Energy 252 (2022) 123990

Contents lists available at ScienceDirect

Energy

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

Interpretable wind speed prediction with multivariate time series and temporal fusion transformers

Binrong Wu [a], Lin Wang [a], Yu-Rong Zeng [b, *]

[a] School of Management, Huazhong University of Science and Technology, Wuhan, 430074, China
[b] School of Information and Communication Engineering, Hubei University of Economics, Wuhan, 430205, China

ARTICLE INFO

Article history:
Received 19 January 2022
Received in revised form
2 April 2022
Accepted 10 April 2022
Available online 14 April 2022

Keywords:
Wind speed forecasting
Interpretable forecasting
Deep learning
Multisource data
Variational mode decomposition

ABSTRACT

Wind power has been utilized well in power systems, so steady and successful wind speed forecasting is crucial to security management power grid market economy. To date, most researchers have often discounted the interpretability of prediction models, leading to obscure forecasts. This study puts forward a unique forecasting methodology that incorporates notable decomposition techniques, multifactor interpretable forecasting models, and optimization algorithms. In the proposed model, variational mode decomposition is employed to break down the raw wind speed sequence into a set of intrinsic mode functions. Adaptive differential evolution is then used for optimizing several parameters of temporal fusion transformers (TFT) to achieve satisfactory forecasting performance. TFT is a new attention-based deep learning model that puts together high-performance multi-horizon prediction and interpretable insights into temporal dynamics. Empirical studies using eight real-world 1-h wind speed data sets in Albert, Canada, and Five Points, USA demonstrate that the system using the proposed model outperforms those employing other comparable models in nearly all performance metrics. Examples of TFT's interpretable outputs are the importance ranking of the decomposed wind speed sub-sequences and meteorological data and attention analysis of different step lengths. The findings signify substantial progress for wind speed prediction and aid policymakers.

1. Introduction

1.1. Motivation and incitement

The application and advancement of clean energy are receiving much attention from governments around the globe owing to the worldwide scarcity of energy and policy calls of different countries to save energy, reduce emissions, and protect the environment [1]. Wind energy is a standard example of renewable resources. Given its potential as a sustainable energy supply, wind energy has been widely utilized in production. Wind power generation maintains rapid growth at the global level. The Global Wind Report 2021 cited that the total installed capacity of wind power in 2020 reached 743 GW, representing a growth of 14.3% compared with that in 2019. Wind power generation is largely influenced by wind speed, and wind speed fluctuations possess the inherent nonlinear characteristics of stochasticity and intermittency. As such, the output power of wind power generation fluctuates considerably, leading to difficulties integrating wind power into a grid [2,3]. Wind speed prediction should be unfailing at all times. Thus, the demand for improving forecast accuracy and expanding forecast horizons is serious. This study aims to produce a more advanced and more robust wind speed forecasting model that is easier to implement than most existing ones. This study fills the gap in the research on the interpretability of multi-factor wind speed forecasting models.

1.2. Review of literature on wind speed prediction

Wind speed prediction can be divided into four categories based on timescale: ultra-short-term (a few seconds to 30 min), short term (30 min–6 h), medium-term (6 h to one day), and long term (more than one day) [4]. The functions of wind speed predictions of different timescales vary. Ultra-short-term wind speed prediction is employed for real-time load tracking and wind turbine control; short-term, load dispatch planning; medium-term, energy trading

* Corresponding author.
E-mail addresses: binronghust@foxmail.com (B. Wu), wanglin@hust.edu.cn (L. Wang), zyr@hbue.edu.cn (Y.-R. Zeng).

https://doi.org/10.1016/j.energy.2022.123990
0360-5442/

Nomenclature

ADE	Adaptive differential evolution
AI	Artificial intelligence
ARIMA	Autoregressive integrate moving average
BPNN	Backpropagation neural networks
CNN	Convolutional neural network
DE	Differential evolution
EEMD	Ensemble empirical mode decomposition
EMD	Empirical mode decomposition
GA	Genetic algorithm
GLUs	Gated linear units
GRN	Gated residual network
GRNN	Generalized regression neural network
GRU	Gated recurrent unit
GW	Gigawatt
LSTM	Long short-term memory
MAE	Mean absolute error
MAPE	Mean absolute percentage error
RMSE	Root mean square error
RNN	Recurrent neural network
SSA	Singular spectrum analysis
SVM	Support vector machines
TFT	Temporal fusion transformers
VMD	Variational mode decomposition
WNN	Wavelet neural network
WT	Wavelet transform

and power system management; long-term, informing the best equipment maintenance plan [5].

Wind speed forecasting techniques have been proposed and applied [6]. According to the calculation mechanism of calculation, wind speed forecasting systems can be grouped into physical, statistical, spatiotemporal correlation, and artificial intelligence (AI) models [7]. The summary of selected wind speed forecasting studies is shown in Table 1.

Physical models predict wind speed by looking at weather factors (e.g., air pressure, humidity, and temperature) [8,9]. The prediction of physical models is more accurate in the medium- and long terms. However, the lack of the trend characteristics of historical data causes such models to perform poorly in the short term [10].

Statistical models include auto-regressive integrated moving average (ARIMA) [11], fractional ARIMA (f-ARIMA) [12], and Hammerstein auto-regressive models [13]. Statistical models are time-series models that can extract the linear characteristics of historical wind speed. They perform well in ultra-short-term and short-term wind speed prediction.

Spatiotemporal correlation models predict wind speed by using the spatiotemporal correlation between the target and neighboring stations. The predictions of spatiotemporal models are more reliable, but the accumulation of multi-regional data increases the prediction complexity and application cost [14].

With the fast advancement of AI technology, various AI models for predicting wind speed have been introduced. Examples are

Table 1
The summary of selected wind speed forecasting studies.

Classification	Forecasted areas	Time interval	Input variables	Forecasting Methods
Physical models	Eastern Liguria (Italy)	3 h	Physical processes in the atmosphere	Kalman filters [8]
	UK	1 h	Data partitioned on atmospheric stability class	Numerical weather prediction model and Gaussian process regression (GPR) model [9]
Spatio-temporal correlation models	Texas (USA)	5 min	Meteorological factors and historical wind speed of various sites	Combination of convolutional neural network and long short-term memory neural network [14]
Statistical models	Baltic Sea area	10 min	Historical wind speed	Auto-regressive integrated moving average (ARIMA) [11]
	Illinois and New Jersey	1 h	Historical wind speed	Hammerstein Auto-Regressive models [13]
	North Dakota	1 h	Historical wind speed	Fractional-ARIMA [12]
Artificial intelligence models	Penglai (China)	10 min and 30 min	Historical wind speed	Elman neural network [3]
	Sotavento Galicia and Beresford	10 min and 1 h	Historical wind speed	Kernel extreme learning machines [15]
	Chinese Qinghai Wind Farm	30 min	Historical wind speed	Artificial neural network [16]
	Fuzhou and Haikou (China)	1 day	Meteorological factors and historical wind speed	Gated recurrent unit [18]
	The United States	10 min	Meteorological factors and historical wind speed	Regularized extreme learning machine [31]
Hybrid models	Hebei wind farm and Inner Mongolia wind farm (China)	1 h	Historical wind speed	An adaptive hybrid model based on variational mode decomposition, fruit fly optimization algorithm, ARIMA, and deep belief network [22]
	Rio Grande do Norte, Northeast of Brazil	10 min	Historical wind speed	An adaptive hybrid model based on the SSA and adaptive neuro-fuzzy inference system [24]
	Inner Mongolia, China	10 min and 1 h	Historical wind speed	A hybrid model combined LSTM, hysteretic extreme learning machine, differential evolution algorithm [28]
	Sweden	10 min and 1 h	Historical wind speed	A hybrid model based on variational mode decomposition method, generalized normal distribution algorithm, and Bi-LSTM [29]
	A wind farm at Parazinho city, Brazil	10 min	Historical wind speed	A decomposition-ensemble learning method that combines stacking-ensemble learning and complete ensemble empirical mode decomposition [25]
	China	15 min	Meteorological factors and historical wind speed	A hybrid model used a highway gate algorithm to optimize the configuration of local convolutional neural networks [30]

Elman neural network (ENN) [3], kernel extreme learning machine [15], and artificial neural network [16]. AI-based forecasting models, especially deep learning ones, exhibit strong data adaptability, and they can build complex feature mapping relationships [17]. In many previous studies, the performance of AI-based models is more competitive than that of traditional statistical models. For example, aiming at the input optimization problem of the network, Wu et al. [18] proposed a novel wind speed prediction system based on a gated recurrent unit network. Their experimental results show that optimizing the input of the neural network improves the performance of wind speed prediction and provides a new idea for selecting the input for wind speed prediction.

As the weather system is volatile and unstable, the above four types of models cannot always successfully extract the complex feature correlations in nonlinear and non-stationary wind speed data [19]. To overcome this problem, many wind speed forecasting studies focus on hybrid models. Hybrid models can unify the advantages of multiple individual models and may provide excellent prediction performance for wind speed datasets [20]. Hybrid models can be divided into two categories based on hybrid strategies: stacking-based and weight-based models. In stacking-based forecasting models, the outputs of one or more basic models are used as features, which are then combined with those of a high-level model. Chen et al. [14] predicted wind speed by using the output of the CNN model as the input of the long short term memory (LSTM) model. CNN models can extract spatiotemporal features from raw wind data. Weight-based forecasting models can be constructed by exploiting the diversity of forecasters [2]. Song, Wang, and Lu [21] proposed to integrate four types of AI models (i.e., Elman, BPNN, GRNN, and WNN) to predict wind speed by using combined weights.

Hybrid models are based on data decomposition methods. Singular spectrum analysis (SSA), variational mode decomposition (VMD), empirical mode decomposition (EMD), and ensemble empirical mode decomposition (EEMD) are used extensively in forecasting wind speed [22]. Data decomposition methods help remove the random interference of the wind speed sequence to enhance wind speed prediction performance [23]. Moreno and dos Santos Coelho [24] introduced a hybrid approach combining SSA and an adaptive neuro-fuzzy inference system for wind speed prediction. De Silva et al. [25] presented a new decomposition-ensemble learning method that combines stacking-ensemble learning and complete ensemble empirical mode decomposition to forecast wind speed. Their experimental results reveal that decomposition methods can substantially improve prediction accuracy.

The forecasting algorithm is the central component of a hybrid prediction model. The configuration settings of the model largely influence the performance of different prediction methods, especially deep learning algorithms. Many researchers used intelligent optimization algorithms to determine the satisfactory configuration setting of different wind speed prediction models [26,27]. Hu and Chen [28] employed a differential evolution (DE) algorithm to decide the optimal number of neurons and hidden layers of the LSTM network for wind speed prediction. Neshat et al. [29] applied the generalized normal distribution algorithm to optimize the configuration of Bi-LSTM. Dong et al. [30] proposed a novel hybrid supervised approach that uses a highway gate algorithm to optimize the configuration of local convolutional neural networks. The experimental results show that using intelligent algorithms to optimize the parameter configuration of prediction models considerably improves their prediction performance.

Existing hybrid wind speed forecasting systems are constrained to the decomposition techniques of historical wind speed data. The coupling relationship between wind speed and other meteorological variables (e.g., air pressure, air temperature, and relative humidity) in the time-frequency domain is overlooked. Studies on wind speed prediction models that are based on multiple weather variables are scant. For example, Shang et al. [31] proposed a combined wind speed forecasting system that considers historical wind speed and weather factors and uses a self-organizing map to cluster samples according to weather factors. Their experimental results show that weather data can significantly improve wind speed prediction performance. However, there is currently little literature on how different meteorological factors contribute to wind speed forecasting. In other words, research on the interpretability of wind speed forecasting is inadequate currently. Compared with previous studies, this study uses an interpretable model to study the coupling relationship between meteorological data and wind speed and gives interpretable results.

1.3. Contributions

This study puts forward a unique system based on VMD, adaptive DE (ADE), and temporal fusion transformer (TFT) algorithms to realize correct and interpretable wind speed prediction. This study considers historical wind speed series and multiple meteorological variables simultaneously and distinguishes their importance to wind speed forecasting. VMD decomposes raw wind speed data into multiple band-limited intrinsic mode functions. For better prediction results, six key parameters of TFT are optimized by applying ADE. The wind speed sub-modes decomposed by VMD are used as historical input. Meteorological data and time indicators (such as “month,” “day,” and “hour”) are then input into the TFT model as known future input. Meteorological data are often input into prediction models as known inputs gave the availability of weather forecasts. To ensure that the VMD-ADE-TFT model is feasible and efficient, eight hourly wind speed data sets in two regions are applied for prediction. The results of the proposed model are compared with those of ten comparable models. The main contributions of this study are as follows:

(a) This study is one of the first attempts to employ TFT for multivariate and interpretable high-performance wind speed prediction. Current deep learning methods often have difficulties in providing interpretability of results. It either fails to provide interpretable results, leads to a decrease in prediction accuracy, or increases the computational effort. Without compromising forecast accuracy and increasing the computational effort, TFT provides two valuable interpretive results to identify important variables for wind speed forecasting and to analyze persistent temporal patterns by paying attention to different lag orders.
(b) This study introduces a wind speed forecasting model named VMD-ADE-TFT. In this model, VMD decomposes raw wind speed data into multiple relatively stable subsequences to minimize the difficulty of wind speed prediction. The TFT parameters improved by ADE are then employed to enhance the prediction performance and stability of the model. The experimental results demonstrate that the proposed model generally outperforms other comparable models, confirming its stability and reliability.
(c) This study assesses the importance of decomposed wind speed subsequences and meteorological data through the interpretability capability of TFT. As a result, it identifies important variables in wind speed forecasting, providing convincing wind speed forecast analysis and decision support for decision-makers.

The rest of the paper is outlined as follows: Sections 2 and 3 describe the basic theories and proposed forecasting model, respectively. Section 4 reports the experiments and results. Section 5 draws the conclusions.

2. Methodology

2.1. Variational mode decomposition

VMD is an adaptive, non-recursive signal decomposition technology that decomposes non-stationary signals into a series of sub-patterns. Its advantages include fast and simple optimization when dealing with data noise and robust decomposition series [32].

VMD aims to decompose raw signal $f(t)$ into multiple sub-modes (u_k, $k = 1, 2, \ldots, K$), each with a center frequency ω_k. The constraint condition of VMD is that the sum of each mode should be equal to the raw signal. The minimum sum of the frequency bandwidth of each sub-mode is regarded as the objective function. The mathematical expression is given as follows:

$$\underset{\{u_k\}\{\omega_k\}}{\text{minimum}}\left\{\sum_{k=1}^{K}\partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)*u_k(t)\right]e^{-j\omega_k t}{}_2^2\right\}$$

$$s.t.\sum_{k=1}^{K}u_k(t)=f(t) \tag{1}$$

where $\delta(t)$ denotes the Dirac distribution. To derive the optimal solution to the above problem, the Lagrangian multiplier $\lambda(t)$ and quadratic penalty term *alpha* are applied to transform the problem into an unconstrained problem. To ensure that the constrained problem is equivalent to the unconstrained problem, $\lambda(t)$ is used. The *alpha* can guarantee that the sub-model can be reconstructed accurately when Gaussian noise exists. The unconstrained problem is expressed as follows:

$$\begin{aligned}\mathrm{L}(\{\mathrm{u_k}\},\{\omega_k\},\lambda)=\mathrm{alpha}\sum_{k=1}^{K}\partial_t\Big[&\\ \times\left(\delta(t)+\frac{j}{\pi t}\right)*\mathrm{u}_k(t)\Big]\mathrm{e}^{-j\omega_k t}{}_2^2+\mathrm{f}(t)-\sum_{k=1}^{K}\mathrm{u}_k(t)_2^2+\lambda(t),\mathrm{f}(t)&\\ -\sum_{k=1}^{K}\mathrm{u}_k(t)&\end{aligned} \tag{2}$$

The saddle point of the Lagrangian function is derived by iterating u_k^{n+1},ω_k^{n+1} and λ^{n+1}. Thereafter, $\widehat{u}_k^{n+1}$ and ω_k^{n+1} can be updated by Eqs. (3) and (4), respectively. $\widehat{u}_i(\omega)$, $\widehat{u}_k^{n+1}(\omega)$, $\widehat{\lambda}(\omega)$, and $\widehat{f}(\omega)$ are the Fourier transform of $u_i(t), u_k^{n+1}(t)$, $\lambda(t)$, and $f(t)$, respectively.

$$\widehat{u}_k^{n+1}(\omega)=\frac{\widehat{f}(\omega)-\sum_{i\neq k}\widehat{u}_i(\omega)+\frac{\widehat{\lambda(\omega)}}{2}}{1+2alpha(\omega-\omega_k)^2} \tag{3}$$

$$\omega_k^{n+1}=\frac{\int_0^\infty\omega|\widehat{u}_k(\omega)|^2d\omega}{\int_0^\infty|\widehat{u}_k(\omega)|^2d\omega} \tag{4}$$

With the alternate direction method of multipliers as a basis, $\widehat{\lambda}^{n+1}(\omega)$ can be obtained by Eq. (5), and τ represents the updated parameter. The termination condition of the iteration is given as Eq. (6), in which ε denotes accuracy.

$$\widehat{\lambda}^{n+1}(\omega)=\widehat{\lambda}^{n}(\omega)+\tau\left(\widehat{f}(\omega)-\sum_{k=1}^{K}\widehat{u}_k^{n+1}\left(\omega\right)\right. \tag{5}$$

$$\sum_{k=1}^{K}\frac{\widehat{u}_k^{n+1}-\widehat{u}_{k2}^{n2}}{\widehat{u}_{k2}^{n2}}<\varepsilon \tag{6}$$

The final output of VMD is u_k^{n+1}, which is converted by the real part of $\widehat{u}_k^{n+1}(\omega)$ by using Fourier transform. The ratio of residual energy $\mathrm{r_{res}}$ is applied to determine the appropriate number of VMD outputs. $\mathrm{r_{res}}$ can be described as follows:

$$\mathrm{r_{res}}=\frac{1}{\mathrm{N_s}}\sum_{t=1}^{\mathrm{N_s}}\left|\frac{\mathrm{f}(t)-\sum_{k=1}^{K}\mathrm{u}_k(t)}{\mathrm{f}(t)}\right| \tag{7}$$

where $\mathrm{f}(t)$ represents the raw wind speed data, $\mathrm{N_s}$ is the sample number; K is the number of decomposed sub-mode, and $\mathrm{u_k}(t)$ portrays the decomposed modes. $\mathrm{r_{res}}$ should have no obvious downward trend to ascertain the number of patterns [33].

