eng

MDPI

Article

Short-Term Wind Direction Forecasting Based on VMD-Transformer with Gated Residual Compensation

Yi Lu [1], Zhishuo Liu [1], Tingyu Yan [1], Dunhui Xiao [2] and Xin Jin [1,*]

[1] College of Electronic Engineering, Chengdu University of Information Technology, Chengdu 610225, China; m15082892869@163.com (Y.L.); 18328575197@163.com (Z.L.); yty@cuit.edu.cn (T.Y.)
[2] School of Mathematical Sciences, Key Laboratory of Intelligent Computing and Applications, Ministry of Education, Tongji University, Shanghai 200092, China; xiaodunhui@tongji.edu.cn
* Correspondence: jinxin1988@cuit.edu.cn

Abstract

Wind direction time series exhibit angular periodic discontinuity, multi-scale non-stationary fluctuations and abrupt wind shifts, which hinder the precision of short-term forecasting for wind turbine yaw control. In this paper, a dual-branch forecasting framework based on Variational Mode Decomposition (VMD) and Transformer is developed to address the above drawbacks, with a hysteresis gating and zoning residual compensation module embedded for targeted error correction. First, sine–cosine encoding is adopted to eliminate the numerical discontinuity between 0° and 360° for wind direction angular data, and valid meteorological input features are screened to discard redundant covariates. Second, the sine–cosine-encoded wind direction sequence is decomposed into multiple band-limited intrinsic mode functions (IMFs) via VMD, extracting frequency-specific features that reduce non-stationarity and facilitate subsequent Transformer modeling. The standard Transformer encoder serves as the normal branch to capture long-range temporal dependencies across the whole time series, while a lightweight multilayer perceptron (MLP) constitutes the compensation branch to learn prediction deviations between baseline predictions and ground-truth values. The hysteresis gating unit activates residual compensation based on historical prediction errors and angular variation, without requiring access to the current ground-truth value, and compensation intensity is adaptively adjusted via the zoning strategy; relevant coefficients are optimized by random search. Verified on a real wind farm dataset consisting of 10,421 15 min sampling points, the proposed model achieves the lowest MAE of 9.64° among six benchmark models. For the improved genuine mutation samples (angle change > 70°), the model achieves a mean improvement of 6.02°. Ablation experiments verify that VMD preprocessing, the MLP compensation branch, and the hysteresis gating mechanism play indispensable roles in forecasting performance. The proposed framework can support accurate yaw control of wind turbines, and the decomposition–compensation workflow can also be generalized to other periodic non-stationary forecasting tasks.

Keywords: wind direction forecasting; Variational Mode Decomposition; Transformer; residual compensation; gating mechanism; residual learning; wind farm

Check for updates

Academic Editors: Liércio André Isoldi and Mateus Das Neves Gomes

Received: 25 July 2026
Revised: 25 August 2026
Accepted: 31 August 2026
Published: 2 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Global energy transition has promoted the large-scale deployment of wind power, making wind energy one of the most widely applied and commercially viable renewable

energy sources [1]. Stochastic and volatile wind characteristics, especially frequent abrupt wind direction shifts, bring continuous operational risks to wind turbine groups. As a key actuator of horizontal-axis wind turbines, the yaw system adjusts the nacelle orientation to align with incoming wind flow, maximizing wind energy capture. Even minor yaw deviations can reduce power generation efficiency and increase turbine mechanical fatigue loads. Accurate short-term wind direction forecasting therefore provides reliable ahead-of-time guidance for closed-loop yaw control, helping improve wind farm energy yield and extend equipment service life. Joint forecasting of wind speed and direction has been recognized as an essential prerequisite for future-aware wind farm control operations [2].

Existing wind direction forecasting approaches are mainly divided into statistical methods and data-driven learning methods, both of which show evident limitations in processing volatile wind sequences. Traditional statistical models such as ARIMA fail to capture strong nonlinear wind variation characteristics [3]. Classical machine learning methods including SVM and BP neural networks improve nonlinear fitting ability but cannot effectively model long-range temporal dependencies in non-stationary wind series [4]. Although modern deep learning models achieve better overall prediction performance, they still suffer from gradient vanishing problems, tedious hyperparameter tuning, and insufficient sensitivity to abrupt wind mutations [5].

Signal decomposition and self-attention mechanisms have proven effective for time-series forecasting. Variational Mode Decomposition (VMD) separates complex non-stationary sequences into stable sub-components and reduces modeling difficulty [6]. Recent studies have demonstrated the effectiveness of combining VMD with advanced neural networks for wind speed forecasting [7]. The Transformer model, equipped with self-attention architecture, excels at capturing global long-range temporal correlations, providing a promising solution for meteorological sequence prediction [8]. Transformer-based models have also been adapted for short-term wind speed forecasting with promising results [9]. Recent advances in multivariate decomposition and Temporal Fusion Transformers have further improved the interpretability of wind forecasting models. The combination of signal decomposition techniques and Transformer architectures has also been explored in recent wind energy forecasting studies [10].

Nevertheless, current wind direction forecasting frameworks still face three key practical bottlenecks. First, few models can effectively handle the inherent angular discontinuity, multi-scale fluctuating features, and sudden mutation characteristics of wind direction data. Second, most existing architectures cannot simultaneously extract global long-term trends and local abrupt variation features. Third, there is a lack of an adaptive and stable end-to-end forecasting design dedicated to wind direction mutation correction.

