2539	0.3458	0.2189	0.1983	0.2542
VMD-BiLSTM	0.2385	0.3231	0.2031	0.1734	0.2345
VMD-DA-RNN	0.2152	0.3015	0.1974	0.1519	0.2165
VBAED	**0.1691**	**0.3004**	**0.1706**	**0.1366**	**0.1941**

experiments during execution and providing the best parameter estimation. In particular, DA-RNN can only be used for the multi-feature dataset, and therefore, DA-RNN is removed from the comparison experiment of the Alabama dataset. It is shown that VBAED obtains the best results in both the BTH and Alabama datasets. In addition, in the BTH dataset, when the VMD decomposition is not adopted, RMSEs of LSTM, BiLSTM and DA-RNN are 0.2093, 0.1657 and 0.1259, respectively. After adopting it, RMSEs of VMD-LSTM, VMD-BiLSTM and VDM-DA-RNN are 0.1688, 0.1475, 0.1156, respectively. In the Alabama dataset, when the VMD decomposition is not adopted, RMSEs of LSTM and BiLSTM are 0.1957 and 0.1866, respectively. After adopting it, RMSEs of VMD-LSTM and VMD-BiLSTM are 0.1724 and 0.1555, respectively. It is shown that VMD effectively grasps the evolution trend of the water quality data and decomposes it into key information modes and noise ones, which is helpful to model training and improves the prediction accuracy. It is also observed that RMSE of LSTM is worse than that of BiLSTM on both the BTH and Alabama datasets, which suggests that the bidirectional LSTM structure overcomes the limitation of traditional LSTM, which tends to ignore the information from back to front, resulting in the loss of correlation information.

To more reliably validate the performance of VBAED on the BTH and Alabama datasets, this work further conducts cross-validation on models that exhibit superior performance.

Since using the future data to predict the past data is meaningless in the time series prediction, we adopt the rolling cross-validation. We divide the dataset into six equal parts sequentially, labeled as P_1, P_2, P_3, P_4, P_5, and P_6. We conduct four rounds of cross-validation. In the first round, the training set consists of $\{P_1, P_2\}$, and the test set is P3. In the second round, the training set includes $\{P_1, P_2, P_3\}$, and the test set is P4. For the third round, the training set is $\{P_1, P_2, P_3, P_4\}$, and the test set is P5. In the fourth round, the training set is $\{P_1, P_2, P_3, P_4, P_5\}$, and the test set is P6. In the end, we calculate the average RMSE across all rounds of the model to further validate its performance. Tables 8 and 9 present the RMSE results of cross-validation for both the BTH and Alabama datasets.

Tables 8 and 9 demonstrate that in all rounds, VMD consistently improves the performance of the original model. This indicates that VMD effectively decomposes the original data, assisting the model in more effective feature extraction. VBAED achieves the best performance in all rounds and in the average results, providing evidence of its strong generalization capability across different test sets.

To further verify the effect of the bidirectional input attention mechanism and the bidirectional temporal attention one, ablation experiments are conducted on two datasets. Tables 10 and 11 show that

Table 10
Ablation experiment given the BTH dataset.

Adopted component	Evaluation metrics		
	RMSE	MAE	R^2
Bidirectional input attention	0.0705	0.0480	0.9949
Bidirectional temporal attention	0.0768	0.0524	0.9940
Complete VBAED	**0.0602**	**0.0404**	**0.9963**

Table 11
Ablation experiment given the Alabama dataset.

Adopted component	Evaluation metrics		
	RMSE	MAE	R^2
Bidirectional input attention	0.1313	0.0936	0.9848
Bidirectional temporal attention	0.1386	0.0986	0.9830
Complete VBAED	**0.1268**	**0.0891**	**0.9858**

the only application of the bidirectional input attention mechanism or the bidirectional time attention one results in significant decrease in the prediction accuracy. It is observed that RMSEs of the model that adopts the bidirectional input attention are 0.0705 and 0.1313, respectively, while RMSEs of the model that adopts bidirectional temporal attention are 0.0768 and 0.1368 given BTH and Alabama datasets, respectively. This demonstrates that the bidirectional input attention mechanism plays a more important role in VBAED than the bidirectional temporal attention mechanism. For the original long-sequence data, it is difficult for the network to directly capture important information. The bidirectional input attention mechanism enables VBAED to distinguish the importance of original features, which strengthens important ones and weakens unimportant ones. In addition, it enables the encoder in VBAED to obtain more useful information.

VBAED adopts the bidirectional input attention mechanism to extract relevant features, and the bidirectional temporal one to select relevant hidden states across all time steps. Thus, VBAED achieves the highest prediction accuracy among all methods in both the BTH and Alabama datasets.