2.2. Temporal fusion transformers

Transformer-based models have been widely used in time series forecasting [34]. The Transformer mainly includes an encoder and a decoder, where the encoder part takes the historical data of the time series as input, and the decoder part predicts future values in an autoregressive manner. The decoder is connected with the encoder using an attention mechanism. In this way, the decoder can learn to focus on the parts of the time series historical values that are most valuable for predicting the target value before making a prediction. The decoder uses masked self-attention so that the network does not acquire future values during training. However, previous methods often fail to take into account the different types of inputs that are common in time series forecasting, or assume that all exogenous inputs are known in the future. TFT with appropriate inductive biases allows the alignment of architectures with unique data characteristics, thus addressing the above problems.

Proposed by the Google Cloud AI team, TFT is an inherently interpretable deep learning model for forecasting multi-horizon time series [35]. The interpretation ability of TFT is stronger than that of the general black-box models. Black-box models such as neural networks or complex ensemble models usually have high accuracy. However, the internal working mechanism is not easy to understand, and it is impossible to estimate the importance of each input feature to the forecasting results of the model, and it is impossible to understand the interaction between different input features. Fortunately, TFT has the above explanatory. Meanwhile, TFT is the latest approach, which goes beyond previous spatial-temporal methods employing LSTM and CNN.

Fig. 1 illustrates the model architecture of TFT. TFT uses established components to successfully construct feature representations for each input type (i.e., static, known future, and observed inputs), thereby ensuring high prediction performance in various forecasting problems. TFT includes five major constituents, namely, gating mechanisms, variable selection networks, static covariate encoders, temporal processing, and prediction intervals. These elements of TFT are detailed below. The remaining parts of TFT are available in Lim et al. [35].

2.2.1. Gating mechanisms

To enable the model to flexibly apply nonlinear processing

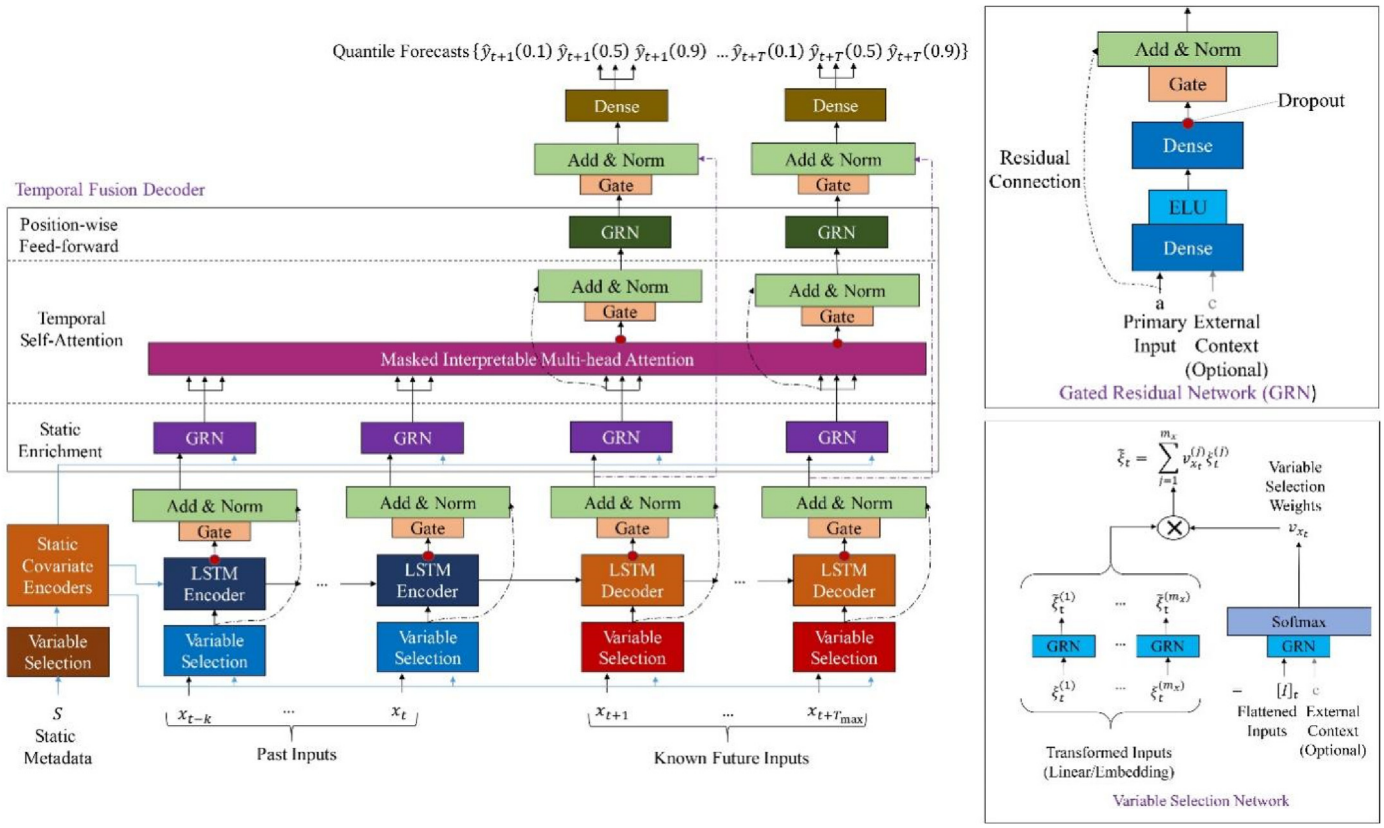

Fig. 1. The model architecture of TFT.

between variables and targets, a gated residual network (GRN) is used. The GRN contains two types of inputs, namely, a primary input a and an optional context vector c. The GRN is described as follows:

$$GRN_{\omega}(a,c) = LayerNorm(a + GLU_{\omega}(\eta_1)), \tag{8}$$

$$\eta_1 = W_{1,\omega}\eta_2 + b_{1,\omega}, \tag{9}$$

$$\eta_2 = ELU(W_{2,\omega}a + W_{3,\omega}c + b_{2,\omega}), \tag{10}$$

where ELU is the activation function of the exponential linear unit, η_1 and $\eta_2 \in \mathbb{R}^{d_{model}}$ denote the intermediate layers, *LayerNorm* is the standard layer normalization, and ω represents weight sharing. Component gating layers based on gated linear units (GLUs) are adopted to provide flexibility for suppressing any part of the architecture that is unnecessary for a given data set. The GLU is described as follows:

$$GLU_{\omega}(\gamma) = \sigma(W_{4,\omega}\gamma + b_{4,\omega}) \odot (W_{5,\omega}\gamma + b_{5,\omega}), \tag{11}$$

where $\gamma \in \mathbb{R}^{d_{model}}$ is the input, $\sigma(.)$ is the sigmoid activation function, $b_{(.)} \in \mathbb{R}^{d_{model}}$ and $W_{(.)} \in \mathbb{R}^{d_{model} \times d_{model}}$ are the biases and weights, d_{model} is the hidden state size, and $\odot$ is the element-wise Hadamard product. The GLU allows TFT to control the degree of contribution of GRN to the original input. If necessary, this layer can be skipped altogether because the outputs of GLUs may all be close to zero to suppress nonlinear contributions.

2.2.2. Variable selection networks

Variable selection networks can offer insights into which variables are the most critical to the prediction problem. They also allow TFT to remove any unnecessary noise input that may curtail forecasting performance. Let $[\mathrm{I}]_t = [\xi_t^{(1)^T}, \ldots, \xi_t^{(m_x)^T}]^T$ represent the flattened vector of all past inputs, with $\xi_t^{(j)}$ the transformed input of the j th variable. As shown in Eq. (12), flattened inputs $[\mathrm{I}]_t$ and external context c_s are inputted into a GRN and then pass through a Softmax layer to derive variable selection weights V_{xt}. Variable selection weights provide explanatory properties for the results of the TFT model. In Eq. (13), each $\xi_t^{(j)}$ is nonlinearly processed by its GRN. Finally, the processed features are weighted by their variable selection weights and combined as shown in Eq. (14).

$$\mathrm{V}_{xt} = Softmax(GRN_{\mathrm{V}_x}([\mathrm{I}]_t, c_s)) \tag{12}$$

$$\tilde{\xi}_t^{(j)} = GRN_{\tilde{\xi}^{(j)}}\left(\xi_t^{(j)}\right) \tag{13}$$

$$\tilde{\xi}_t = \sum_{j=1}^{m_x} v_{xt}^{(j)} \tilde{\xi}_t^{(j)} \tag{14}$$

where $v_{xt}^{(j)}$ denotes the j th element of the vector V_{xt}.

2.2.3. Interpretable multi-head attention

TFT employs a self-attention mechanism that modifies the transformer-based multi-head attention structure to increase interpretability and learn the long-term relationship between different time steps. With the relationships between queries $\mathrm{Q} \in \mathbb{R}^{N \times d_{attn}}$ and keys $\mathrm{K} \in \mathbb{R}^{N \times d_{attn}}$ as a basis, attention mechanisms scale values $\mathrm{V} \in \mathbb{R}^{N \times d_V}$ as follows:

$$\text{Attention}(\text{Q},\ \text{K},\ \text{V}) = \text{A}(\text{Q},\ \text{K})\text{V} \tag{15}$$

where N is the number of time steps inputted to the attention layer, and $A()$ is a normalization function. For attention values, the scaled dot-product is commonly given as follows:

$$\text{A}(\text{Q},\ \text{K}) = \text{Softmax}\left(\text{Q}K^T \Big/ \sqrt{d_{attn}}\right) \tag{16}$$

For the learning capacity of the attention mechanism, multi-head attention is adopted to employ different heads for different representation subspaces.

$$\text{MultiHead}(\text{Q},\ \text{K},\ \text{V}) = [H_1, \ldots, H_{m_H}]W_H \tag{17}$$

$$H_h = Attention\left(QW_Q^{(h)}, KW_K^{(h)}, VW_V^{(h)}\right) \tag{18}$$

$W_Q^{(h)} \in \mathbb{R}^{d_{model} \times d_{attn}}$, $W_K^{(h)} \in \mathbb{R}^{d_{model} \times d_{attn}}$, and $W_V^{(h)} \in \mathbb{R}^{d_{model} \times d_v}$ denote head-specific weights for queries, keys, and values, respectively. $W_H \in \mathbb{R}^{(m_H \cdot d_V) \times d_{model}}$ linearly combines outputs concatenated from all heads H_h.

Each head uses different values, so only attention weight cannot indicate the importance of a particular feature. The multi-head attention is thus changed to the shared value in each head, and the additive aggregation of all heads is employed.

$$\text{InterpretableMultiHead}(\text{Q},\ \text{K},\ \text{V}) = \tilde{H}W_H \tag{19}$$

$$\tilde{H} = \tilde{A}(Q,K)VW_V \tag{20}$$

$$= \left\{\frac{1}{m_H}\sum_{h=1}^{m_H} A\left(QW_Q^{(h)}, KW_K^{(h)}\right)\right\} VW_V \tag{21}$$

$$= \left\{\frac{1}{m_H}\sum_{h=1}^{m_H} Attention\left(QW_Q^{(h)}, KW_K^{(h)}, VW_V\right)\right\} \tag{22}$$

$W_H \in \mathbb{R}^{d_{attn} \times d_{model}}$ is adopted for final linear mapping, and $W_V \in \mathbb{R}^{d_{model} \times d_V}$ denotes the value weights shared across all heads. $\tilde{A}(Q,K)$ successfully enhances presentation ability. Still, simple interpretability studies can be performed by analyzing a set of attention weights.

2.2.4. Quantile outputs and loss functions

TFT produces prediction intervals with point prediction as a basis. TFT simultaneously predicts different percentiles (e.g., 10th, 50th, and 90th) at each time step. Quantile prediction is achieved using the linear transformation output by the temporal fusion decoder.

TFT is trained through joint minimization of quantile loss. The outputs of all quantiles are then added, with the formula given as follows:

$$\mathscr{L}(\Omega,\ \text{W}) = \sum_{y_t \in \Omega} \sum_{q \in \varrho} \sum_{T=1}^{T_{max}} \frac{QL(y_t, \hat{y}(q, t-T, T), q)}{MT_{max}} \tag{23}$$

$$\text{QL}(\text{y}, \hat{y}, \text{q}) = q(y - \hat{y})_+ + (1-q)(\hat{y} - y)_+ \tag{24}$$

where Ω is the domain of training data containing M samples, ϱ is the set of output quantiles, W is the weights of TFT, and $(.)_+$ means $\max(0, .)$.

2.3. Adaptive differential evolution

Basic DE is a stochastic model that replicates biological evolution through repeated iterations. This model can efficiently solve nonlinear and global optimization problems [36]. To balance the global search capability and convergence efficiency of basic DE, ADE is constructed. S-shape adaptive mutation factor is introduced to the DE algorithm. The steps of ADE are detailed below.

Step 1: Population initialization. The following parameters are initialized: population size (M), genetic dimension (D), gene value range $[x_{max}^j, x_{min}^j]$, the maximum number of iterations (T), mutation operator (F), and crossover probability (CR). Eq. (25) is used to derive the initial population.

$$x_{i,t}^j = x_{min}^j + rand(0,1) \cdot \left(x_{max}^j - x_{min}^j\right) \times (i = 1,\ 2,\ \ldots,\ M,\ j = 1,\ 2,\ \ldots,\ D) \tag{25}$$

Step 2: Mutation operation. For each target vector, Eq. (26) is used to generate a corresponding mutation vector. $x_{r1,t}^j$ and $x_{r2,t}^j$ are randomly selected individuals. The mutation operator F establishes the degree of mutation and is expressed as Eq. (27). The value range of F is [0,1]. t is the current number of iterations. In the early stage of the algorithm, a larger mutation factor guarantees the diversity of the population and helps figure out the global optimal solution. In the later stage of the algorithm, a smaller mutation factor can retain excellent individuals, exhibit strong local searchability, and secure the convergence ability of the algorithm.

$$V_{i,t}^j = x_{i,t}^j + F^*\left(x_{r1,t}^j - x_{r2,t}^j\right) \tag{26}$$

$$F = F_{min} + (F_{max} - F_{min})^* \frac{1}{1 + e^{10(t - 0.5^*T)/T}} \tag{27}$$

Step 3: Cross operation. The target vector and mutation vector are crossed to generate a test vector. The cross rule is shown in Eq. (28). CR is distributed at intervals [0, 1].

$$u_{i,t}^j = \begin{cases} V_{i,t}^j\ ,\ if\ rand(j) \le CR\ or\ j = randn(t) \\ x_{i,t}^j\ ,\ otherwise \end{cases} \tag{28}$$

Step 4: Selecting operation. The fitness value of the target vector is compared with its corresponding test vector. The better vector is chosen as the individual entering the next-generation population.

3. Overall rationale of interpretable wind speed forecasting and procedure of the proposed model

3.1. Rationale of the proposed model

In the VMD-ADE-TFT forecasting system, VMD is used to decompose raw wind speed sequence into sub-sequences; ADE optimizes several key parameters of TFT. To forecast wind speed, the sub-sequences and meteorological data are input to the improved TFT. Interpretable prediction results are finally derived. The VMD-ADE-TFT model combines the advantages of VMD, ADE, and TFT. The rationality of using VMD and ADE is explained in detail below.