Recent studies have explored dual-branch architectures for wind power forecasting, demonstrating the effectiveness of separate feature extraction pathways [11]. Dual-path networks with adaptive compensation have also been validated for multivariate time-series forecasting [12]. This study constructs a dual-branch VMD-Transformer framework embedded with a hysteresis gating and zoning residual compensation module. Random Forest (RF) is first adopted for feature screening to eliminate redundant meteorological covariates. VMD then decomposes the sine–cosine-encoded wind direction sequences into multiple band-limited intrinsic mode functions (IMFs). It extracts frequency-specific features from these sub-components to reduce sequence non-stationarity and ease subsequent Transformer modeling. The standard Transformer branch captures global long-range temporal dependencies from the decomposed components. A compact multilayer perceptron (MLP) serves as the compensation branch to specifically rectify large local prediction deviations induced by abrupt wind shifts via residual learning [13].

Unlike traditional fixed compensation schemes, the embedded hysteresis gating mechanism only triggers residual correction under severe wind mutation conditions to prevent over-compensation during steady wind phases. Thanks to the coordinated optimization of global trend fitting and targeted error rectification for abrupt shifts, the proposed framework achieves balanced forecasting performance under both stable and volatile wind conditions. High-quality wind-field observation data is the foundation of reliable wind-direction prediction, and virtual sensing techniques provide an alternative way to supplement field-measured wind information [14]. Experiments on a real wind farm dataset containing 10,421 15 min samples verify that the proposed model with RF-based feature screening outperforms Pure Transformer, VMD-Transformer, and other mainstream baselines, achieving the lowest test MAE of 9.64° [15]. For the improved genuine mutation samples (angle change >70°), the model achieves a mean improvement of 6.02°. The proposed framework provides a feasible intelligent yaw control solution for practical wind farm operation, and the decomposition–compensation dual-branch structure is also generalizable to other periodic non-stationary time-series forecasting tasks, such as short-term power load prediction [16].

2. Materials and Methods

2.1. Random Forest Algorithm

As a mainstream ensemble machine learning algorithm suitable for classification and regression tasks, RF features easy implementation and high flexibility. In contrast to conventional statistical models, RF achieves superior predictive performance, mitigates overfitting risks, and captures complex interactions among predictor variables. As an improved variant of bagging, RF trains a set of uncorrelated decision trees by simultaneously adopting bootstrap row sampling and replacement random feature sampling; although each individual decision tree has high variance, RF averages the output results of all trees to obtain low-variance prediction results. As theoretically established by Breiman, stable regression performance of such an ensemble requires weak residual correlation and low single-tree prediction bias. For the wind direction prediction task in this work, RF was only utilized for data preprocessing instead of direct angle forecasting and fulfilled two core functions: (i) combined with sine–cosine encoding to eliminate the inherent circular discontinuity of wind direction angles; (ii) quantified feature importance through reductions in the Gini index to screen influential meteorological predictors and reduce input dimensionality.

The RF training workflow was implemented as follows: (i) equal-sized T bootstrap samples were extracted from the raw training set D; (ii) a separate decision tree was trained on each sample; (iii) during each node split, $m = \sqrt{M}$ candidate features (where M is the total number of candidate predictors) were randomly chosen to locate the optimal split threshold; (iv) the outputs of all decision trees were averaged to produce the final regression outputs. All data-processing procedures were implemented in MATLAB (MathWorks, Natick, MA, USA, R2023a). Deep-learning network training was carried out using PyCharm (JetBrains, Prague, Czech Republic, 2024.1).

RF quantified the contribution of each feature by summing the reductions in the Gini index over all split nodes. The global importance score of feature X_j is defined as follows:

$$V_{\mathrm{IM}j}^{(\mathrm{Gini})} = \sum_{i=1}^{T} \sum_{m \in M_{ij}} V_{\mathrm{IM}jm}^{(\mathrm{Gini})} \tag{1}$$

where $V_{\mathrm{IM}j}^{(\mathrm{Gini})}$ is the global Gini importance of feature X_j, T is the number of trees, and $V_{\mathrm{IM}jm}^{(\mathrm{Gini})}$ denotes the Gini reduction at node m when splitting on X_j, with M_{ij} being the set of such nodes on the i-th tree.

The effectiveness of RF-based feature selection has been validated in wind turbine fault diagnosis and predictive maintenance applications [17]. Hybrid wind direction forecasting algorithms have also been developed using data-analysis-based approaches, further confirming the utility of feature selection in wind direction forecasting [18].

2.2. Variational Mode Decomposition

After key auxiliary features were screened via RF, the raw wind direction sequence still presented obvious non-stationarity and multi-scale fluctuations. Directly inputting raw data would raise modeling difficulty; thus, VMD was introduced to preprocess wind direction data [6].