5. Conclusions and future work

This work aims at implementing accurate water indicator prediction for real-world water quality data. To achieve it, this work adopts variational mode decomposition (VMD) to decompose the water quality data and combines a bidirectional input attention mechanism with bidirectional long short-term memory (BiLSTM) as an encoder to extract the hidden information. The bidirectional temporal attention mechanism is combined with BiLSTM as a decoder to capture long-term dependency. To demonstrate the effectiveness of the proposed VBEAD, two experiments based on real-world water quality data are conducted. The experimental results support the following conclusions: (1) VMD effectively reduces the negative impact of mode mixing on prediction and decreases the nonlinearity of time series; (2) The bidirectional input attention mechanism adaptively selects relevant features; (3) The bidirectional temporal attention mechanism adaptively selects important hidden states in the time dimension; (4) and BiLSTM captures long-term dependency and hidden information in two directions and outperforms other baseline methods.

In the future work, VBAED can be further extended to solve other different problems, such as wind speed prediction (Gao et al., 2021). In addition, since missing values of water quality data occur frequently and have negative impact on prediction, we plan to further explore more effective data completion methods (Pan et al., 2021) to complement missing values.

CRediT authorship contribution statement

Jing Bi: Conceptualization, Supervision, Funding acquisition, Writing – original draft. **Zexian Chen:** Formal analysis, Methodology, Validation, Data curation, Software. **Haitao Yuan:** Resources, Project administration, Visualization, Investigation. **Jia Zhang:** Investigation, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Haitao Yuan reports financial support was provided by National Natural Science Foundation of China.

Data availability

Data will be made available on request.

References

Bae, K. Y., Jang, H. S., & Sung, D. K. (2017). Hourly solar irradiance prediction based on support vector machine and its error analysis. *IEEE Transactions on Power Systems*, *32*, 935–945.

Baigang, D., Qiliang, Z., & Jun, G. (2021). Deep learning with long short-term memory neural networks combining wavelet transform and principal component analysis for daily urban water demand forecasting. *Expert Systems with Applications*, *171*, Article 114571.

Bandara, K., Bergmeir, C., & Hewamalage, H. (2021). LSTM-MSNet: Leveraging forecasts on sets of related time series with multiple seasonal patterns. *IEEE Transactions on Neural Networks and Learning Systems*, *32*, 1586–1599.

Bi, J., Chen, Z., Yuan, H., Lin, Y., & Qiao, J. (2022). Hybrid prediction for water quality with bidirectional LSTM and temporal attention. In *Proc. international conference on systems, man, and cybernetics* (pp. 1–6).

Bi, J., Lin, Y., Dong, Q., Yuan, H., & Zhou, M. (2020). An improved attention-based LSTM for multi-step dissolved oxygen prediction in water environment. In *Proc. 2020 IEEE int. conf. on networking, sensing and control* (pp. 1–6).

Bi, J., Lin, Y., Dong, Q., Yuan, H., & Zhou, M. (2021). Large-scale water quality prediction with integrated deep neural network. *Information Sciences*, *571*, 191–205.

Box, G. E. P., & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, *65*, 1509–1526.

Buhan, S., & Çadırcı, I. (2015). Multistage wind-electric power forecast by using a combination of advanced statistical methods. *IEEE Transactions on Industrial Informatics*, *11*, 1231–1242.

Chang, F. J., Tsai, Y. H., Chen, P. A., Coynel, A., & Vachaud, G. (2015). Modeling water quality in an urban river using hydrological factors data driven approaches. *Journal of Environmental Management*, *151*, 87–96.

Ding, C., Duan, J., Zhang, Y., Wu, X., & Yu, G. (2019). Using an ARIMA-GARCH modeling approach to improve subway short-term ridership forecasting accounting for dynamic volatility. *IEEE Transactions on Intelligent Transportation Systems*, *19*, 1054–1064.

Dong, Q., Lin, Y., Bi, J., & Yuan, H. (2019). An integrated deep neural network approach for large-scale water quality time series prediction. In *Proc. IEEE int. conf. on systems, man and cybernetics* (pp. 3537–3542).

Dragomiretskiy, K., & Zosso, D. (2022). Variational mode decomposition. *IEEE Transactions on Signal Processing*, *62*, 531–544.

Eseye, A. T., & Lehtonen, M. (2020). Short-term forecasting of heat demand of buildings for efficient and optimal energy management based on integrated machine learning models. *IEEE Transactions on Industrial Informatics*, *16*, 7743–7755.

Fortino, G., Savaglio, C., Spezzano, G., & Zhou, M. (2021). Internet of things as system of systems: A review of methodologies, frameworks, platforms, and tools. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *51*, 223–236.