3.1.1. Rationale of using VMD to reduce the wind speed sequence

As a new and effective decomposition technique, VMD decomposes raw wind speed sequences into several sub-sequences. Doing so eliminates the noise of the original sequence and mine

its main characteristics. Among the decomposition techniques, EMD-based and WT-based methods are the most extensively applied. EMD-based techniques are adaptive and require fewer parameters to be adjusted. These methods are more sensitive to noise and sampling though, so they decompose easily but inadequately. WT-based methods possess outstanding localization characteristics in the time and frequency domains. However, the performance of WT-based methods largely depends on the number of decomposition layers and the structure of the decomposition binary tree [37]. In comparison, VMD can achieve precise separation of signals and has higher computational performance owing to its strong mathematical theoretical foundation [38]. The related experiments demonstrate that the decomposition performance of the VMD technique is more excellent than those of the EMD- and WT-based techniques [39].

VMD can fully decompose raw wind speed series into various sub-sequences, thereby minimizing the difficulty of prediction due to the large volatility and increased randomness of the original wind speed series. Meanwhile, TFT can well select sub-sequences that contribute more to wind speed prediction and sort different sub-sequences according to their importance to wind speed prediction, leading to a more enlightening analysis. Therefore, this study utilizes VMD to decompose the raw wind speed series.

3.1.2. Rationale of using ADE to optimize TFT parameters

A common intelligent optimization algorithm, the DE algorithm has the advantages of easy implementation and efficiency [40]. Numerous studies reveal that DE outperforms other common algorithms (e.g. genetic algorithm) on many problems, such as global optimization problems [41]. Thus, DE-optimized TFT is likely to perform well. The DE algorithm has been recently employed to solve many types of problems [42], but it is rarely utilized for forecasting wind speed. In particular, ADE has better global search capabilities and search efficiency than basic DE.

The key parameters of the TFT include the number of time steps, number of batch sizes, learning rates, number of hidden layers, number of attention heads, and number of hidden layer neurons. All these parameters significantly influence the performance of the TFT. However, in certain applications, setting appropriate values for these parameters is challenging. Therefore, this study uses the ADE algorithm to determine the optimal value of these six key parameters.

3.2. Forecasting procedure of the proposed model

Fig. 2 depicts the forecasting procedure of the proposed model. The procedure involves data collection, decomposition of wind speed series, ADE optimization, TFT forecasting, and interpretable analysis. In the optimization module, each individual in the ADE population corresponds to a solution vector, that is, the value of the six parameters of TFT; hence, the gene dimension of each individual is six. At the end of the iterative process of the ADE algorithm, the optimized parameter values can be derived according to the current best individual. MAPE is selected as the fitness function. The steps of the prediction process of the VMD-ADE-TFT model are summarized below.

Step 1: VMD is adopted to decompose the raw wind speed sequence into multiple sub-sequences. VMD can clean the noise of the original sequence and mine its main characteristics. The number of sub-sequences is figured out based on r_{res}.

Step 2: All sub-sequences obtained by decomposition and meteorological data are inputted into the ADE-optimized TFT (ADE-TFT) as past inputs for prediction. Given the availability of weather forecasts, meteorological data are also inputted into the TFT model as future known variables. In ADE-TFT, the six parameters of TFT are optimized through ADE, which then further improves prediction accuracy.

Step 3: The forecasting performance and interpretable results from the VMD-ADE-TFT model are analyzed. Eight hourly wind speed data sets in two regions are adopted. Mean absolute error (MAE), root means square error (RMSE), and mean absolute percentage error (MAPE) are used to evaluate the prediction results. Interpretable results are mainly separated into three parts: the importance order of past variables, the importance order of future variables, and the attention of different lags orders. Finally, research on the interpretability of input variables for wind speed prediction is enriched.

4. Experimental study

As case studies, two sets of wind speed and meteorological data gathered from Saskatchewan, Central Canada, and the western coast of the United States are applied to assess the forecasting performance of the proposed combined model. Since the locations where the two sets of data are located have different climatic characteristics, the generalization ability of the proposed model can be better verified.

This section presents the implementation of deep learning models by *Python 3.8* with TensorFlow 2.2.0, PyTorch-forecasting 0.9.0, PyTorch-lightning 1.5.0, Torch 1.10.0. The Python library "TimeSeriesDataSet" is used to split the data. The TFT model uses an ADAM optimizer. Early Stopping is used to prevent overfitting. The TFT model is trained on the CPU. The computation is evaluated on an efficient computer with an *Intel (R) Core (TM) i7-10700K CPU, 3.*80 GHz, 32 GB RAM*, and Windows 10 system*.

4.1. Case A: Prince Albert wind speed forecasting

4.1.1. Data retrieval and pre-processing

Located in Saskatchewan, Central Canada, Prince Albert possesses rich wind energy resources. Its hourly wind speed and meteorological data (i.e., wind direction, temperature, dew point temperature, relative humidity, barometric pressure) are sourced from the official website (https://www.canada.ca/en/services/environment/weather.html) [43]. Wind speed data vary substantially with the seasons. For analysis, this study breaks down the data for one year into four seasons: spring (February 2021 to April 2021), summer (May 2021 to July 2021), autumn (August 2021 to October 2021), and winter (November 2020 to January 2021). These four wind speed series are depicted in Fig. 3. The original wind speed data are separated into training, validation, and testing sets, and the training-to-validation-to-testing ratio is 80%:10%:10%. For example, in the spring data set, the training, validation, and testing sets include 1736, 200, and 200 observations, respectively. The three groups are adopted to build models, select hyperparameters, and verify the final model in sequence.

To distinctly show the relationship between meteorological data and wind speed, these time series are linearly scaled between 0.1 and 0.9. With the spring data set as an example, Fig. 4 presents the relationship between meteorological data and wind speed. Table 2 presents the Pearson Correlation Coefficient between meteorological data and wind speed. The results of the Pearson test show that, in the four data sets, temperature and wind speed have a positive correlation, and relative humidity and barometric pressure have a negative correlation with wind speed. The Pearson correlation between dew point temperature and wind speed fluctuates with the seasons. Pearson correlation between wind direction and wind speed is only significant in the autumn dataset.

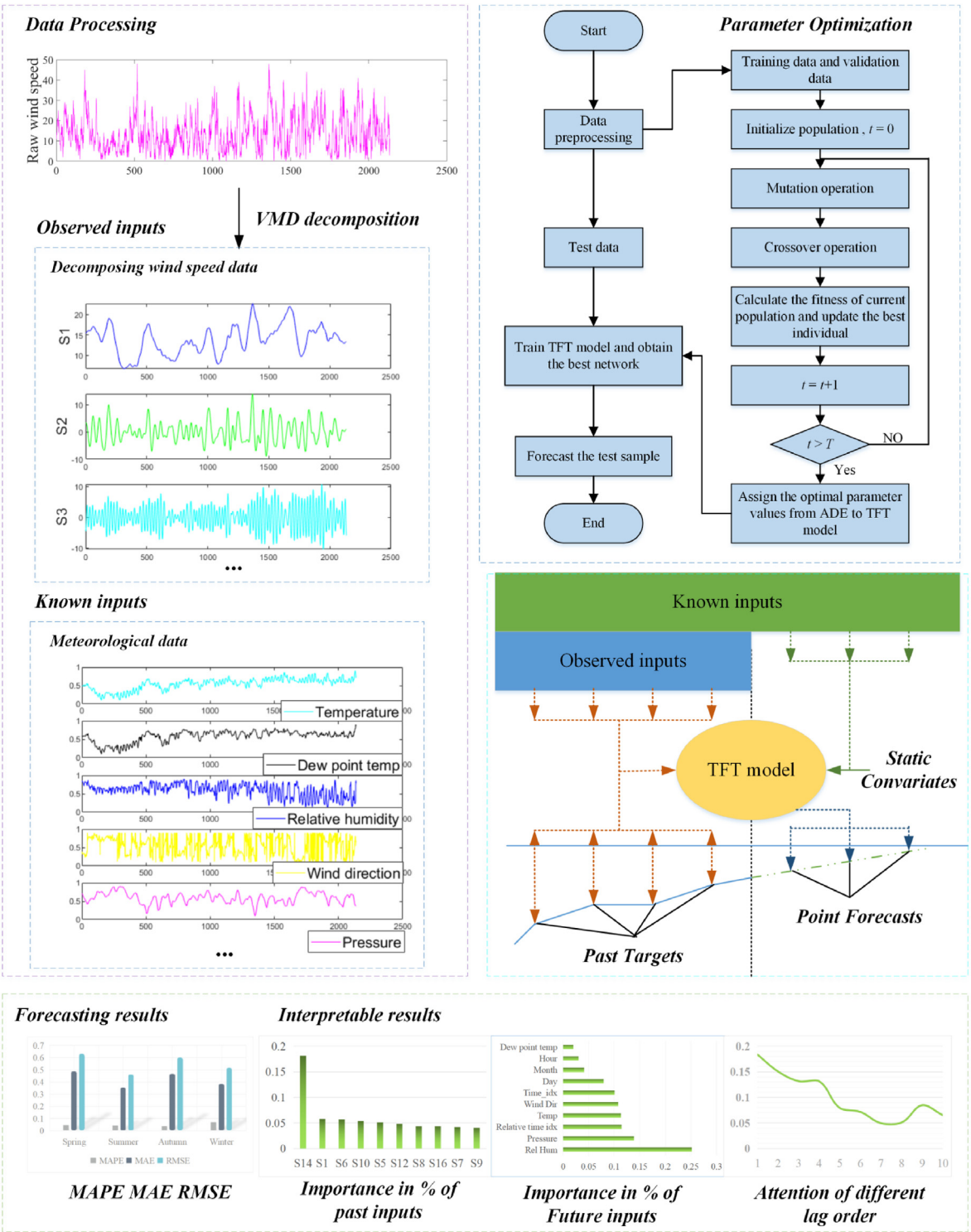

Fig. 2. Flowchart of the VMD-ADE-TFT model.

4.1.2. Evaluation metrics

Two scale-dependent errors and a percentage error, namely, MAE, RMSE, and MAPE are utilized to evaluate forecasting performance. The calculation of the three evaluation metrics is given by Eqs. (29)–(31).

$$MAPE = \frac{\sum_{t=1}^{k} |\hat{y}_t - y_t| / y_t}{k} \tag{29}$$

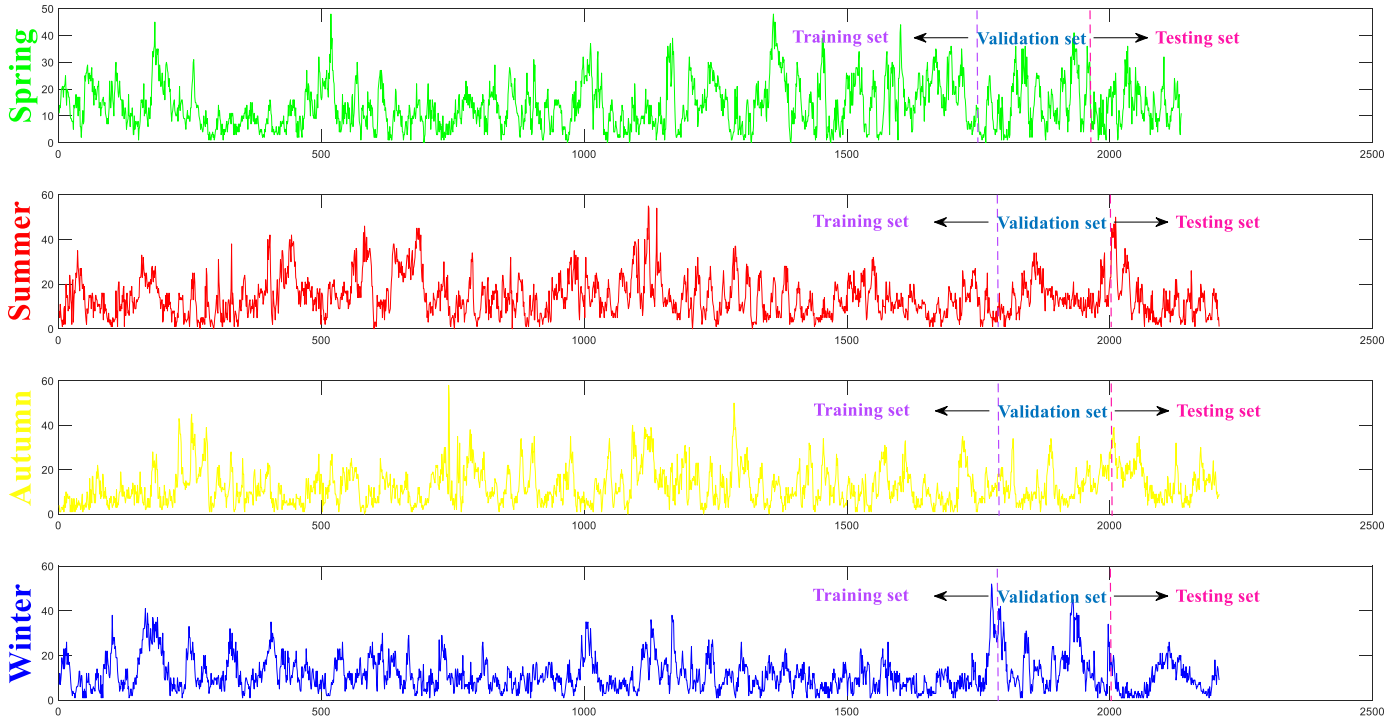

Fig. 3. Wind speed series collected from Albert.

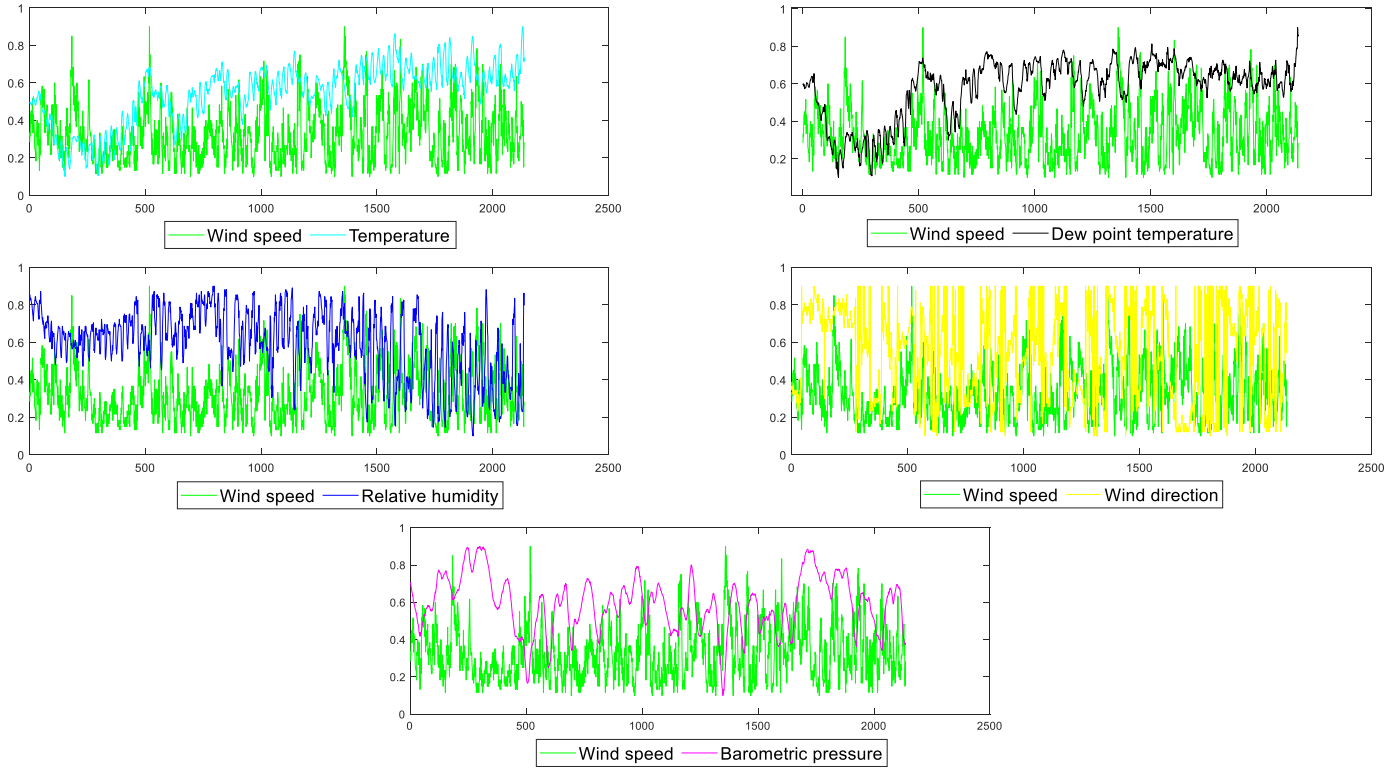

Fig. 4. Comparison chart of meteorological data and wind speed data.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{k}(\hat{y}_t - y_t)^2}{k}}, \tag{30}$$

$$MAE = \frac{1}{k}\sum_{t=1}^{k}\left|(\hat{y}_t - y_t)\right| \tag{31}$$

Table 2
Pearson Correlation Coefficient between meteorological data and wind speed in Albert's data set.