VMD is an advanced signal processing method recognized for its reliable decomposition performance. In signal analysis, VMD adaptively searches for optimal central frequencies and reduces the bandwidth of each mode, separating band-limited intrinsic mode functions (IMFs) and realizing frequency-domain partitioning. Compared with empirical mode decomposition and wavelet analysis, VMD offers better noise immunity and more stable frequency-domain decomposition. As an adaptive non-recursive signal decomposition algorithm, it decomposes complex non-stationary and nonlinear signals into several band-limited IMFs with independent central frequencies. Recent studies have demonstrated the effectiveness of combining VMD with advanced neural networks for short-term wind direction forecasting [7]. The constrained variational formulation of VMD is given by Equation (2).

$$\begin{cases} \min\limits_{\{u_k\},\{\omega_k\}} \left\{ \sum\limits_{k=1}^{K} \left\| \partial_t \left[\left(\delta(t) + \dfrac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \\ \text{s.t.} \quad \sum\limits_{k=1}^{K} u_k(t) = f(t) \end{cases} \tag{2}$$

In Equation (2), $f(t)$ denotes the preprocessed raw wind direction time series, and K the total number of band-limited intrinsic mode functions decomposed from the original signal. The k-th decomposed mode component is expressed as $u_k(t)$ with corresponding central frequency ω_k. $\delta(t)$ stands for the Dirac delta function; the symbol * denotes the convolution operator, and ∂_t represents the partial derivative with respect to time. j is the imaginary unit in the complex frequency domain; $\lambda(t)$ is the time-domain Lagrange multiplier for the constrained optimization, and α is the bandwidth penalty coefficient controlling the spectral width of each IMF component.

To solve this constrained variational optimization problem, VMD adopts an iterative updating strategy. The mode functions, corresponding central frequencies, and Lagrange multiplier are iteratively optimized via the recursive formulas shown in Equations (3)–(5) until the algorithm meets the preset convergence threshold.

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

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

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

where $\hat{u}_k(\omega)$ is the Fourier spectrum of the k-th IMF component; $\hat{f}(\omega)$ represents the Fourier transform of the target wind direction series; $\hat{\lambda}(\omega)$ denotes the Fourier spectrum of

the Lagrange multiplier; τ is the step size for updating the Lagrange multiplier; n indicates the current iteration index of the alternating direction method of multipliers (ADMM). The convergence criterion for VMD is defined in Equation (6):

$$\sum_{k=1}^{K} \frac{\|\hat{u}_k^{n+1} - \hat{u}_k^n\|_2^2}{\|\hat{u}_k^n\|_2^2} < \varepsilon \tag{6}$$

where ϵ is the preset tolerance threshold.

Before decomposition, raw wind angle data was converted via sine–cosine encoding to remove periodic discontinuity. Each transformed sequence was decomposed into four IMFs; eight IMF subsequences were obtained after concatenating sine and cosine branches. All IMF components were merged into one feature vector, which enabled the Transformer to extract multi-scale temporal features and improve prediction accuracy.

In this study, VMD is employed as a multi-scale feature extractor rather than a signal reconstruction module: the Transformer operates directly on the decomposed IMF subsequences, and the reconstructed signal is not used as an intermediate input. While reconstruction fidelity is reported in Section 3.2.1 for reference, the predictive validity of the IMF features is independently verified by the ablation study in Section 3.2.5, where removing VMD increases the test MAE by 67.0%. This confirms that the frequency-domain information captured by VMD carries substantial predictive value, despite the moderate reconstruction deviation.

2.3. Transformer

The multi-scale subseries decomposed via Variational Mode Decomposition (VMD) exhibit complex long-range temporal dependencies and nonlinear cross-scale coupling characteristics. Traditional recurrent neural networks (RNNs) and convolutional neural networks (CNNs) are incapable of simultaneously modeling such multi-frequency feature information [19]. Benefiting from the self-attention mechanism, the Transformer model can capture pairwise dependency relationships between arbitrary time steps within a sequence, which renders it applicable for processing VMD-based multi-scale subseries.

As a prevalent deep learning architecture based on the self-attention mechanism, the Transformer implements sequence-to-sequence prediction mapping through an encoder–decoder framework. Its core components include multi-head self-attention (MHSA), positional encoding, and feed-forward networks (FFNs), which collaboratively extract global trend characteristics and local abrupt fluctuation features of time-series data [8]. For the wind direction prediction task in this study, the Transformer constructs a nonlinear mapping between historical and future wind direction observations and serves as the core feature extraction and prediction backbone. Recent studies have demonstrated the effectiveness of dual-branch architectures for wind power forecasting [11]. To further improve prediction accuracy, this paper designs a dual-branch prediction structure integrated with a gated fusion module. The proposed structure independently models low-frequency trend components and high-frequency fluctuation components and adaptively fuses multi-scale feature outputs to achieve high-precision short-term wind direction forecasting. Transformer-based models have shown competitive performance in multi-step wind power forecasting tasks [20].

The self-attention mechanism constitutes the fundamental component of the Transformer architecture. It calculates the correlation weight between each pair of time steps to dynamically aggregate global contextual information of time series. To enhance the feature representation capability of the model, the multi-head self-attention (MHSA) mechanism projects input features into multiple independent latent subspaces, performs parallel attention calculation in each subspace, and finally concatenates the parallel outputs and executes

linear transformation for feature fusion. The specific mathematical expressions are defined in Equations (7) and (8):

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O \quad (7)$$

$$\text{head}_i = \text{Attention}(QW_{i1}, KW_{i2}, VW_{i3}) \quad (8)$$

where h denotes the total number of attention heads; Q, K and V represent the query, key matrix, and value matrix of the input meteorological time series, respectively; W_{i1}, W_{i2} and W_{i3} denote the projection matrices that map original input features to the i-th latent subspace; W_O denotes the output linear matrix for fusing multi-view attention features. For VMD-decomposed subseries, the multi-head structure enables the model to synchronously extract low-frequency trend components and high-frequency fluctuation components, thereby enriching the diversity and completeness of temporal feature extraction.