Gao, C., Zhang, N., Li, Y., Lin, Y., Cheng, & Wan, H. (2023). Adversarial self-attentive time-variant neural networks for multi-step time series forecasting. *Expert Systems with Applications*, *231*, Article 120722.

Gao, S., Zhou, Y., Cheng, J., Yachi, H., & Wang, J. (2019). Dendritic neuron model with effective learning algorithms for classification, approximation, and prediction. *IEEE Transactions on Neural Networks and Learning Systems*, *30*, 601–614.

Gao, S., Zhou, M., Wang, Z., Sugiyama, D., Cheng, J., Wang, J., & Todo, Y. (2021). Fully complex-valued dendritic neuron model. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

Guo, J., He, H., & Sun, C. (2019). ARIMA-based road gradient and vehicle velocity prediction for hybrid electric vehicle energy management. *IEEE Transactions on Vehicular Technology*, *68*, 5309–5320.

Guo, Y., Zhang, S., Yang, J., Yu, G., & Wang, Y. (2022). Dual memory scale network for multi-step time series forecasting in thermal environment of aquaculture facility: A case study of recirculating aquaculture water temperature. *Expert Systems with Applications*, *208*, Article 118218.

Hochreiter, S., & Schmidhuber, J. (2019). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Hou, X., Wang, K., Zhong, C., & Wei, Z. (2021). ST-trader: A spatial-temporal deep neural network for modeling stock market movement. *IEEE/CAA Journal of Automatica Sinica*, *8*, 1015–1024.

Hsu, C., Lu, Y., & Yan, J. (2022). Temporal convolution-based long-short term memory network with attention mechanism for remaining useful life prediction. *IEEE Transactions on Semiconductor Manufacturing*, *35*, 220–228.

Imran, S., Mahmood, T., Morshed, A., & Sellis, T. (2021). Big data analytics in healthcare - a systematic literature review and roadmap for practical implementation. *IEEE/CAA Journal of Automatica Sinica*, *8*, 1–22.

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. In *Proc. of the 3rd international conference for learning representations* (pp. 1–15).

Kong, W., Dong, Z. Y., Jia, Y., Hill, D. J., Xu, Y., & Zhang, Y. (2019). Short-term residential load forecasting based on LSTM recurrent neural network. *IEEE Transactions on Smart Grid*, *10*, 841–851.

Liu, P., Fu, B., Yang, S. X., Deng, L., Zhong, X., & Zheng, H. (2021). Optimizing survival analysis of xgboost for ties to predict disease progression of breast cancer. *IEEE Transactions on Biomedical Engineering*, *68*, 148–160.

Liu, P., Wang, J., Sangaiah, A. K., Xie, Y., & Yin, X. (2019). Analysis and prediction of water quality using LSTM deep neural networks in IoT environment. *Sustainability*, *11*, 1–14.

Lu, Y., Panneerselvam, J., Liu, L., & Wu, Y. (2016). RVLBPNN: A workload forecasting model for smart cloud computing. *Scientific Programming*, *2016*, 1–9.

Moeeni, H., Bonakdari, H., & Fatemi, S. E. (2017). Stochastic model stationarization by eliminating the periodic term and its effect on time series prediction. *Journal of Hydrology*, *547*, 348–364.

Najah, A., El-Shafie, A., Karim, O. A., Jaafar, O., & El-Shafie, Amr H. (2011). An application of diferent artificial intelligences techniques for water quality prediction. *International Journal of Physical Sciences*, *6*, 5298–5308.

Pan, J., Li, C., Tang, Y., Li, W., & Li, X. (2021). Energy consumption prediction of a CNC machining process with incomplete data. *IEEE/CAA Journal of Automatica Sinica*, *8*, 987–1000.

Principi, E., Rossetti, D., Squartini, S., & Piazza, F. (2019). Unsupervised electric motor fault detection by using deep autoencoders. *IEEE/CAA Journal of Automatica Sinica*, *6*, 441–451.

Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., & Cottrell, G. (2017). A dual-stage attention-based recurrent neural network for time series prediction. In *International joint conference on artificial intelligence* (pp. 1–7).

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, *36*, 1627–1639.

Sharma, R. R., Kumar, M., Maheshwari, S., & Ray, K. P. (2021). EVDHM-ARIMA-based time series forecasting model and its application for COVID-19 cases. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–10.

Sun, Z., Zhao, S., & Zhang, J. (2019). Short-term wind power forecasting on multiple scales using VMD decomposition, K-means clustering and LSTM principal computing. *IEEE Access*, *7*, 166917–166929.

Wang, J., & Li, Y. (2018). Multi-step ahead wind speed prediction based on optimal feature extraction long short term memory neural network and error correction strategy. *Applied Energy*, *230*, 429–443.