	Spring	Summer	Autumn	Winter
Temperature	0.269**	0.195**	0.140**	0.286**
Dew point temperature	0.185**	− 0.085**	− 0.090**	0.251**
Relative humidity	− 0.284**	− 0.341**	− 0.330**	− 0.192**
Wind direction	−0.025	0.041	0.144**	0.034
Barometric pressure	− 0.206**	− 0.251**	− 0.145**	− 0.200**

Note: ** denotes meteorological data and wind speed have a significant correlation at the 1% level (two-tailed).

k is the size of output samples, y_t is the actual value, and $\widehat{y}_t$ is the forecasting value.

The improvement rate (IR) is introduced to compare the forecasting performance of two different models. The three improved percentage metrics are calculated as follows:

$$IR_{MAPE} = \frac{MAPE_A - MAPE_B}{MAPE_B} \times 100\% \tag{32}$$

$$IR_{MAE} = \frac{MAE_A - MAE_B}{MAE_B} \times 100\% \tag{33}$$

$$IR_{RMSE} = \frac{RMSE_A - RMSE_B}{RMSE_B} \times 100\% \tag{34}$$

where IR_{MAE}, IR_{MAPE}, and IR_{MAE} represent the IRs of model A compared with model B in terms of MAE, MAPE, and MAE, respectively.

4.1.3. Comparable models and parameter setting

To confirm the stability and accuracy of the proposed forecasting model, its performance is compared with those of ten comparable models, namely, VMD-ADE-LSTM, VMD-ADE-RNN, VMD-ADE-GRU, VMD-ADE-BPNN, VMD-DE-TFT, VMD-GA-TFT, EMD-ADE-TFT, EEMD-ADE-TFT, the persistence model, and VMD-ADE-TFT without future meteorological data. Among them, LSTM, recurrent neural network (RNN), gated recurrent unit (GRU), backpropagation neural network (BPNN), and TFT are compared as individual forecasting models. They are recognized models for forecasting wind speed [44,45]. For a fair comparison, the intelligent algorithm for parameter optimization is adopted for all individual models. GA, DE, and ADE are employed as different intelligent optimization algorithms to search the parameters of the TFT model to identify a more satisfying optimization algorithm. In addition, this study compares the effects of VMD and other decomposition methods (i.e., EMD and EEMD). Prediction models using different decomposition techniques, that is, EMD-ADE-TFT, EEMD-ADE-TFT, and VMD-ADE-TFT are used for comparison.

The raw wind speed data are decomposed into several sub-modes by using VMD to reduce the non-stationary characteristics. The appropriate number of sub-modes can be specified by the ratio of residual energy r_{res} (see Equation (7)). The appropriate number of sub-modes must be determined. If the number of sub-modes is small, then the original sequence may be incompletely decomposed, resulting in inaccurate predictions. If the signal is decomposed too much, then the difference between the sub-modalities becomes very small, thus reducing the accuracy and increasing unnecessary computational overhead. When r_{res} is less than 3% and the downward trend is not significant, the number of sub-modes can be decided. Table 3 gives the r_{res} corresponding to the different number of sub-modes. The suitable number of sub-modes of the four data sets are 16, 15, 16, and 16. Fig. 5 illustrates the sub-modes of the decomposition results of the spring data set. The relatively low-frequency sub-modes represent the overall trend of the original wind speed; the higher frequency, the local fluctuation trend. The extracted sub-sequence is smoother than the original data, which helps enhance wind speed prediction performance.

Table 3
r_{res} corresponding to the different numbers of sub-modes in Albert's data set.

K	Spring (r_{res}, %)	Summer (r_{res}, %)	Autumn (r_{res}, %)	Winter (r_{res}, %)
2	16.47	22.08	15.53	17.02
3	14.96	13.87	12.78	15.76
4	12.79	11.74	11.26	13.44
5	11.25	10.60	10.12	11.66
6	9.57	9.64	8.86	10.05
7	8.67	8.13	7.77	8.64
8	7.38	7.04	7.02	7.78
9	6.40	6.23	6.36	6.86
10	5.47	5.72	5.82	6.05
11	4.97	5.03	5.43	4.94
12	4.43	4.74	5.21	4.32
13	3.75	4.47	4.87	3.75
14	3.43	3.94	4.72	3.37
15	3.03	**2.82**	4.62	2.94
16	**2.85**	2.67	**2.65**	**2.62**
17	2.61	2.43	2.50	2.42
18	2.51	2.22	2.43	2.24
19	2.45	2.11	2.27	2.10
20	2.23	2.03	2.22	1.85

In this study, the grid search method, which features the advantages of high efficiency and simplicity [46–48], is adopted to optimize the parameters of DE, ADE, and GA. These evolutionary algorithms are subsequently used to search for the optimal parameters of TFT. The search scope of TFT parameters is as follows: the range of the number of time steps is set within [1,24]; number of batch sizes, [16,64]; learning rates, [0.001,0.1]; number of hidden layers, [2,32]; number of attention heads, [1,4]; number of hidden layer neurons, [2,32]. The final parameters of the VMD-ADE-TFT in the four data sets are shown in Table 4. Table 5 gives the input details of the TFT model, with the spring data set as an example. The dropout rate and the max gradient norm are selected by the grid search.

4.1.4. Experimental results and analysis

The prediction results of VMD-ADE-TFT on the four-season data sets are compared with those of other prediction methods regarding MAPE, MAE, and RMSE. Table 6 presents the forecasting performance of each model. The results are analyzed in detail below.

(a) Compared with those of the persistence model, VMD-ADE-LSTM, VMD-ADE-RNN, VMD-ADE-GRU, VMD-ADE-BPNN, the error value of TFT is smaller than the above models on all four data sets. The MAPE value of the spring data set is the smallest among the six models. The MAPE values of the persistent model, VMD-ADE-LSTM, VMD-ADE-RNN, VMD-ADE-GRU, VMD-ADE-BPNN, and VMD-ADE-TFT are 47.9%,

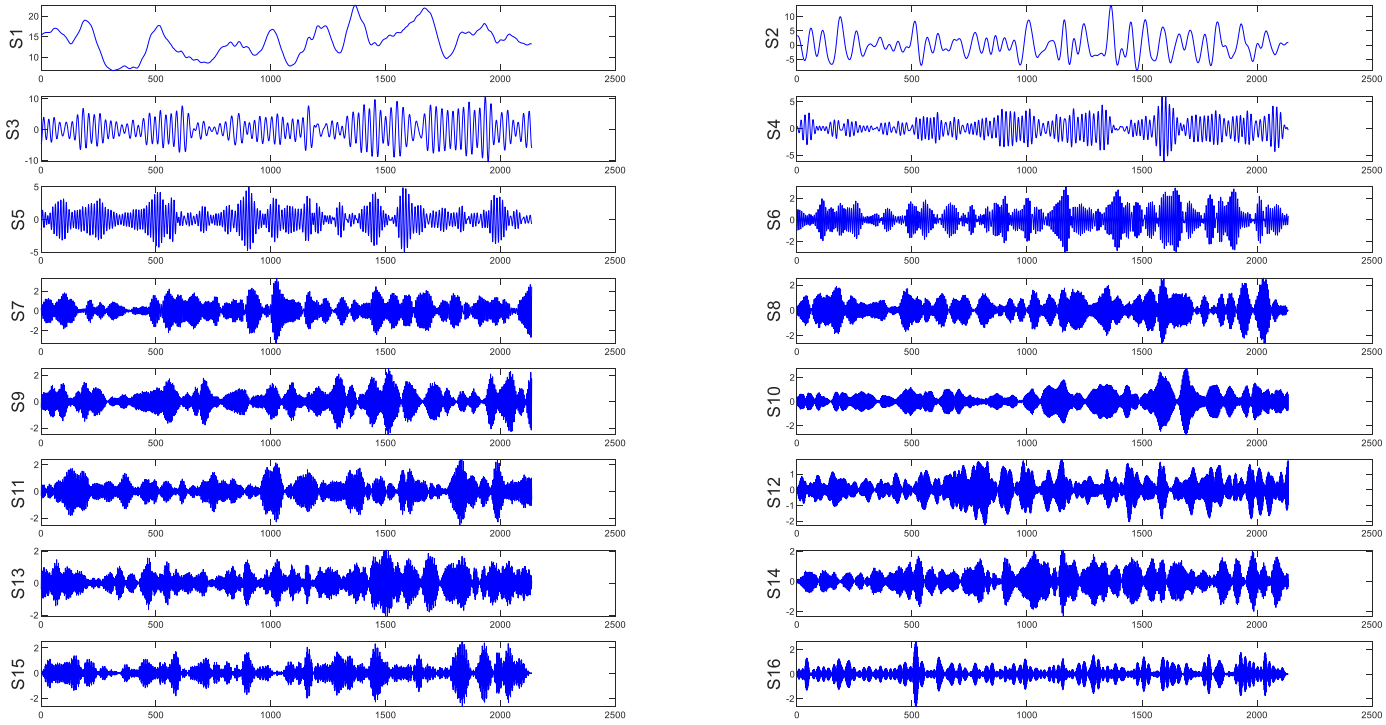

Fig. 5. Decomposition results by VMD for Albert's spring data set.

9.4%, 12.1%, 12.8%, 12.8%, and 4.6%, respectively. The results reveal that the TFT model is the best individual prediction model.

(b) VMD-ADE-TFT outperforms VMD-GA-TFT and VMD-DE-TFT in each evaluation index on the spring, summer, and winter data sets. For example, Table 6 indicates that the MAPE value of VMD-ADE-TFT in the spring data set is 42.5% and 41.8% higher than that of VMD-GA-TFT and VMD-DE-TFT, respectively. In the autumn data set though, VMD-GA-TFT performs better than VMD-ADE-TFT in respect of MAE. This outcome is due to the contingency of parameter optimization. The results show that, in most cases, the prediction model based on the ADE optimization of TFT parameters can achieve better results than the TFT model optimized by GA or DE.

(c) Tables 6 and 7 show that the performance of VMD-ADE-TFT is significantly better than that of EMD-ADE-TFT and EEMD-ADE-TFT. This result indicates that VMD can decompose the raw wind speed sequence more fully, thereby obtaining better prediction results.

(d) Given the availability of weather forecasts, future meteorological data are typically input into forecasting models as a known variable. Therefore, this study compares two scenarios: knowing only the past meteorological data and setting the meteorological data as a future known variable. Tables 6 and 7 reveal that setting meteorological data as known variables in the future significantly improves the prediction performance of the TFT model. For example, the MAPE value of VMD-ADE-TFT in the spring data set is 54.9% higher than that of VMD-ADE-TFT without future meteorological data. Fig. 6 depicts the forecasting results of the two models. The correlation coefficient R is also given, which describes the degree of collinearity between the actual values and predicted results. The R-value of VMD-ADE-TFT with future meteorological data is the largest, and the scatter points are most evenly distributed near the regression line. The prediction results of VMD-ADE-TFT with future meteorological data are closer to the true value than that of VMD-ADE-TFT without future meteorological data. Furthermore, the results of Table 6 show that VMD-ADE-TFT without future meteorological data underperforms comparatively to several other models on some datasets because the other models take future meteorological data as input variables into the prediction models. This further verifies the important role of future meteorological data in wind speed prediction.

(e) This study examines the predictive ability of each model for peak wind speed periods. The wind speed fluctuations near the peak of the wind speed are very violent, so the accurate prediction of the wind speed during the peak period can better measure the performance of each model. In this study, peak wind speed periods consist of three moments before and after the peak wind speed in the test set, a total of seven moments. The forecasting performances of each model are shown in Table 7 and Fig. 7. In the spring, summer, and winter data set, the MAPE of VMD-ADE-TFT is smaller than the above models, and its predicted value is also closer to the true value, which suggests a higher performance than other models.

Fig. 8 presents the interpretable results of the VMD-ADE-TFT model in four seasons. The results are broken down into three parts: the importance order of past variables, the importance order of future variables, and the attention of different lags orders. Given the large number of variables entered in the past, the figure only illustrates the top ten variables in importance. The interpretable results are detailed below.

(a) Among the past variables, the decomposed partial wind speed sub-modes are more helpful for wind speed prediction

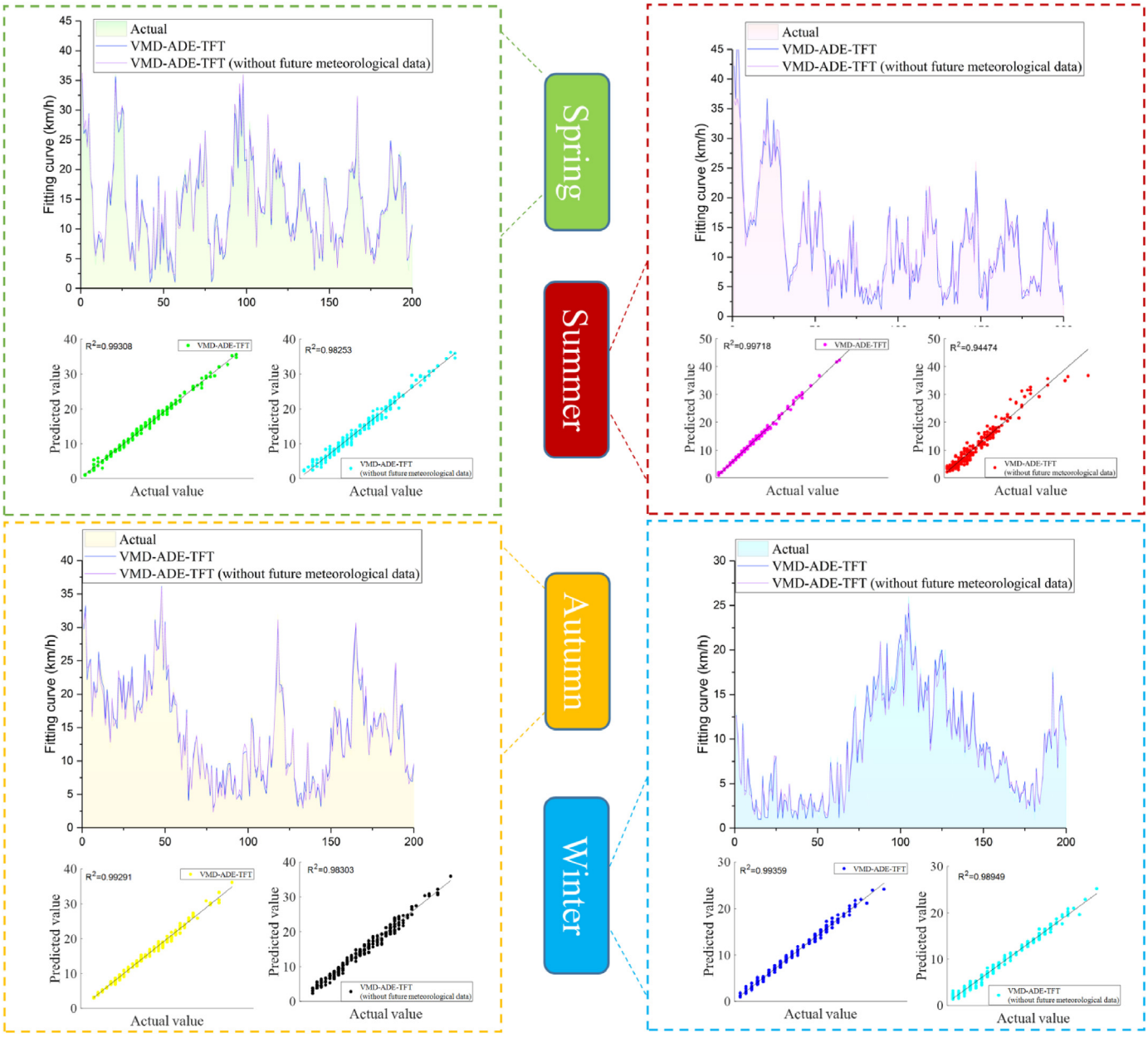

Fig. 6. Forecasting results of the VMD-ADE-TFT model with (or without) future meteorological data in four seasons.

than meteorological data. Low-frequency and high-frequency sub-modes help better in the prediction than intermediate-frequency sub-modes. For example, S2 is the most important variable in the summer, autumn, and winter data sets, and S16 (highest frequency sub-mode) is the most important variable in the spring data set. S2 can better represent the overall trend of the original wind speed than S1. Compared with the intermediate-frequency sub-modes, the high-frequency ones can better reflect the local fluctuation trend of wind speed and contribute more to wind speed prediction.

(b) Regarding the importance of future inputs, the importance of variables varies greatly depending on the season. The relative humidity is the most important future variable in the spring data set. Hour is the most important one in the summer data set, showing that wind speed is very sensitive to the time of day in summer. In the autumn and winter data sets, the temperature is the most important variable as it varies greatly in these two seasons. Wind direction and barometric pressure are relatively important variables as well. In the four data sets, the contribution of dew point temperature to wind speed predictions is low.