Unlike RNNs with recursive temporal propagation structures, the vanilla Transformer lacks inherent sequential order perception capability. Positional encoding is therefore embedded into input feature embeddings to retain the temporal order information of time-series data. The Transformer encoder is composed of N stacked identical layers, and each layer contains an MHSA sublayer and an FFN sublayer. Residual connections and layer normalization are configured after each sublayer to optimize model training efficiency and stability. The Transformer decoder adopts masked multi-head self-attention to mask future time step information, which effectively avoids data leakage in the prediction process. In this work, the encoder is employed to extract fine-grained temporal features from each VMD-decomposed subseries, and the decoder completes multi-step wind direction prediction based on the extracted multi-scale encoder features.

2.4. Construction of the Proposed VMD-Transformer with Gated Error Compensation

2.4.1. Overall Architecture and Workflow

The overall workflow of the proposed VMD-Transformer with Gated Error Compensation consists of four core stages: preprocessing and feature selection, normal branch prediction, compensation branch training, and gated error compensation, as illustrated in Figure 1. The detailed procedures are as follows:

a. Preprocessing and feature selection: Sine–cosine encoding is applied to wind direction angles to address the circular periodicity problem. RF is employed to select the top 30 auxiliary features. VMD is performed on the sine and cosine components of the encoded wind direction signal, respectively. Each sine–cosine-encoded signal is decomposed into four IMFs, generating eight IMF subsequences in total. These eight IMF subsequences are concatenated with the 30 screened auxiliary features to construct a 38-dimensional input vector. Similar VMD-based decomposition and feature construction strategies have been successfully applied in short-term wind power prediction [21].

b. Normal branch prediction: A vanilla Transformer encoder forms the normal branch to extract global temporal trends. This branch captures long-term dependencies and periodic characteristics within wind direction series. The network adopts a fixed complete time window and is trained on all training samples to produce baseline predictions. Transformer-based models have shown competitive performance in multi-step wind power forecasting tasks [20].

c. Compensation branch training: Based on prediction errors yielded by the normal branch on the training set, error jump samples are screened using the following criteria: the error at the preceding time step e_{t-2} is less than 5°, the error at the previous time step e_{t-1} exceeds 10°, and the error jump amplitude is larger than 8°. A lightweight MLP serves as the compensation branch, which learns to fit the prediction bias of the normal branch

via residual loss optimization. Dual-path networks with adaptive compensation have been validated for multivariate time-series forecasting [12].

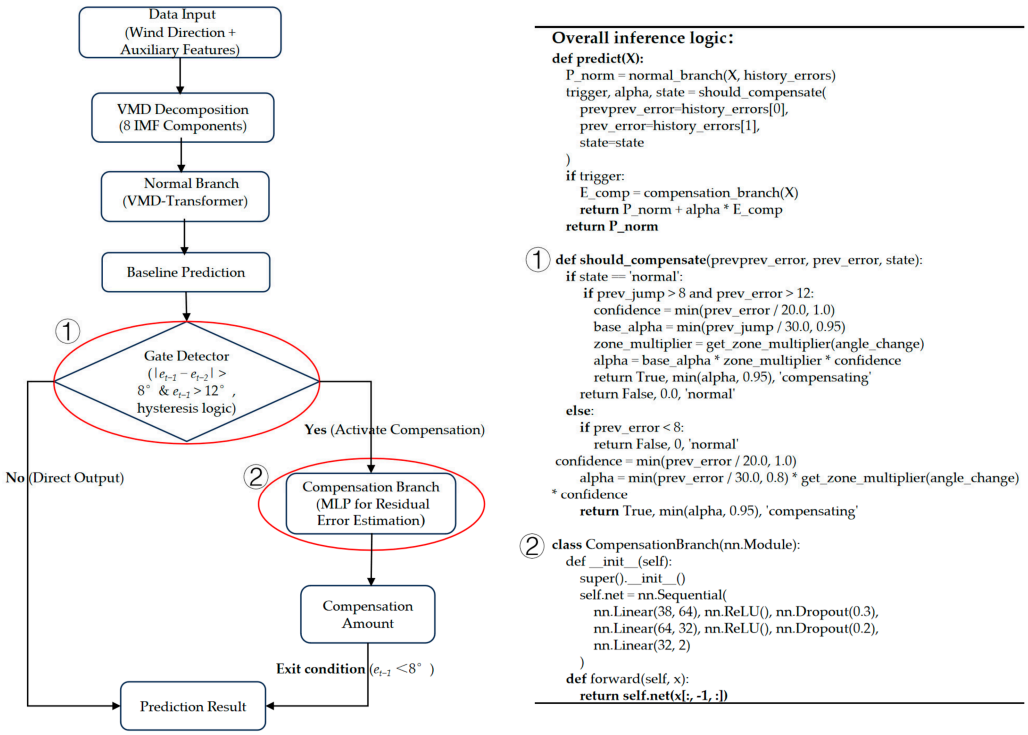


Figure 1. Flowchart of the proposed VMD-Transformer with Gated Error Compensation for short-term wind direction forecasting.

d. Gated error compensation: During inference, the hysteresis gate determines whether to activate residual compensation according to prediction error metrics. The triggering rules are defined as follows.

Under the normal state, compensation is activated when the error jump from the preceding step $|e_{t-1} - e_{t-2}|$ exceeds 8° and the prediction error at the previous time step e_{t-1} exceeds 12°.

Once the gate switches to the compensating state, residual compensation remains enabled until the prediction error at the previous time step e_{t-1} drops below 8°.