Wang, Z., Su, X., & Ding, Z. (2021). Long-term traffic prediction based on LSTM encoder-decoder architecture. *IEEE Transactions on Intelligent Transportation Systems*, *22*, 6561–6571.

Wu, D., Wang, H., Mohammed, H., & Seidu, R. (2020). Quality risk analysis for sustainable smart water supply using data perception. *IEEE Transactions on Sustainable Computing*, *5*, 377–388.

Xia, M., Shao, H., Ma, X., & de Silva, C. W. (2021). A stacked GRU-RNN-based approach for predicting renewable energy and electricity load for smart grid operation. *IEEE Transactions on Industrial Informatics*, *17*, 7050–7059.

Xie, Y., Liu, G., Yan, C., Jiang, C., & Zhou, M. (2022). Time-aware attention-based gated network for credit card fraud detection by extracting transactional behaviors. *IEEE Transactions on Computational Social Systems*, 1–13.

Yang, L., He, M., Zhang, J., & Vittal, V. (2015). Support-vector-machine-enhanced Markov model for short-term wind power forecast. *IEEE Transactions on Sustainable Energy*, *6*, 791–799.

Yang, Z., Yan, W., Huang, X., & Mei, L. (2022). Adaptive temporal-frequency network for time-series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, *34*, 1576–1587.

Yule, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society B Biological Sciences*, *226*, 267–298.

Zhang, X., Mohanty, S. N., Parida, A. K., Pani, S. K., Dong, B., & Cheng, X. (2020). Annual and non-monsoon rainfall prediction modelling using SVR-MLP: An empirical study from odisha. *IEEE Access*, *8*, 30223–30233.

Zhang, J., Zhu, Y., Zhang, X., Ye, M., & Yang, J. (2019). Developing a long short-term memory (LSTM) based model for predicting water table depth in agricultural areas. *Journal of Hydrology*, *561*, 918–929.

Zheng, H., Lin, F., Feng, X., & Chen, Y. (2021). A hybrid deep learning model with attention-based conv-LSTM networks for short-term traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, *22*, 6910–6920.

Zhou, J., Ding, D., Wu, Z., & Xiu, Y. (2023). Spatial context-aware time-series forecasting for QoS prediction. *IEEE Transactions on Network and Service Management*, *20*, 918–931.

Zhou, B., Ma, X., Luo, Y., & Yang, D. (2019). Wind power prediction based on LSTM networks and nonparametric kernel density estimation. *IEEE Access*, *7*, 165279–165292.

Zou, M., Holjevac, N., Daković, J., Kuzle, I., Langella, R., Giorgio, V. D., & Djokic, S. Z. (2022). Bayesian CNN-BiLSTM and vine-GMCM based probabilistic forecasting of hour-ahead wind farm power outputs. *IEEE Transactions on Sustainable Energy*, *13*, 1169–1187.

Jing Bi is currently an Associate Professor with the Faculty of Information Technology, School of Software Engineering, Beijing University of Technology, Beijing, China. She has over 80 publications including journal and conference papers. Her research interests include distributed computing, cloud computing, large-scale data analysis, machine learning and performance optimization. Dr. Bi was the recipient of the IBM Fellowship Award and the recipient of the Best Paper Award-Finalist in the 16th IEEE International Conference on Networking, Sensing and Control. She is now an Associate Editor of IEEE ACCESS. She is a senior member of the IEEE.

Zexian Chen is currently a Master student in the Faculty of Information Technology, School of Software Engineering, Beijing University of Technology, Beijing, China. Before that, he received his B.E. degree in Water Supply & Sewerage Science and Project from Beijing University of Technology in 2021. His research interests include time series prediction, intelligent optimization algorithms and machine learning.

Haitao Yuan received the Ph.D. degree in Computer Engineering from New Jersey Institute of Technology (NJIT), Newark, NJ, USA in 2020. He is currently an Associate Professor with the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. His research interests include cloud computing, edge computing, data centers, big data, machine learning, deep learning and optimization algorithms. He received the Chinese Government Award for Outstanding Self-Financed Students Abroad, the 2021 Hashimoto Prize from NJIT, and the Best Paper Award in the 17th ICNSC.

Jia Zhang received the Ph.D. degree in computer science from the University of Illinois at Chicago. She is currently the Cruse C. and Marjorie F. Calahan Centennial Chair in Engineering, Professor of Department of Computer Science in the Lyle School of Engineering at Southern Methodist University. Her research interests emphasize the application of machine learning and information retrieval methods to tackle data science infrastructure problems, with a recent focus on scientific workflows, provenance mining, software discovery, knowledge graph, and interdisciplinary applications of all of these interests in earth science. She is a senior member of the IEEE.