(c) The interpretability results show the general trend of attention changes, that is, the smaller the lag order is, the greater the contribution to wind speed prediction. Larger lag order sometimes also corresponds to larger attention. For example, in the autumn data set, the lag order of 9 achieves 11.1% attention. The predictive model must have the memory capacity to retrieve long-term inputs.

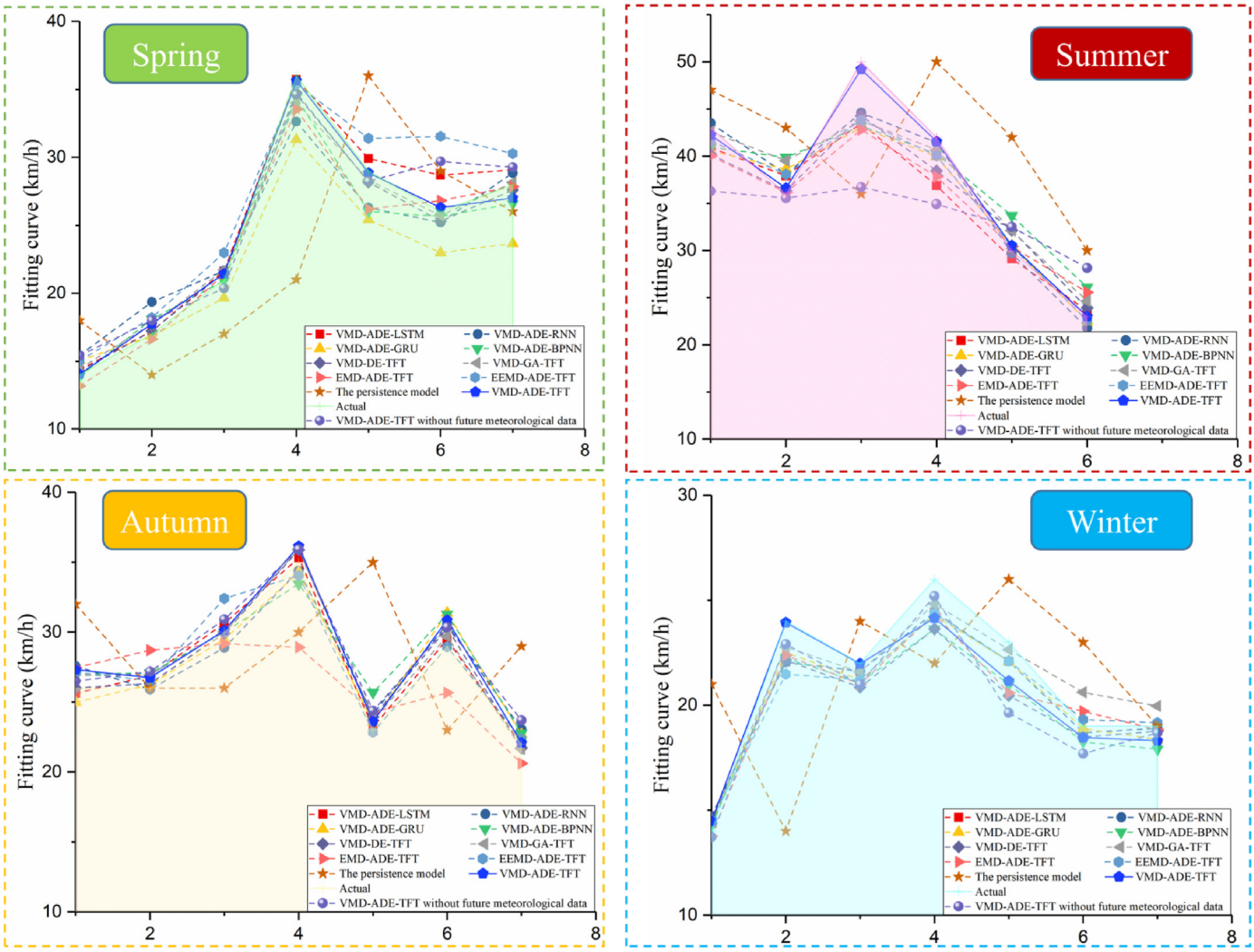

Fig. 7. Peak wind speed forecasting test results of different models in four seasons.

Table 4
Parameters of the VMD-ADE-TFT in the four data sets.

	Parameter	Spring	Summer	Autumn	Winter
VMD	Number of sub-modes (K)	16	15	16	16
ADE	Population size (M)	15	15	20	20
	Maximum number of iterations (T)	20	30	25	30
	Crossover probability (CR)	0.2	0.4	0.6	0.4
	Mutation operator (F)	[0,1]	[0,1]	[0,1]	[0,1]
TFT	Number of time steps	10	10	12	11
	Number of batch sizes	32	35	48	41
	Learning rates	0.048	0.029	0.086	0.01
	Number of hidden layers	24	23	16	18
	Number of attention heads	1	1	1	1
	Number of hidden layer neurons	13	10	8	7
	Dropout rate	0.1	0.1	0.1	0.1
	Max gradient norm	0.1	0.1	0.1	0.1

4.1.5. Extended experiment

To further verify the robustness of the proposed model, this study collected Albert's wind speed and meteorological data from November 2019 to October 2020. This study still disassembles the data of one year into four seasons. As shown in Fig. 9, there are certain differences in the fluctuation range and trend of wind speed series in different years, so the extended experiments can also prove the generalization ability of the proposed model. The wind speed data of the four data sets are both separated into training, validation, and testing sets and the training-to-validation-to-testing ratio is 80%:10%:10%.

After a series of experiments, the final parameters of the VMD-ADE-TFT in the four data sets are presented in Table 8. As shown in Table 9, in most cases, the proposed model outperforms other

Table 5
Inputs of the VMD-ADE-TFT in the spring data set.

Static covariates	Past inputs	Future inputs
ID (name of wind speed series)	S1−S16	Wind direction
–	Wind direction	Temperature
–	Temperature	Dew point temperature
–	Dew point temperature	Relative humidity
–	Relative humidity	Barometric pressure
–	Barometric pressure	Hour
–	Hour	Day
–	Day	Month
–	Month	Time index
–	Time index	Relative time index
–	Relative time index	–

Table 6
Forecasting results of different models in the Albert data set.

Model	Spring			Summer			Autumn			Winter		
	MAPE/ IR_{MAPE}	MAE/ IR_{MAE}	RMSE/ IR_{RMSE}	MAPE/ IR_{MAPE}	MAE/ IR_{MAE}	RMSE/ IR_{RMSE}	MAPE/ IR_{MAPE}	MAE/ IR_{MAE}	RMSE/ IR_{RMSE}	MAPE/ IR_{MAPE}	MAE/ IR_{MAE}	RMSE/ IR_{RMSE}
VMD-ADE-LSTM	9.4% (51.1%)	0.911 (46.4%)	1.183 (46.6%)	11.8% (64.4%)	0.753 (53.3%)	1.060 (56.5%)	8.0% (56.3%)	0.939 (50.3%)	1.175 (48.9%)	12.0% (42.5%)	0.423 (9.5%)	0.559 (7.7%)
VMD-ADE-RNN	12.1% (62.0%)	1.157 (57.8%)	1.446 (56.3%)	10.4% (59.6%)	0.628 (43.9%)	0.863 (46.6%)	7.4% (52.7%)	0.790 (40.9%)	0.979 (38.6%)	12.1% (43.0%)	0.424 (9.7%)	0.560 (7.9%)
VMD-ADE-GRU	12.8% (64.1%)	1.425 (65.8%)	1.759 (64.1%)	11.4% (63.2%)	0.726 (51.5%)	1.023 (54.9%)	8.3% (57.8%)	0.907 (48.5%)	1.118 (46.2%)	12.3% (43.9%)	0.415 (7.7%)	0.542 (4.8%)
VMD-ADE-BPNN	12.4% (62.9%)	1.248 (60.1%)	1.530 (58.7%)	10.6% (60.4%)	0.815 (56.8%)	1.122 (58.9%)	8.1% (56.8%)	0.826 (43.5%)	1.032 (41.8%)	13.7% (49.6%)	0.555 (31.0%)	0.706 (26.9%)
VMD-DE-TFT	7.9% (41.8%)	0.713 (31.6%)	0.919 (31.2%)	9.9% (57.6%)	0.770 (54.3%)	1.032 (55.3%)	5.7% (38.6%)	0.590 (20.8%)	0.743 (19.1%)	12.7% (45.7%)	0.507 (24.5%)	0.668 (22.8%)
VMD-GA-TFT	8.0% (42.5%)	0.697 (30.0%)	0.881 (28.3%)	10.6% (60.4%)	0.662 (46.8%)	0.931 (50.5%)	3.6% (2.7%)	**0.457** (−2.2%)	0.640 (6.1%)	17.4% (60.3%)	0.564 (32.1%)	0.710 (27.3%)
EMD-ADE-TFT	12.0% (61.7%)	1.231 (60.4%)	1.524 (58.5%)	12.4% (66.1%)	0.793 (55.6%)	1.135 (59.4%)	14.4% (75.7%)	1.842 (74.6%)	2.370 (74.6%)	12.5% (44.8%)	0.456 (16.0%)	0.585 (11.8%)
EEMD-ADE-TFT	16.0% (71.3%)	1.508 (67.6%)	1.798 (64.8%)	12.1% (65.3%)	0.702 (49.9%)	0.951 (51.5%)	10.5% (66.7%)	0.984 (52.5%)	1.254 (52.1%)	12.1% (43.0%)	0.391 (2.0%)	0.527 (2.1%)
The persistence model	47.9% (90.4%)	4.480 (89.1%)	5.872 (89.2%)	53.1% (92.1%)	3.88 (90.9%)	5.054 (90.9%)	30.7% (88.6%)	3.480 (86.6%)	4.695 (87.2%)	48.2% (85.7%)	2.385 (83.9%)	3.161 (83.7%)
VMD-ADE-TFT (without future meteorological data)	10.2% (54.9%)	0.794 (38.5%)	1.005 (37.1%)	21.6% (80.6%)	1.432 (75.4%)	2.026 (77.2%)	6.8% (48.5%)	0.772 (39.5%)	0.936 (35.8%)	16.1% (57.1%)	0.664 (42.3%)	0.852 (39.4%)
VMD-ADE-TFT	**4.6**%	**0.488**	**0.632**	**4.2**%	**0.352**	**0.461**	**3.5**%	0.467	**0.601**	**6.9**%	**0.383**	**0.516**

Note: Bold values indicate the best forecasting performance. *IR* means the proposed model v.s. the current model.

Table 7
Peak wind speed forecasting test results of different models in the Albert data set.

Model	Spring			Summer			Autumn			Winter		
	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE
VMD-ADE-LSTM	3.6%	0.880	1.185	6.9%	2.865	3.609	**1.5**%	**0.426**	**0.496**	3.8%	0.846	1.057
VMD-ADE-RNN	7.4%	1.739	2.004	4.2%	1.679	2.439	3.8%	1.023	1.133	3.9%	0.856	1.069
VMD-ADE-GRU	9.6%	2.595	3.062	5.9%	2.513	3.317	3.2%	0.879	1.093	3.8%	0.857	1.008
VMD-ADE-BPNN	4.1%	1.090	1.456	9.8%	3.545	3.913	5.1%	1.337	1.584	5.8%	1.131	1.485
VMD-DE-TFT	2.3%	0.597	0.685	6.4%	2.578	3.129	1.9%	0.544	0.709	5.9%	1.321	1.580
VMD-GA-TFT	2.9%	0.713	0.905	6.7%	2.540	3.173	2.4%	0.660	0.733	4.2%	0.886	**0.995**
EMD-ADE-TFT	4.3%	1.121	1.499	7.3%	2.920	3.764	8.4%	2.437	2.969	4.9%	1.113	1.350
EEMD-ADE-TFT	9.0%	2.150	2.630	4.8%	2.065	2.836	2.8%	0.797	1.097	4.3%	0.980	1.244
The persistence model	21.4%	5.428	6.845	24.3%	8.666	9.291	22.1%	5.714	6.611	21.4%	4.285	5.264
VMD-ADE-TFT (without future meteorological data)	6.3%	1.449	1.737	15.1%	5.859	7.124	4.3%	1.137	1.195	5.3%	1.168	1.514
VMD-ADE-TFT	**1.8%**	**0.431**	**0.532**	**1.3**%	**0.523**	**0.584**	3.0%	0.840	1.033	**3.6%**	**0.783**	1.057

models in terms of MAPE, MAE, and RMSE. Although the VMD-ADE-TFT is not the best model on the winter data set, it is the top three model. The extended experiment further verifies the robustness and generalization ability of the proposed model.

4.2. Case B: Five Points wind speed forecasting

4.2.1. Data retrieval

Five Points is located in San Joaquin Valley, Fresno, California, USA. The hourly wind speed and meteorological data (i.e., air temperature, dew point temperature, soil temperature, relative

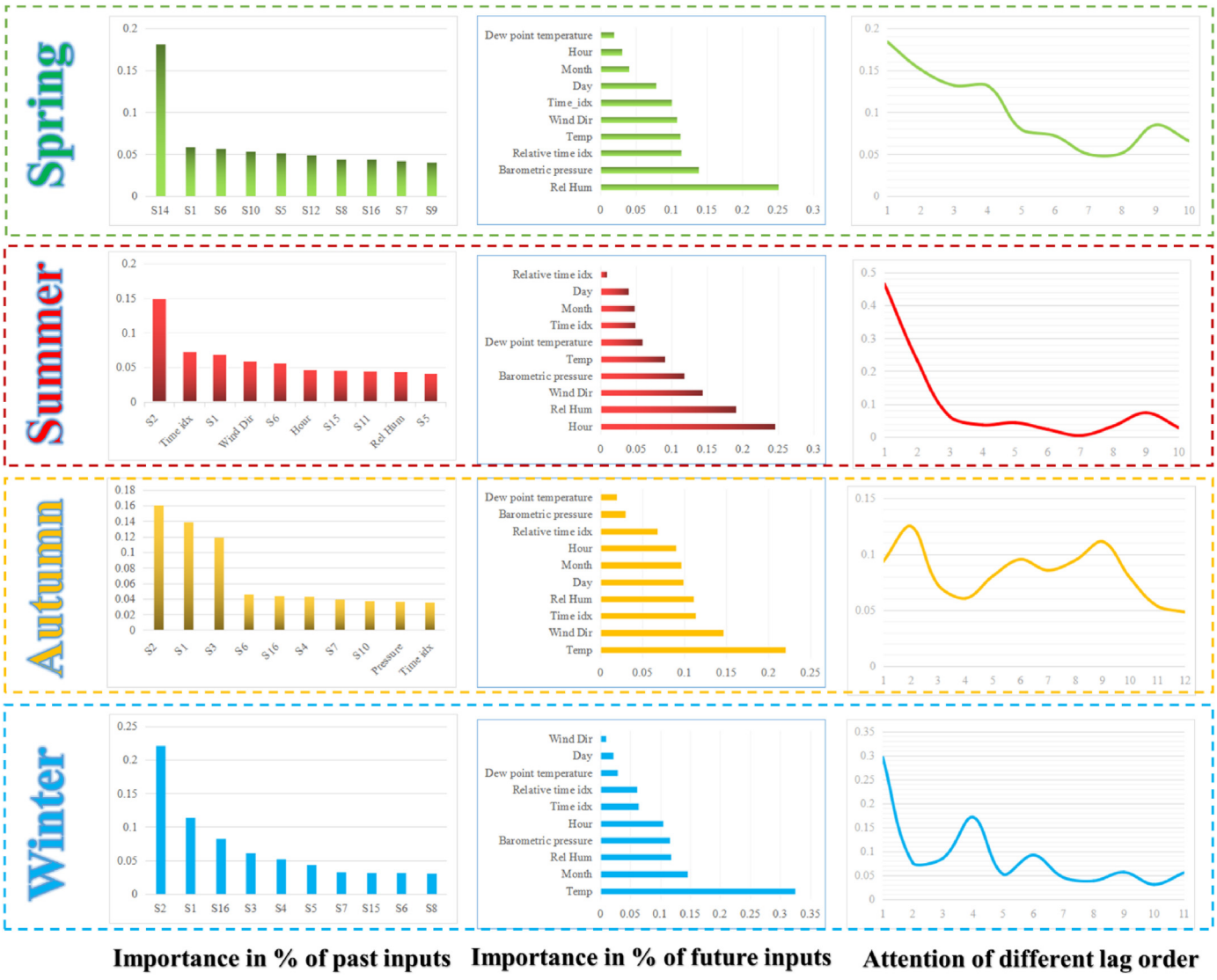

Fig. 8. Interpretable results of the VMD-ADE-TFT model in four seasons.

humidity, vapor pressure, and wind direction) are collected from California Irrigation Management Information System (https://cimis.water.ca.gov) [49]. The wind speed and weather data for a year are also divided into four data sets: spring (February 2021 to April 2021), summer (May 2021 to July 2021), autumn (August 2021 to October 2021), and winter (November 2020 to January 2021). Fig. 10 depicts these four wind speed series. The training-to-validation-to-testing ratio is 80%:10%:10%. In the summer data set, the training, validation, and testing sets consist of 1808, 200, and 200 observations, respectively. With the summer data set as an example, Fig. 11 illustrates the relationship between meteorological data and wind speed. Table 10 shows the Pearson Correlation Coefficient between meteorological data and wind speed. The results show that, in the four data sets, wind direction and air temperature have a positive correlation with wind speed, and relative humidity has a negative correlation with wind speed. In the spring data set, dew point temperature, vapor pressure, and soil temperature are not significantly correlated with wind speed.