The overall forward inference procedure implemented by the predict function follows the pipeline illustrated in Figure 1. When compensation is triggered, the weighted residual error output from the compensation branch is superimposed on the baseline prediction of the normal branch. Otherwise, the baseline prediction is directly adopted as the final forecasting result. Dynamic gating strategies have also been explored in advanced Transformer architectures for wind power monitoring [22].

2.4.2. Data Preprocessing and VMD Configuration

Wind direction data used in this study were collected from a wind farm in Sichuan Province, provided by the open dataset of the State Grid Renewable Energy Forecasting Competition [15], with a sampling interval of 15 min. A total of 10,421 valid samples were retained and divided into three independent continuous time periods: training, validation, and test sets, as summarized in Table 1.

Table 1. Partition statistics of the 15 min wind direction dataset.

Dataset	Sample Count	Proportion	Start Date	End Date	Duration (Days)
Training Set	6410	61.5%	May 11	July 17	68
Validation Set	2095	20.1%	Aug 22	Sep 13	23
Test Set	1916	18.4%	Sep 14	Oct 4	22

Wind direction measurements belong to angular data with inherent periodicity and circular continuity. Direct modeling using raw angular values ranging from 0° to 360° introduces numerical discontinuity defects: on a circular scale, 359° and 1° are only 2° apart, whereas their direct numerical difference reaches 358°. Such inconsistency breaks the continuity of data distribution and increases the difficulty of model training. To address this issue, trigonometric encoding maps angular values onto a continuous two-dimensional space [18]. Let θ denote the raw wind direction angle (in radians); its sine–cosine encoding is defined as:

$$x_{\sin} = \sin\left(\frac{2\pi\theta}{360}\right), x_{\cos} = \cos\left(\frac{2\pi\theta}{360}\right) \tag{9}$$

After encoding, the original wind direction angle series is transformed into two orthogonal time series x_{sin} and x_{cos}. Following model prediction, outputs are converted back to angular values using the four-quadrant arctangent function:

$$\hat{\theta} = \arctan2(\hat{x}_{\sin}, \hat{x}_{\cos}) \times \frac{180^\circ}{\pi} \tag{10}$$

where $\hat{x}_{\sin}$ and $\hat{x}_{\cos}$ are the predicted sine and cosine values, and $\hat{\theta}$ is the recovered wind direction angle.

A feature selection strategy based on feature importance is adopted for wind direction forecasting [17]. Specifically, an RF model is trained to quantify the importance of all candidate features. All features are sorted in descending order of importance, and the top 30 features with the highest importance scores are retained. Correlation analysis is further carried out on the screened features to alleviate multicollinearity. Ultimately, 30 core auxiliary features are preserved, covering sine–cosine-encoded wind direction, multi-scale lag features, sliding average features, temporal periodic features and meteorological variables. Z-score standardization is uniformly applied to all 30 auxiliary features:

$$z_i = \frac{x_i - \mu_i}{\sigma_i}, i = 1, 2, \ldots, 30 \tag{11}$$

where x_i is the original value of the i-th feature, μ_i and σ_i are its mean and standard deviation computed over the training set, and z_i is the standardized result.

An adaptive parameter tuning strategy is applied for the sine–cosine-encoded components. We evaluated K from 2 to 6 on the validation set, and selected $K = 4$ as it provides a balanced trade-off between reconstruction fidelity (RMSE: 28.54° for $K = 4$ vs. 22.21° for $K = 6$) and model parsimony. Hyperparameter searches yield the optimal configuration: the number of modes $K = 4$ (shared by sine and cosine series), penalty coefficient $\alpha = 2000$, Lagrange multiplier update step $\tau = 0.0$, and convergence tolerance $\varepsilon = 1.0 \times 10^{-7}$. Similar hybrid architectures combining VMD with Transformer-based models have been applied to short-term wind power prediction [21]. Two-layer decomposition strategies combined with attention-enhanced temporal networks have also shown promising results in extracting multi-scale features from non-stationary wind sequences [23].

VMD is applied to the sine–cosine-encoded components of each data subset with K = 4. The decomposition preserves the original sample count of each subset (6410 for

training, 2095 for validation, and 1916 for test; see Table 1), as no boundary cropping is performed after decomposition.

To intuitively demonstrate the complete preprocessing workflow, Figure 2 illustrates the full data pipeline from raw wind direction series to the final 38-dimensional input vector. As shown in the flowchart, the raw dataset first undergoes RF-based feature screening to extract 30 auxiliary features from all candidates. Meanwhile, the target sine and cosine components are decomposed via VMD with $K = 4$, producing eight IMF subsequences in total (four for each encoded component). These eight IMF subsequences are concatenated with the 30 selected auxiliary features to construct the 38-dimensional input vector for the proposed VMD-Transformer with Gated Error Compensation.