4.2.2. Experimental design and results

The forecasting performances of VMD-ADE-TFT are compared with those of ten comparable models without future meteorological data. According to the r_{res}, the suitable number of sub-modes of the four data sets is 22, 23, 21, and 23. The grid search method is used to collect the parameters of all models. Table 11 lists the parameters of the VMD-ADE-TFT.

One-hour wind speed data during the four seasons is applied to prove the feasibility and validity of the proposed model. RMSE, MAE, and MAPE are used to examine the prediction performance of the proposed model and other comparable models. The prediction results are detailed in Tables 12 and 13 and Fig. 12.

(a) Compared with the persistence model, VMD-ADE-LSTM, VMD-ADE-RNN, VMD-ADE-GRU, and VMD-ADE-BPNN, the VMD-ADE-TFT model obtains almost all the best evaluation metrics (e.g., MAPE, RMSE, and MAE) in all data sets. The proposed model obtains the best prediction results. Simply put, the TFT model outperforms other individual methods in terms of prediction performance.

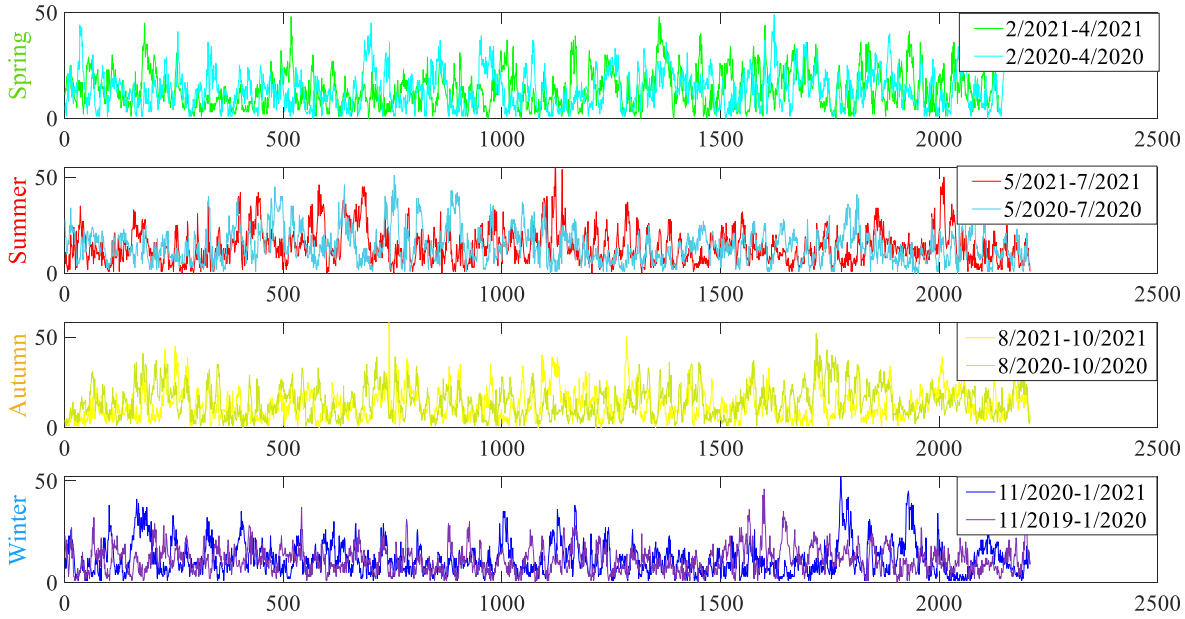

Fig. 9. Comparison chart of wind speed series in different years.

Table 8
Parameters of the VMD-ADE-TFT in the four data sets.

	Parameter	Spring	Summer	Autumn	Winter
VMD	Number of sub-modes (K)	32	30	32	23
ADE	Population size (M)	16	20	17	15
	Maximum number of iterations (T)	25	35	20	30
	Crossover probability (CR)	0.3	0.4	0.5	0.2
	Mutation operator (F)	[0,1]	[0,1]	[0,1]	[0,1]
TFT	Number of time steps	12	14	17	13
	Number of batch sizes	36	39	42	40
	Learning rates	0.065	0.061	0.083	0.059
	Number of hidden layers	15	19	16	18
	Number of attention heads	1	1	1	1
	Number of hidden layer neurons	8	7	8	9
	Dropout rate	0.1	0.1	0.1	0.1
	Max gradient norm	0.1	0.1	0.1	0.1

Table 9
Forecasting results of different models in the Albert data set.

Model	Spring			Summer			Autumn			Winter		
	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE
VMD-ADE-LSTM	10.4%	0.738	1.098	1.6%	0.047	0.074	4.9%	0.536	0.729	4.6%	0.362	**0.517**
VMD-ADE-RNN	11.0%	0.703	1.006	1.7%	**0.044**	0.076	5.0%	0.515	0.733	5.2%	0.447	0.713
VMD-ADE-GRU	11.1%	0.834	1.155	1.9%	0.067	0.085	4.4%	0.451	0.620	5.3%	0.451	0.751
VMD-ADE-BPNN	13.7%	0.912	1.256	2.0%	0.077	0.135	5.5%	0.507	0.659	8.8%	0.743	1.370
VMD-DE-TFT	8.5%	0.567	0.849	1.8%	0.081	0.108	4.8%	0.499	0.692	**3.7%**	**0.359**	0.652
VMD-GA-TFT	8.6%	0.552	0.824	1.7%	0.056	0.117	4.1%	0.436	0.610	6.4%	0.589	1.072
EMD-ADE-TFT	16.4%	1.172	1.814	2.1%	0.074	0.119	8.2%	0.814	1.066	9.8%	0.897	1.532
EEMD-ADE-TFT	11.8%	0.686	0.993	1.7%	0.074	0.129	4.2%	0.429	0.576	6.7%	0.604	1.042
The persistence model	51.2%	3.665	4.842	50.2%	3.495	4.954	27.3%	3.255	4.425	39.0%	3.125	4.074
VMD-ADE-TFT (without future meteorological data)	8.7%	0.562	0.857	1.9%	0.087	0.127	5.1%	0.598	0.907	6.5%	0.554	1.005
VMD-ADE-TFT	**6.0%**	**0.445**	**0.639**	**1.3%**	0.052	**0.073**	**3.3%**	**0.397**	**0.573**	4.4%	0.377	0.592

Note: Bold values indicate the best forecasting performance.

(b) The prediction performance of the VMD-ADE-VMD method is better than those of the VMD-GA-TFT and VMD-DE-TFT models in the spring, summer, and autumn data sets. The MAE and RMSE of VMD-ADE-TFT are 19.9% and 24.7% lower than those of VMD-GA-TFT in the winter data set. Nevertheless, the proposed model outperforms VMD-GA-TFT and VMD-DE-TFT in most cases, highlighting the effectiveness of ADE optimization.

(c) The VMD-ADE-TFT method outperforms the EMD-ADE-TFT and EEMD-ADE-TFT methods in all four data sets. For example, the MAPE of VMD-ADE-TFT decreases by 44.3% and

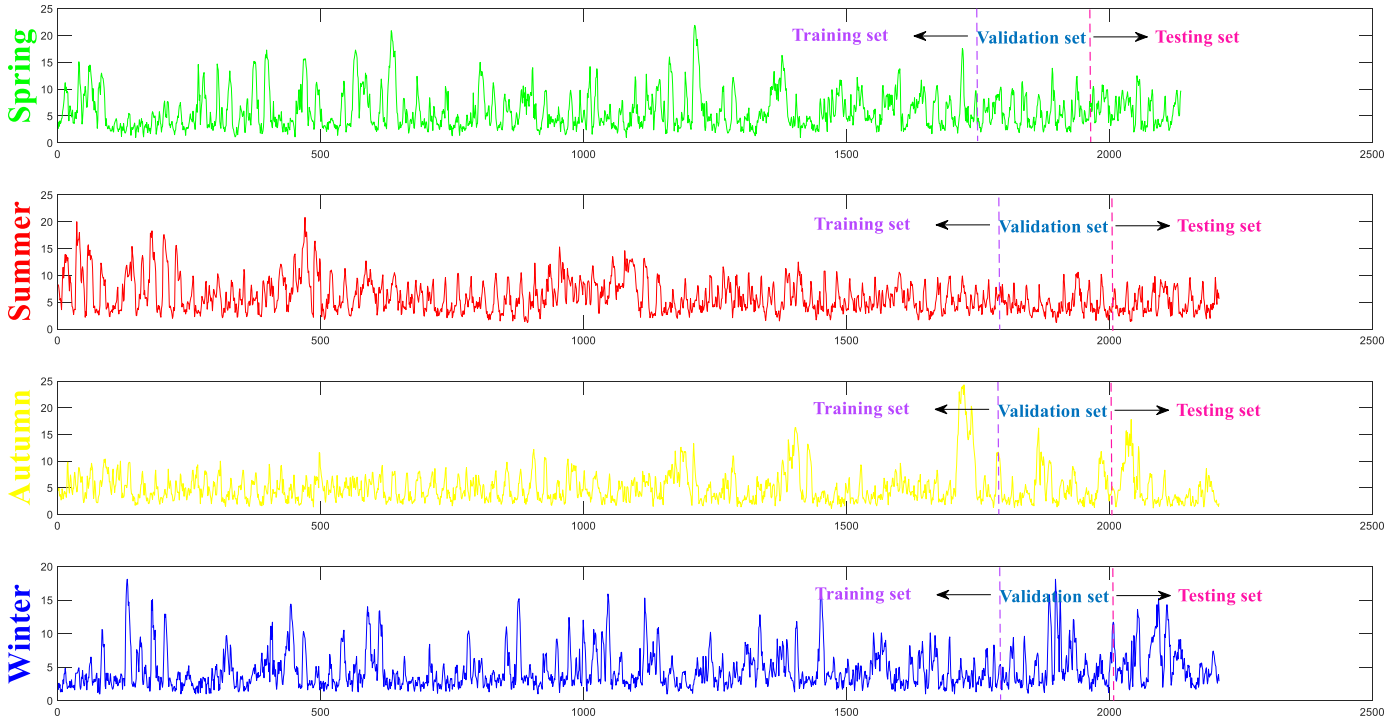

Fig. 10. Wind speed series collected from Five Points.

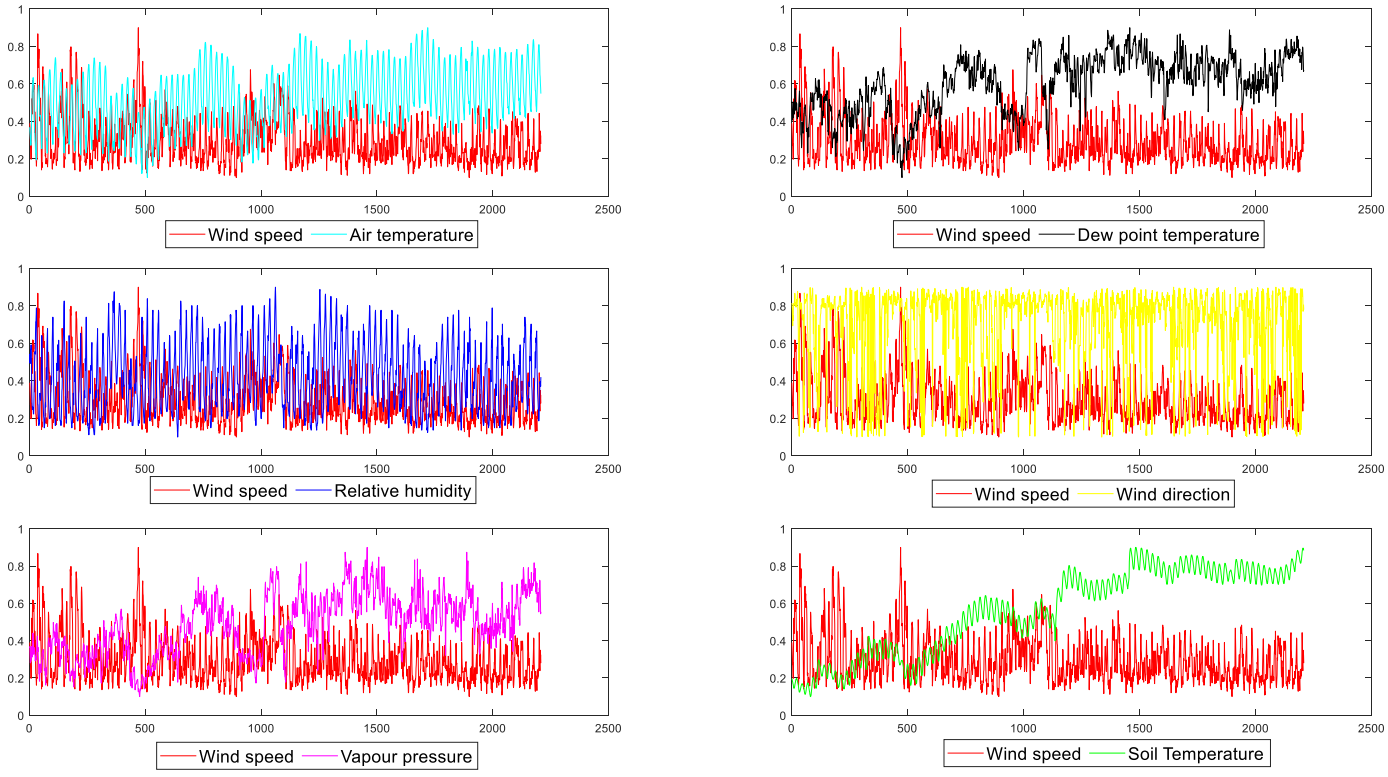

Fig. 11. Comparison chart of meteorological data and wind speed data.

44.7% in the spring data set, demonstrating the effectiveness and contribution of VMD decomposition.

(d) In the four data sets, setting future weather data as a known variable leads to better prediction performance than feeding only past weather data. Fig. 12 presents the forecasting results of the VMD-ADE-TFT model with (or without) future meteorological data. In all data sets, the R-value of VMD-ADE-TFT with future meteorological data is larger than that

Table 10
Pearson Correlation Coefficient between meteorological data and wind speed in Five Points' data set.

	Spring	Summer	Autumn	Winter
Air temperature	0.375**	0.204**	0.232**	0.433**
Dew point temperature	−0.038	− 0.281**	− 0.168**	0.191**
Relative humidity	− 0.371**	− 0.428**	− 0.332**	− 0.326**
Wind direction	0.411**	0.276**	0.239**	0.310**
Vapor pressure	−0.021	− 0.259**	− 0.119**	0.209**
Soil Temperature	0.037	− 0.285**	− 0.065**	− 0.078**

Note: ** denotes meteorological data and wind speed have a significant correlation at the 1% level (two-tailed).

Table 11
Parameters of the VMD-ADE-TFT in the four data sets.

	Parameter	Spring	Summer	Autumn	Winter
VMD	Number of sub-modes (K)	22	23	21	23
ADE	Population size (M)	30	20	15	25
	Maximum number of iterations (T)	25	25	25	30
	Crossover probability (CR)	0.3	0.6	0.4	0.3
	Mutation operator (F)	[0,1]	[0,1]	[0,1]	[0,1]
TFT	Number of time steps	18	12	15	12
	Number of batch sizes	40	32	48	39
	Learning rates	0.030	0.012	0.011	0.02
	Number of hidden layers	11	26	29	18
	Number of attention heads	1	1	1	1
	Number of hidden layer neurons	9	12	12	10
	Dropout rate	0.1	0.1	0.1	0.1
	Max gradient norm	0.1	0.1	0.1	0.1

Table 12
Forecasting results of different models in the Five Points data set.