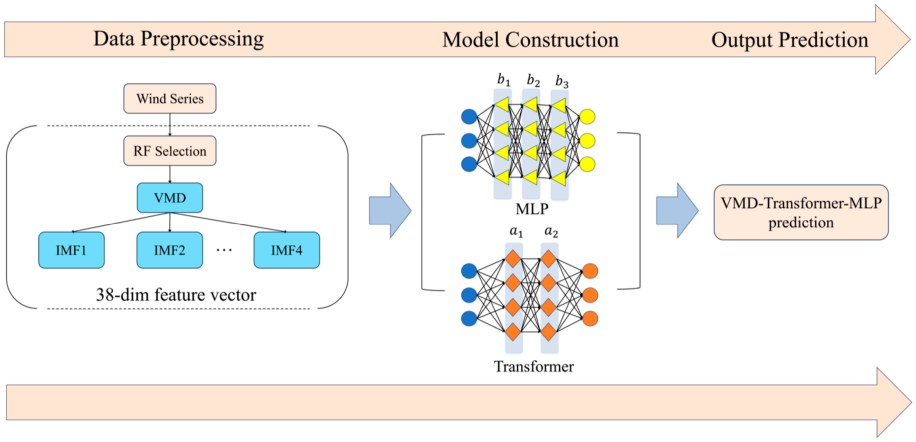


Figure 2. Schematic diagram of data preprocessing and feature extraction workflow incorporating RF feature screening and VMD multi-scale decomposition. Yellow and orange nodes denote outputs of the MLP branch and Transformer branch, respectively.

VMD decomposes each encoded wind direction component into four IMFs using the configuration detailed above. The decomposed IMF subsequences are concatenated with RF-selected auxiliary features to form the 38-dimensional input vector. The reconstruction fidelity of VMD is reported in Section 3.2.1 for reference, while the predictive validity of the IMF features is verified through ablation experiments in Section 3.2.5.

2.4.3. Normal Branch and Compensation Branch Design

The normal branch adopts a standard Transformer encoder. It receives 38-dimensional feature vectors, and the input time window size is set to 24, corresponding to the preceding 6 h observation data. The hyperparameters of the normal branch are configured as follows: encoder dimension of 128, 8 attention heads, 4 encoder layers, feed-forward dimension of 256, and a dropout rate of 0.3. The normal branch is trained on the full training dataset using the Adam optimizer, with a learning rate of 1.0×10^{-4}, a batch size of 32, and an early stopping patience of 30 epochs. Upon convergence, the normal branch outputs baseline predictions.

The core idea of the compensation branch lies in learning the residual errors generated by the normal branch rather than directly forecasting wind direction. This converts the complex wind direction prediction task into a residual error fitting task and alleviates the learning burden of the network [24].

(1) Error Jump Sample Generation

The trained normal branch is employed to make predictions on the training set, and the angular prediction error is calculated for each sample. Samples exhibiting obvious error jumps are selected as the training dataset for the compensation branch. The screening criteria are defined as follows: the prediction error at the previous time step is less than 5°,

the error at the current time step exceeds 10°, and the error jump amplitude is larger than 8°. A total of 324 error jump samples is screened, with an average error jump of 14.69°.

(2) Compensation Branch Network Structure

The compensation branch adopts a lightweight MLP with the following structure [12]. The input layer receives 38-dimensional feature vectors consistent with the input of the normal branch. The first hidden layer linearly maps the 38-dimensional input to 64 dimensions, followed by the ReLU activation function and a dropout rate of 0.3. The second hidden layer transforms the 64-dimensional features into 32 dimensions with ReLU activation and a dropout rate of 0.2. The output layer, equipped with the Tanh activation function, linearly maps the 32-dimensional features to 2 dimensions and outputs the residual error corrections for sine and cosine components.

Training hyperparameters are set as follows: learning rate of 1.0×10^{-3}, early stopping patience of 15 epochs, and MSE adopted as the loss function.

(3) Training Objective

The compensation branch does not directly learn the wind direction values; instead, it learns the prediction errors of the normal branch, i.e., the difference between the true values and the normal branch's predictions. In this way, the compensation branch only needs to focus on "where the normal branch makes mistakes" rather than "what the wind direction should be", thereby reducing the learning difficulty. After training, the compensation branch can predict the potential error of the normal branch for a given sample based on the input features [24].

2.4.4. Hysteresis-Driven Gating Fusion Mechanism

During the inference stage, this study develops a hysteresis-driven gated fusion mechanism. A rule-based detector identifies wind direction abrupt change moments and dynamically fuses outputs from two branches, enabling smooth adaptive switching between steady-state forecasting mode and mutation compensation mode [22].

The residual error e_t at time step t is defined as the angular deviation between baseline predictions from the normal Transformer branch and ground-truth wind direction values. To avoid frequent state jitter near threshold boundaries, a hysteresis state machine is designed for abrupt change judgment, with decision rules described as follows:

Activation condition: The gating signal $m_t = 1$ is triggered when the error jump from the preceding step $|e_{t-1} - e_{t-2}|$ exceeds 8° and the prediction error at the previous time step e_{t-1} exceeds 12°.

Deactivation condition: The compensation state is terminated once e_{t-1} falls below 8°.

State retention rule: After activation, the state $m_t = 1$ remains valid even if e_{t-1} falls within 8° to 12°, and compensation does not stop until the deactivation condition is met. Because all three conditions rely solely on e_{t-1} and e_{t-2}, both of which are available at inference time, the gating decision at step t requires no access to the current ground-truth value f_t. Hysteresis-driven switching strategies have been validated in various time-series forecasting applications to prevent frequent mode transitions [25].

Once the gating signal m_t is enabled, the model enters the wind mutation compensation mode [24]. Compensation strength is dynamically adjusted according to mutation severity, quantified by the maximum wind direction variation $\Delta\theta_{\max}$ within the preceding four time steps, corresponding to a one-hour sliding window. The variable $\Delta\theta_{\max}$ is discretized into six intervals to assign independent compensation multipliers.

Two separate threshold standards are applied for wind abrupt change identification. The 5°/10° threshold combination screens error jump samples for training the compensation branch, while the 8°/12° error thresholds serve as the activation criterion of the hysteresis gate during inference.