Model	Spring			Summer			Autumn			Winter		
	MAPE/ IR_{MAPE}	MAE/ IR_{MAE}	RMSE/ IR_{RMSE}	MAPE/ IR_{MAPE}	MAE/ IR_{MAE}	RMSE/ IR_{RMSE}	MAPE/ IR_{MAPE}	MAE/ IR_{MAE}	RMSE/ IR_{RMSE}	MAPE/ IR_{MAPE}	MAE/ IR_{MAE}	RMSE/ IR_{RMSE}
VMD-ADE-LSTM	8.2% (58.5%)	0.471 (58.6%)	0.563 (55.6%)	8.8% (48.9%)	0.362 (46.4%)	0.464 (44.4%)	8.1% (42.0%)	0.346 (39.0%)	0.584 (47.6%)	8.8% (52.3%)	0.493 (52.3%)	0.672 (52.7%)
VMD-ADE-RNN	11.1% (69.4%)	0.673 (71.0%)	0.867 (71.2%)	10.3% (56.3%)	0.399 (51.4%)	0.516 (50.0%)	8.6% (45.3%)	0.375 (43.7%)	0.581 (47.3%)	12.2% (65.6%)	0.749 (68.6%)	1.001 (68.2%)
VMD-ADE-GRU	3.5% (2.9%)	**0.194** (−0.5%)	0.255 (2.0%)	14.7% (69.4%)	0.649 (70.1%)	0.848 (70.0%)	8.9% (47.2%)	0.358 (41.1%)	0.558 (45.2%)	4.4% (4.5%)	0.198 (−18.7%)	0.258 (−23.3%)
VMD-ADE-BPNN	4.7% (27.7%)	0.251 (22.3%)	0.315 (20.6%)	21.4% (79.0%)	0.924 (79.0%)	1.195 (78.4%)	11.8% (60.2%)	0.582 (63.7%)	1.001 (69.4%)	17.9% (76.5%)	1.093 (78.5%)	1.678 (81.0%)
VMD-DE-TFT	4.8% (29.2%)	0.252 (22.6%)	0.324 (22.8%)	6.4% (29.7%)	0.250 (22.4%)	0.316 (18.4%)	9.4% (50.0%)	0.347 (39.2%)	0.547 (44.1%)	6.9% (37.7%)	0.359 (34.5%)	0.491 (35.2%)
VMD-GA-TFT	7.0% (51.4%)	0.327 (40.4%)	0.425 (41.2%)	8.9% (49.4%)	0.377 (48.5%)	0.470 (45.1%)	8.3% (43.4%)	0.369 (42.8%)	0.645 (52.6%)	4.3% (2.3%)	**0.196** (−19.9%)	**0.255** (−24.7%)
EMD-ADE-TFT	6.1% (44.3%)	0.316 (38.3%)	0.391 (36.1%)	10.4% (56.7%)	0.422 (54.0%)	0.546 (52.7%)	9.3% (49.5%)	0.451 (53.2%)	0.726 (57.9%)	16.7% (74.3%)	0.874 (73.1%)	1.205 (73.6%)
EEMD-ADE-TFT	6.5% (47.7%)	0.358 (45.5%)	0.448 (44.2%)	11.6% (61.2%)	0.477 (59.3%)	0.604 (57.3%)	9.4% (50.0%)	0.420 (49.8%)	0.708 (56.8%)	11.1% (61.3%)	0.639 (63.2%)	0.899 (64.6%)
The persistence model	20.3% (83.3%)	1.095 (82.2%)	1.436 (82.6%)	22.2% (79.7%)	0.994 (80.5%)	1.297 (80.1%)	23.0% (79.6%)	0.959 (78.0%)	1.410 (78.3%)	23.0% (81.3%)	1.089 (78.4%)	1.521 (79.1%)
VMD-ADE-TFT (without future meteorological data)	8.2% (58.5%)	0.449 (56.6%)	0.539 (53.6%)	18.3% (75.4%)	0.752 (74.2%)	0.979 (73.6%)	16.3% (71.2%)	0.611 (65.5%)	0.843 (63.7%)	6.3% (31.7%)	0.306 (23.2%)	0.407 (21.9%)
VMD-ADE-TFT	**3.4**%	0.195	**0.250**	**4.5**%	**0.194**	**0.258**	**4.7**%	**0.211**	**0.306**	**4.2**%	0.235	0.318

Note: Bold values indicate the best forecasting performance. *IR* means the proposed model v.s. the current model.

of VMD-ADE-TFT without future meteorological data. This result signifies the contribution of future meteorological data to wind speed prediction. Furthermore, as shown in Table 12, the results show that VMD-ADE-TFT without future meteorological data underperforms comparatively to several other models on some datasets. For example, in terms of MAPE, the prediction performance of VMD-ADE-TFT (without future meteorological data) is worse than VMD-ADE-GRU, VMD-ADE-BPNN, VMD-DE-TFT, VMD-GA-TFT, EMD-ADE-TFT, and EEMD-ADE-TFT in the spring dataset. This further confirms the contribution of future meteorological data to wind speed prediction.

(e) In the four seasons, the MAPE of the VMD-ADE-TFT model is 3.4%, 4.5%, 4.7%, and 4.2%. Compared with other models, the proposed model has the smallest change in MAPE. This result indicates that seasonal changes affect the performance of wind speed prediction. Nevertheless, the prediction performance of the VMD-ADE-TFT model is more stable, and the prediction ability has less fluctuation.

(f) The peak wind speed forecasting performances of each model are shown in Table 13 and Fig. 13. In the summer and

Table 13
Peak wind speed forecasting test results of different models in the Five Points data set.

Model	Spring			Summer			Autumn			Winter		
	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE
VMD-ADE-LSTM	9.3%	0.879	0.956	12.6%	0.613	0.705	8.8%	1.220	1.703	**3.3%**	**0.392**	**0.403**
VMD-ADE-RNN	8.4%	0.923	1.173	14.1%	0.704	0.758	7.8%	1.102	1.551	15.0%	1.915	2.079
VMD-ADE-GRU	6.1%	0.576	0.674	19.6%	1.072	1.229	8.3%	1.156	1.437	3.4%	0.397	0.413
VMD-ADE-BPNN	4.8%	0.362	0.386	23.2%	1.248	1.524	16.0%	2.366	2.993	3.8%	5.071	5.506
VMD-DE-TFT	3.8%	0.257	0.298	8.1%	0.357	0.423	7.9%	1.198	1.651	6.3%	0.785	0.903
VMD-GA-TFT	**2.6%**	**0.199**	**0.266**	11.1%	0.607	0.616	9.9%	1.356	1.626	8.6%	1.162	1.491
EMD-ADE-TFT	3.4%	0.315	0.433	11.7%	0.529	0.586	11.2%	1.697	2.207	21.0%	2.715	3.027
EEMD-ADE-TFT	5.0%	0.441	0.490	14.5%	0.651	0.834	9.6%	1.469	2.018	22.5%	2.839	3.046
The persistence model	19.3%	1.628	2.069	35.6%	1.871	2.298	17.8%	2.457	2.655	24.4%	2.700	2.821
VMD-ADE-TFT (without future meteorological data)	3.5%	0.251	0.323	24.8%	1.244	1.344	12.0%	1.617	1.788	4.5%	0.576	0.673
VMD-ADE-TFT	3.3%	0.281	0.316	**5.0%**	**0.273**	**0.305**	**3.8%**	**0.524**	**0.644**	3.6%	0.492	0.616

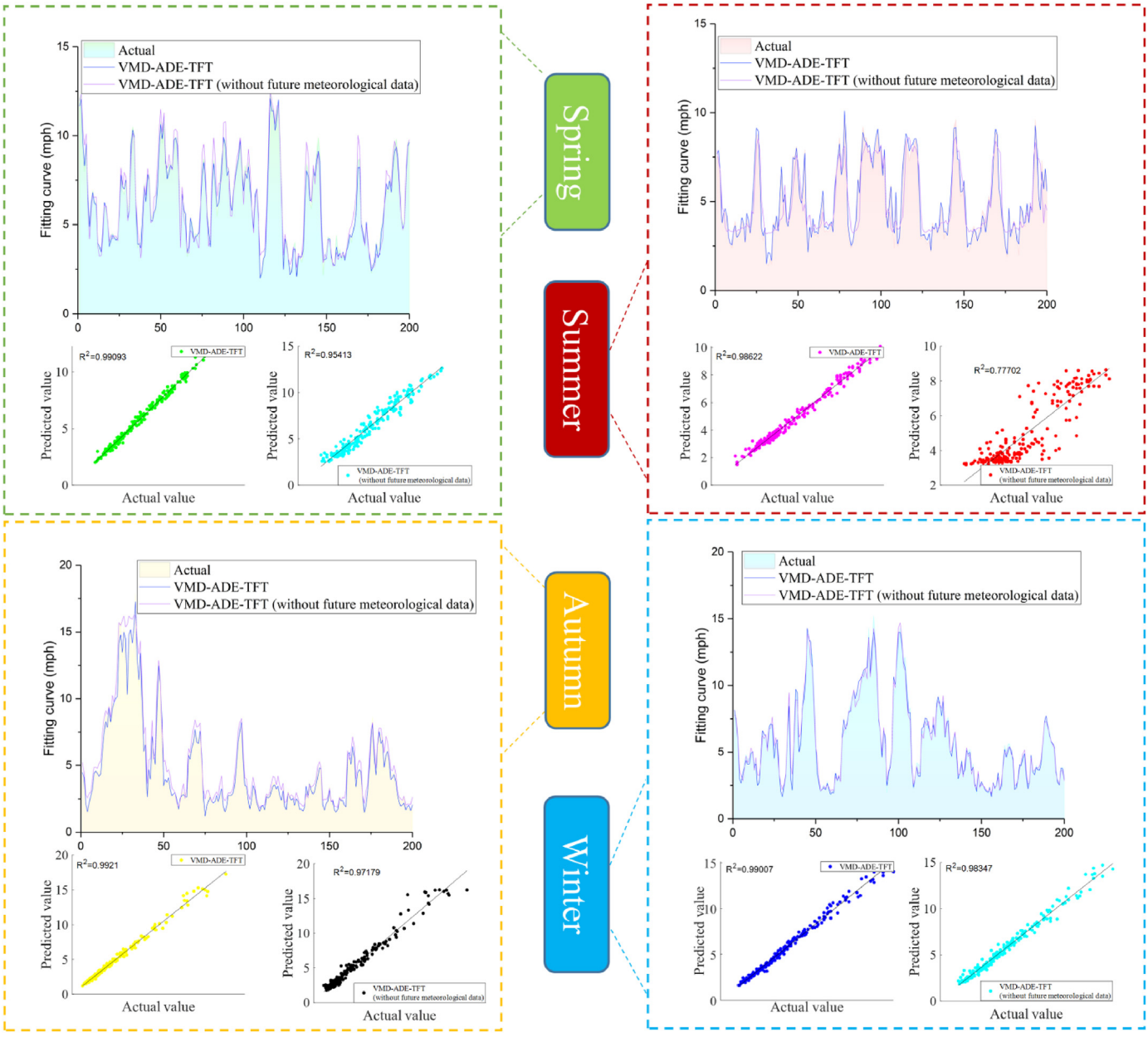

Fig. 12. Forecasting results of the VMD-ADE-TFT model with (or without) future meteorological data models in four seasons.

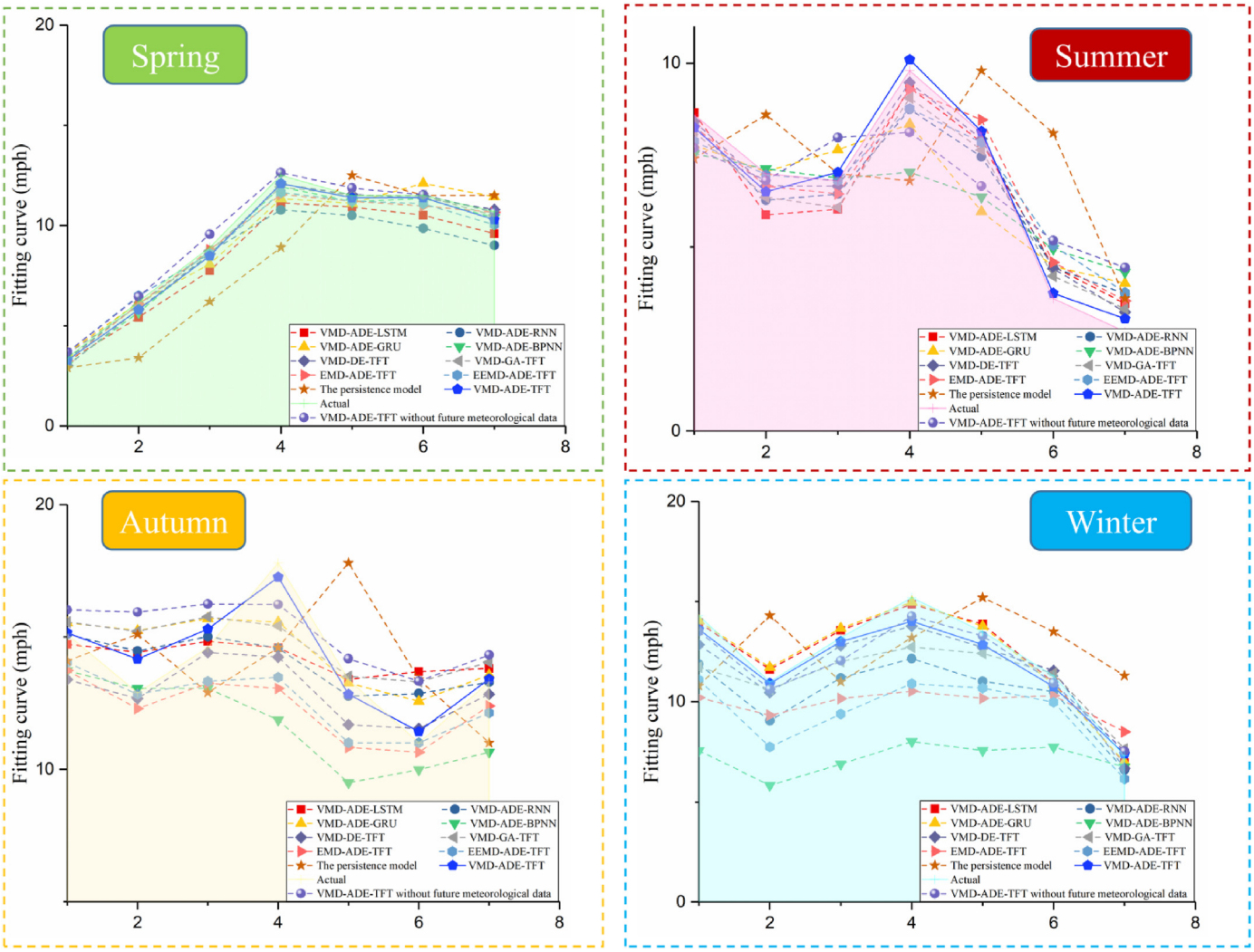

Fig. 13. Peak wind speed forecasting test results of different models in four seasons.

autumn data set, the error values of VMD-ADE-TFT are smaller than the above models. Meanwhile, in the spring and winter data set, the peak wind speed forecasting performance of VMD-ADE-TFT ranks in the top three among all models. Thus, the high forecasting performance of the proposed model can be better demonstrated by detecting the prediction performance during the peak wind speed periods.

The importance order of past variables, the importance order of future variables, and the attention of different lags orders are depicted in Fig. 14. The interpretable results are detailed below.

(a) Similar to the Albert data set, S2 has the largest contribution to wind speed prediction in terms of past input data. In particular, S2 is the most important variable in the winter, summer, and autumn data sets. S2 can better represent the overall trend of the original wind speed than S1. Compared with mid-frequency and high-frequency sequences, low-frequency ones play a greater role in forecasting wind speed prediction. In the summer and winter data sets, vapor pressure also contributes to wind speed prediction. This result suggests that meteorological change in Five Points affects wind speed more than that in Albert.

(b) Unlike the Albert data set, the most important future variable in the Five Points data set is the wind direction in the spring, summer, and autumn data sets. Geographic location significantly influences the prediction of wind speed. Air temperature is another important future variable in the four data sets. Vapor pressure and relative humidity are moderately important variables, whereas dew point temperature and soil temperature have a low contribution. Time variables (i.e., hour, day, month, time index, and relative time index) has a lower contribution to the prediction of wind speed. The periodicity of wind speed fluctuations in Five Points is not strong with time, and it is more affected by changes in meteorological data.

(c) The fluctuation of attention with different lag orders in this region is more severe. This is unlike the rule of the smaller lag order, the greater the attention that emerged from the Albert data set. For example, in the autumn data set, the larger the lag order, the greater the attention. Determining the appropriate lag order is crucial. This proves that ADE is required to optimize the TFT parameters proposed in this study.

Fig. 14. Interpretable results of the VMD-ADE-TFT model in four seasons.

4.2.3. Extended experiment

This study collected Five Points' wind speed and meteorological data from November 2019 to October 2020. Fig. 15 shows the comparison chart of Five Points' wind speed series in different years. The training-to-validation-to-testing ratio is 80%:10%:10%.

After several experiments, the parameters of the VMD-ADE-TFT are shown in Table 14. As shown in Table 15, The results show that the proposed model obtained the best forecasting performance in spring, autumn, and winter data sets. Meanwhile, the forecasting performance of VMD-ADE-TFT in the summer dataset is only worse than that of VMD-GA-TFT and EEMD-ADE-TFT. Overall, the extended experiment further demonstrates the effectiveness and superiority of the proposed model.