Adaptive weighted compensation strategies have been validated in multi-branch time-series forecasting architectures [12]. To realize adaptive weighted compensation, $\Delta\theta_{\max}$ is discretized into six continuous intervals: [0°,20°), [20°,60°), [60°,90°), [90°,120°), [120°,150°), [150°,180°]. Each interval k is matched with a compensation multiplier $\alpha_k \in [0,0.95]$. All multipliers are optimized via 50 iterations of random search on the validation set to strengthen error correction performance under drastic wind shifts.

Figure 3 plots the convergence curve of the random search optimization process. The optimal multiplier group is obtained at the 11th iteration with a validation MAE of 8.49°, and the overall MAE converges steadily. This result verifies the effectiveness of the validation-set parameter searching strategy and the stability of the optimized multiplier set. The optimized multipliers for each interval are listed below: 0.0 for [0°,20°), 1.2 for [20°,60°), 2.0 for [60°,90°), 1.0 for [90°,120°), 0.8 for [120°,150°), and 0.0 for [150°,180°].

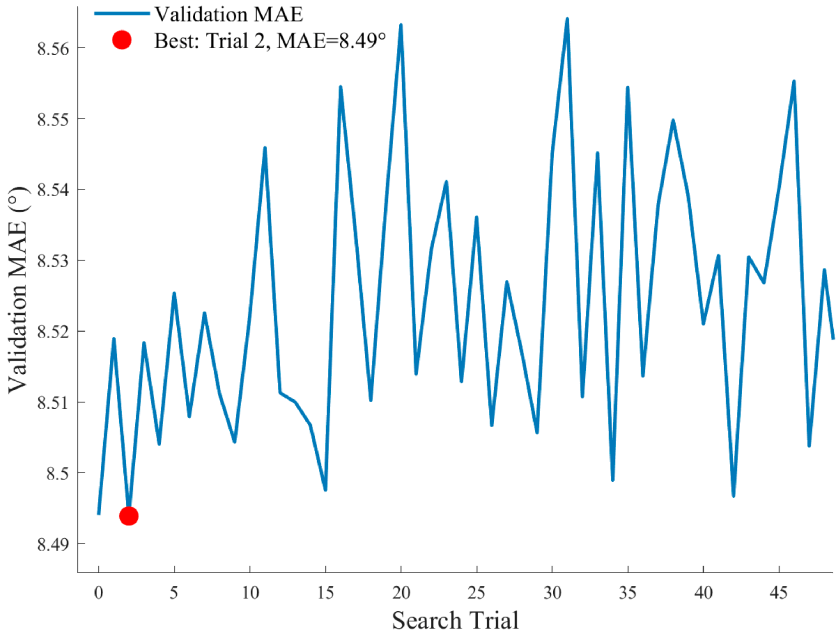


Figure 3. Convergence curve of random search for interval compensation multipliers on the validation set.

The maximum compensation weight of 2.0 is configured for the 60°~90° interval with moderate wind abrupt changes, whereas the multiplier is set to 0 for extreme wind reversal ranging from 150° to 180°. This configuration originates from the data distribution characteristics of actual wind farms: extreme reversal samples are extremely scarce in the training set, and blind compensation introduces extra prediction bias instead of improving accuracy.

Combining the hysteresis gating signal m_t and interval-adaptive multiplier α_k, the final fused prediction output of the dual-branch model is defined in Equation (12).

$$P_{\text{final}} = \begin{cases} P_{norm} + \alpha_k E_{\text{mlp}}, m_t = 1 \\ P_{norm}, m_t = 0 \end{cases} \quad (12)$$

where P_{norm} denotes the sine–cosine baseline prediction vector output by the normal branch, and E_{mlp} represents the 2-dimensional residual error vector predicted by the compensation branch.

For model training, the compensation branch does not directly regress wind direction angles. Instead, it learns the residual gap between baseline forecasts and ground-truth values to capture systematic prediction bias of the normal branch during abrupt wind

changes. This residual learning scheme allows the lightweight MLP to focus on local error features without overlapping the fitting capability of the normal branch, which achieves targeted error correction with very low parameter overhead [24].

In the inference phase, ground-truth labels are unavailable. The compensation branch estimates residual errors merely from input features, and the estimated residuals are fused with baseline outputs to generate compensated wind direction forecasts. To restrict sine and cosine predictions within the theoretical range of trigonometric functions, hard clipping is performed on fused outputs. After clipping, bounded sine and cosine values are converted back to the final wind direction angle via the four-quadrant arctangent function.

Differentiated training schemes are designed for the two branches. The normal branch is pre-trained on the full training set first. Once stable baseline accuracy is achieved, all weights of the Transformer encoder are frozen and no longer updated. By contrast, the compensation branch is only trained on screened error jump samples to fit residual mapping relationships. The overall loss function integrates angular prediction error, sine–cosine encoding loss and a range constraint penalty term to jointly optimize the entire model.

2.4.5. Model Training Strategy and Forecasting Procedure

The model takes a 38-dimensional preprocessed feature vector as input. Sliding windows with a length of 24 (corresponding to six hours of historical wind data) are adopted to construct sequential training samples. This work adopts a two-stage training scheme for the dual-branch architecture.

In the first stage, the Transformer-based normal branch is pre-trained on the full training dataset with the Adam optimizer [26]. The hyperparameter setup includes a learning rate of 1.0×10^{-4}, a batch size of 32, and a maximum training epoch of 200. An early stopping strategy with patience set to 15 is adopted to mitigate overfitting risks. Once the network converges and reaches stable baseline forecasting accuracy, all parameters of the normal branch are frozen and no longer updated.