5. Conclusions

Wind power is receiving growing attention in the global renewable energy market. As wind power is mainly affected by wind speed, accurate and successful forecasting of wind speed prediction is critical to the safety of power dispatch and the normal operation of the power grid. However, wind speed forecasting is challenging given the random and intermittent characteristics of wind speed [50]. To advance the research on wind speed prediction, this study introduces a unique interpretable prediction system, namely, VMD-ADE-TFT, that accounts for historical wind speed and other meteorological data. The prediction model first uses VMD to decompose the historical wind speed series. The decomposed sub-modes are then adopted as the historical input of the TFT model. Meteorological data and time data (e.g., "month," "day," "hour," etc.) are inputted into the TFT model as future known variables. Next, the ADE algorithm is employed to optimize the parameter combination of the TFT model to enhance the performance and stability of the proposed model. Eight experiments are conducted on two cases. The experimental results indicate that the prediction system proposed in this study outperforms nearly all other comparable models in different indicators. The stability and accuracy of VMD-ADE-TFT in forecasting wind speed are satisfactory.

The interpretable results are analyzed in connection with the importance order of past variables, the importance order of future variables, and the attention of different lags orders. Low-frequency

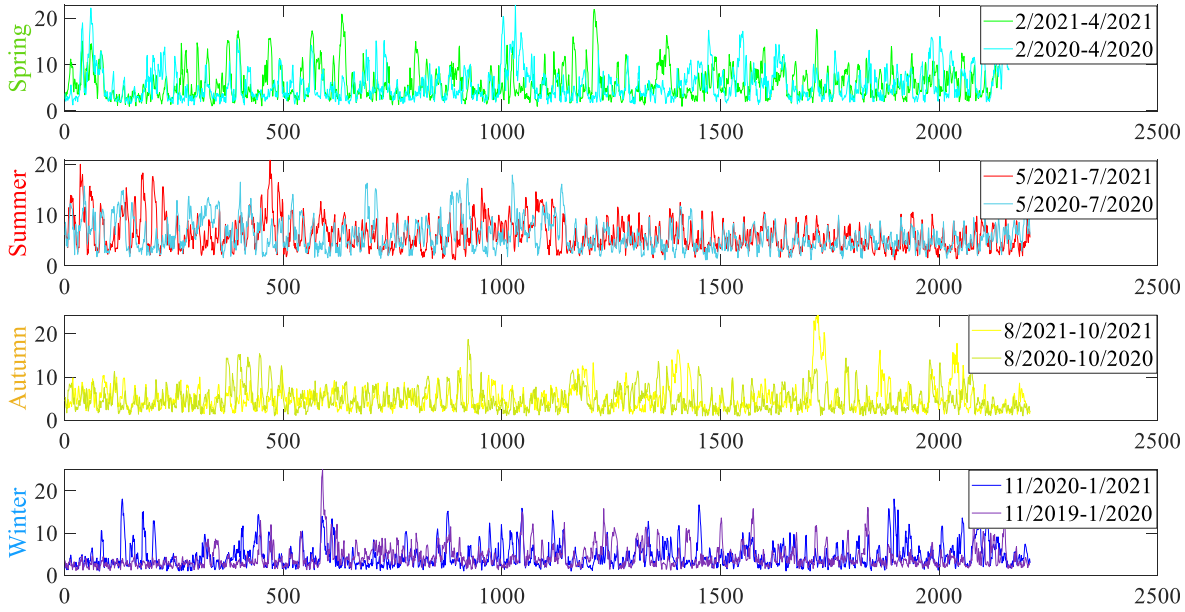

Fig. 15. Comparison chart of wind speed series in different years.

Table 14
Parameters of the VMD-ADE-TFT in the four data sets.

	Parameter	Spring	Summer	Autumn	Winter
VMD	Number of sub-modes (K)	19	22	20	18
ADE	Population size (M)	30	25	10	25
	Maximum number of iterations (T)	30	25	20	30
	Crossover probability (CR)	0.4	0.5	0.3	0.4
	Mutation operator (F)	[0,1]	[0,1]	[0,1]	[0,1]
TFT	Number of time steps	18	14	12	9
	Number of batch sizes	41	42	37	31
	Learning rates	0.032	0.029	0.009	0.013
	Number of hidden layers	12	20	19	16
	Number of attention heads	1	1	1	1
	Number of hidden layer neurons	6	9	10	8
	Dropout rate	0.1	0.1	0.1	0.1
	Max gradient norm	0.1	0.1	0.1	0.1

Table 15
Forecasting results of different models in the Five Points data set.

Model	Spring			Summer			Autumn			Winter		
	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE
VMD-ADE-LSTM	6.1%	0.350	0.436	7.2%	0.382	0.455	10.4%	0.298	0.369	7.4%	0.274	0.444
VMD-ADE-RNN	5.8%	0.356	0.457	5.3%	0.287	0.357	7.8%	0.253	0.324	6.4%	0.266	0.434
VMD-ADE-GRU	4.9%	0.281	0.361	4.8%	0.230	0.296	7.7%	0.248	0.324	6.1%	0.233	0.340
VMD-ADE-BPNN	7.8%	0.463	0.561	7.5%	0.359	0.465	11.1%	0.339	0.430	7.4%	0.293	0.444
VMD-DE-TFT	4.6%	0.282	0.336	4.6%	0.222	0.279	8.1%	0.262	0.326	5.7%	0.213	0.307
VMD-GA-TFT	5.7%	0.321	0.406	**3.4%**	**0.164**	**0.204**	9.8%	0.292	0.371	5.6%	0.211	0.305
EMD-ADE-TFT	7.3%	0.438	0.552	5.3%	0.260	0.325	12.7%	0.360	0.500	5.5%	0.217	0.326
EEMD-ADE-TFT	6.7%	0.389	0.509	4.2%	0.197	0.242	9.3%	0.274	0.361	5.4%	0.210	0.291
The persistence model	20.4%	1.204	1.669	19.7%	1.003	1.282	27.1%	0.876	1.139	23.4%	0.976	1.387
VMD-ADE-TFT (without future meteorological data)	5.8%	0.344	0.434	4.8%	0.236	0.298	8.6%	0.257	0.339	7.2%	0.279	0.430
VMD-ADE-TFT	**3.6%**	**0.233**	**0.299**	4.5%	0.211	0.278	**5.8%**	**0.196**	**0.248**	**5.2%**	**0.201**	**0.276**

Note: Bold values indicate the best forecasting performance.

and high-frequency sub-modes obtained after decomposing the original wind speed series by VMD help better in prediction than the intermediate-frequency sub-modes. In addition, the meteorological variables influencing wind speed vary in different regions, such as temperature in Albert and wind direction in Five Points. Moreover, the importance of each meteorological variable differs across seasons. Researchers can perform an interpretable analysis of specific wind speed sequences to improve prediction models. They can use credible interpretable results, such as selection and collection of data or design of specific functional engineering, as bases. Furthermore, the interpretable analysis of wind speed forecasts can provide decision-makers with reliable analysis that can

aid them in producing accurate forecasts and devising meticulous plans.

Future research can expand the scale of the data or try different timescales to consider longer time steps. Attention weight patterns can be used as well to judge the persistent temporal patterns of wind speed series (e.g., seasonal fluctuation trends). These future research directions can be meaningful in establishing trust with human experts.

Credit author statement

Binrong Wu: Conceptualization, Methodology, Software, Investigation, Writing- Original draft preparation. Lin Wang: Conceptualization, Methodology, Writing-Reviewing and Editing, Supervision, Funding acquisition. Yu-Rong Zeng: Conceptualization, Investigation, Validation, Visualization, Supervision.

Declaration of competing interest

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

Acknowledgment

This research is partially supported by the National Natural Science Foundation of China (Nos: 71771095; 71810107003).

References

[1] Khaloie H, Abdollahi A, Shafie-khah M, Anvari-Moghaddam A, Nojavan S, Siano P, et al. Coordinated wind-thermal-energy storage offering strategy in energy and spinning reserve markets using a multi-stage model. Appl Energy 2020;259:114168.
[2] Wang H, Han S, Liu Y, Yan J, Li L. Sequence transfer correction algorithm for numerical weather prediction wind speed and its application in a wind power forecasting system. Appl Energy 2019;237:1–10.
[3] Wang J, Wang S, Yang W. A novel non-linear combination system for short-term wind speed forecast. Renew Energy 2019;143:1172–92.
[4] Liu H, Chen C. Data processing strategies in wind energy forecasting models and applications: a comprehensive review. Appl Energy 2019;249:392–408.
[5] Zhao W, Wei Y-M, Su Z. One day ahead wind speed forecasting: a resampling-based approach. Appl Energy 2016;178:886–901.
[6] Hu J, Heng J, Wen J, Zhao W. Deterministic and probabilistic wind speed forecasting with de-noising-reconstruction strategy and quantile regression based algorithm. Renew Energy 2020;162:1208–26.
[7] Wang Y, Zou R, Liu F, Zhang L, Liu Q. A review of wind speed and wind power forecasting with deep neural networks. Appl Energy 2021;304:117766.
[8] Cassola F, Burlando M. Wind speed and wind energy forecast through Kalman filtering of Numerical Weather Prediction model output. Appl Energy 2012;99:154–66.
[9] Hoolohan V, Tomlin AS, Cockerill T. Improved near surface wind speed predictions using Gaussian process regression combined with numerical weather predictions and observed meteorological data. Renew Energy 2018;126:1043–54.
[10] Jung J, Broadwater RP. Current status and future advances for wind speed and power forecasting. Renew Sustain Energy Rev 2014;31:762–77.
[11] Yunus K, Thiringer T, Chen P. ARIMA-based frequency-decomposed modeling of wind speed time series. IEEE Trans Power Syst 2016;31:2546–56.
[12] Kavasseri RG, Seetharaman K. Day-ahead wind speed forecasting using f-ARIMA models. Renew Energy 2009;34:1388–93.
[13] Maatallah OA, Achuthan A, Janoyan K, Marzocca P. Recursive wind speed forecasting based on Hammerstein Auto-Regressive model. Appl Energy 2015;145:191–7.
[14] Chen Y, Zhang S, Zhang W, Peng J, Cai Y. Multifactor spatio-temporal correlation model based on a combination of convolutional neural network and long short-term memory neural network for wind speed forecasting. Energy Convers Manag 2019;185:783–99.
[15] Fu W, Wang K, Tan J, Zhang K. A composite framework coupling multiple feature selection, compound prediction models and novel hybrid swarm optimizer-based synchronization optimization strategy for multi-step ahead short-term wind speed forecasting. Energy Convers Manag 2020;205:112461.
[16] Liu H, Tian H, Pan D, Li Y. Forecasting models for wind speed using wavelet, wavelet packet, time series and Artificial Neural Networks. Appl Energy 2013;107:191–208.
[17] Wan J, Liu J, Ren G, Guo Y, Yu D, Hu Q. Day-Ahead prediction of wind speed with deep feature learning. Int J Pattern Recogn Artif Intell 2016;30:1650011.
[18] Wu J, Li N, Zhao Y, Wang J. Usage of correlation analysis and hypothesis test in optimizing the gated recurrent unit network for wind speed forecasting. Energy 2022;242:122960.
[19] Shi J, Guo J, Zheng S. Evaluation of hybrid forecasting approaches for wind speed and power generation time series. Renew Sustain Energy Rev 2012;16:3471–80.
[20] Jiang P, Liu Z, Niu X, Zhang L. A combined forecasting system based on statistical method, artificial neural networks, and deep learning methods for short-term wind speed forecasting. Energy 2021;217:119361.
[21] Song J, Wang J, Lu H. A novel combined model based on advanced optimization algorithm for short-term wind speed forecasting. Appl Energy 2018;215:643–58.
[22] Zhang J, Wei Y, Tan Z. An adaptive hybrid model for short term wind speed forecasting. Energy 2020;190:115615.
[23] Tascikaraoglu A, Sanandaji BM, Poolla K, Varaiya P. Exploiting sparsity of interconnections in spatio-temporal wind speed forecasting using Wavelet Transform. Appl Energy 2016;165:735–47.
[24] Moreno SR, Coelho L, dos S. Wind speed forecasting approach based on singular spectrum analysis and adaptive neuro fuzzy inference system. Renew Energy 2018;126:736–54.
[25] da Silva RG, Dal Molin Ribeiro MH, Moreno SR, Mariani VC, Coelho L dos S. A novel decomposition-ensemble learning framework for multi-step ahead wind energy forecasting. Energy 2021;216:119174.
[26] Xiao L, Shao W, Jin F, Wu Z. A self-adaptive kernel extreme learning machine for short-term wind speed forecasting. Appl Soft Comput 2021;99:106917.
[27] Chen Y, Dong Z, Wang Y, Su J, Han Z, Zhou D, et al. Short-term wind speed predicting framework based on EEMD-GA-LSTM method under large scaled wind history. Energy Convers Manag 2021;227:113559.
[28] Hu Y-L, Chen L. A nonlinear hybrid wind speed forecasting model using LSTM network, hysteretic ELM and Differential Evolution algorithm. Energy Convers Manag 2018;173:123–42.
[29] Neshat M, Nezhad MM, Abbasnejad E, Mirjalili S, Tjernberg LB, Garcia DA, et al. A deep learning-based evolutionary model for short-term wind speed forecasting: a case study of the Lillgrund offshore wind farm. Energy Convers Manag 2021;236:114002.
[30] Dong Y, Wang J, Xiao L, Fu T. Short-term wind speed time series forecasting based on a hybrid method with multiple objective optimization for non-convex target. Energy 2021;215:119180.
[31] Shang Z, He Z, Chen Y, Chen Y, Xu M. Short-term wind speed forecasting system based on multivariate time series and multi-objective optimization. Energy 2022;238:122024.
[32] Dragomiretskiy K, Zosso D. Variational mode decomposition. IEEE Trans Signal Process 2014;62:531–44.
[33] Liu Y, Yang C, Huang K, Gui W. Non-ferrous metals price forecasting based on variational mode decomposition and LSTM network. Knowl Base Syst 2020;188:105006.
[34] Sharda S, Singh M, Sharma K. RSAM: robust self-attention based multi-horizon model for solar irradiance forecasting. IEEE Trans Sustain Energy 2021;12:1394–405.
[35] Lim B, Arik SO, Loeff N, Pfister T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. Int J Forecast 2021;37:1748–64.
[36] Neshat M, Nezhad MM, Abbasnejad E, Mirjalili S, Groppi D, Heydari A, et al. Wind turbine power output prediction using a new hybrid neuro-evolutionary method. Energy 2021;229:120617.
[37] Liu H, Chen C. Multi-objective data-ensemble wind speed forecasting model with stacked sparse autoencoder and adaptive decomposition-based error correction. Appl Energy 2019;254:113686.
[38] Zhang C, Zhou J, Li C, Fu W, Peng T. A compound structure of ELM based on feature selection and parameter optimization using hybrid backtracking search algorithm for wind speed forecasting. Energy Convers Manag 2017;143:360–76.
[39] Wang X, Yu Q, Yang Y. Short-term wind speed forecasting using variational mode decomposition and support vector regression. J Intell Fuzzy Syst 2018;34:3811–20.
[40] Zhong S, Xie X, Lin L, Wang F. Genetic algorithm optimized double-reservoir echo state network for multi-regime time series prediction. Neurocomputing 2017;238:191–204.
[41] Fu CM, Jiang C, Chen GS, Liu QM. An adaptive differential evolution algorithm with an aging leader and challengers mechanism. Appl Soft Comput 2017;57:60–73.
[42] Abd Elaziz M, Li L, Jayasena KPN, Xiong S. Multiobjective big data optimization based on a hybrid salp swarm algorithm and differential evolution. Appl Math Model 2020;80:929–43.
[43] Zhu L, Lian C, Zeng Z, Su Y. A broad learning system with ensemble and classification methods for multi-step-ahead wind speed prediction. Cogn Comput 2020;12:654–66.
[44] Liu X, Lin Z, Feng Z. Short-term offshore wind speed forecast by seasonal ARIMA-A comparison against GRU and LSTM. Energy 2021;227:120492.
[45] Emeksiz C, Tan M. Multi-step wind speed forecasting and Hurst analysis using novel hybrid secondary decomposition approach. Energy 2022;238:121764.
[46] Qu Z, Mao W, Zhang K, Zhang W, Li Z. Multi-step wind speed forecasting based on a hybrid decomposition technique and an improved back-

propagation neural network. Renew Energy 2019;133:919–29.
[47] Wu BR, Wang L, Wang SR, Zeng YR. Forecasting the U.S. oil markets based on social media information during the COVID-19 pandemic. Energy 2021;226:120403.
[48] Lv SX, Wang L. Deep learning combined wind speed forecasting with hybrid time series decomposition and multi-objective parameter optimization. Appl Energy 2022;311:118674.
[49] Zhang S, Chen Y, Xiao J, Zhang W, Feng R. Hybrid wind speed forecasting model based on multivariate data secondary decomposition approach and deep learning algorithm with attention mechanism. Renew Energy 2021;174:688–704.
[50] Hu HL, Wang L, Lv SX. Forecasting energy consumption and wind power generation using deep echo state network. Renew Energy 2020;154:598–613.