In the second stage, screened error jump samples form a dedicated training subset exclusively for the MLP compensation branch. The compensation branch adopts independent hyperparameter configurations rather than copying the settings of the normal branch.

The overall loss function integrates angular prediction error and sine–cosine encoding error to jointly optimize the two branches. During the validation phase, threshold parameters of the hysteresis gate and six interval-adaptive compensation multipliers α_k are tuned through 50 random search trials on the validation set [22], aiming to boost error correction performance under wind direction abrupt changes. After hyperparameter optimization, the model is tested on an independent test set. MAE and RMSE are computed to quantitatively evaluate forecasting accuracy under both steady and drastic wind fluctuation scenarios [18].

3. Results

3.1. Experimental Setup

3.1.1. Benchmark Models and Comparative Configurations

To verify the prediction advantages of the proposed framework, seven benchmark models covering Transformer architectures, recurrent neural networks (RNNs) and reservoir computing methods are established for a comparative ablation study [27]. Prior research has confirmed that Transformer-based and tree-based models deliver complementary forecasting capabilities under different time horizons for wind direction prediction [28]. All comparison models adopt a unified 38-dimensional input feature space, and the training and testing processes are carried out with consistent dataset splitting schemes and identical hardware environments.

Detailed settings of each comparative model are elaborated as follows:

VMD-Transformer: Transformer encoder integrated with VMD multi-scale decomposition. Wind direction time series are decomposed into four IMFs, which are concatenated with 30 auxiliary features to form the 38-dimensional input vector. This configuration isolates the performance improvement contributed solely by VMD preprocessing.

VMD-GRU: Lightweight mainstream recurrent architecture embedded with VMD preprocessing.

VMD-LSTM: Classical LSTM network integrated with VMD [26].

Pure Transformer: Standard Transformer encoder without VMD preprocessing or the residual compensation branch. Raw 38-dimensional features are fed directly into the network. This model acts as the fundamental ablation baseline to quantify the independent gains brought by VMD and the hysteresis compensation module.

Pure ESN: Classical reservoir computing model without VMD, representing typical reservoir learning frameworks.

VMD-ESN: Echo State Network equipped with VMD on input sequences, adopted to verify the generalizability of VMD across different model paradigms [27].

Proposed Model: The complete dual-branch framework composed of the frozen Transformer normal branch, lightweight MLP residual compensation branch, and hysteresis gate fusion module.

Early hybrid models combining signal decomposition with evolutionary optimization have also demonstrated the potential of decomposition-based strategies for wind forecasting [29]. All benchmark models are optimized via the Adam optimizer with a unified learning rate of 1.0×10^{-4} and a batch size of 32. The maximum training epoch is limited to 200, and an early stopping strategy with patience = 15 epochs is adopted to mitigate overfitting risks. Hyperparameters of each comparative model are independently tuned on the validation set. Each experiment group is repeated five times with distinct random seeds, and averaged forecasting metrics are reported to guarantee statistical reliability.

3.1.2. Ablation Experiment Design

A series of ablation tests are carried out to distinguish how each core component contributes to overall forecasting accuracy. Three simplified model variants are constructed: each removes one dedicated functional module while all hyperparameter settings and preprocessing pipelines remain consistent with the full dual-branch model.

(1) Without VMD: The VMD multi-scale decomposition module is removed. The model only takes the 30 auxiliary features screened by RF as inputs, without the eight IMF subsequences decomposed from the wind direction series. This variant shares an identical Transformer encoder structure and training hyperparameters with the proposed model, corresponding to the Pure Transformer benchmark. Contrast experiments against the VMD-Transformer baseline reveal how much VMD frequency decomposition lifts prediction accuracy.

(2) Without Compensation: This test excludes the entire lightweight MLP residual compensation branch, leaving the VMD-Transformer as the sole forecasting backbone. Its output follows $P_{\text{final}} = P_{\text{norm}}$ without any residual error compensation. Test metric gaps between this simplified version and the complete dual-branch architecture reflect the accuracy promotion brought by the compensation branch.

After comparing test set metrics across all simplified variants and the full model, we can separately analyze the individual effects of VMD and the MLP compensation branch (coupled with the hysteresis gate).

3.1.3. Evaluation Metrics

Three error evaluation metrics are adopted to comprehensively assess model prediction performance: Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), defined as follows:

$$\text{MAPE} = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{13}$$

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{14}$$

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i| \tag{15}$$

where m denotes the total number of test samples; y_i represents the observed ground-truth wind direction value at the i-th time step; $\hat{y}_i$ stands for the corresponding predicted value output by the forecasting model.

3.2. Experimental Results and Analysis

3.2.1. VMD Validation

Prior to comparative forecasting experiments, the VMD preprocessing scheme was validated from the perspectives of waveform hierarchical separation and reconstruction fidelity, as illustrated in Figures 4 and 5. Figure 4 shows the temporal curves of four intrinsic mode function (IMF) components decomposed from the original wind direction time series under unified coordinate axes, which facilitates direct comparison of the amplitude and frequency characteristics among different subsequences. In detail, IMF 1 corresponds to the component with the highest oscillation frequency and relatively small overall amplitude, while IMF 4 exhibits the lowest frequency together with the largest fluctuation amplitude. Such hierarchical frequency division is consistent with the basic decomposition principle of VMD.

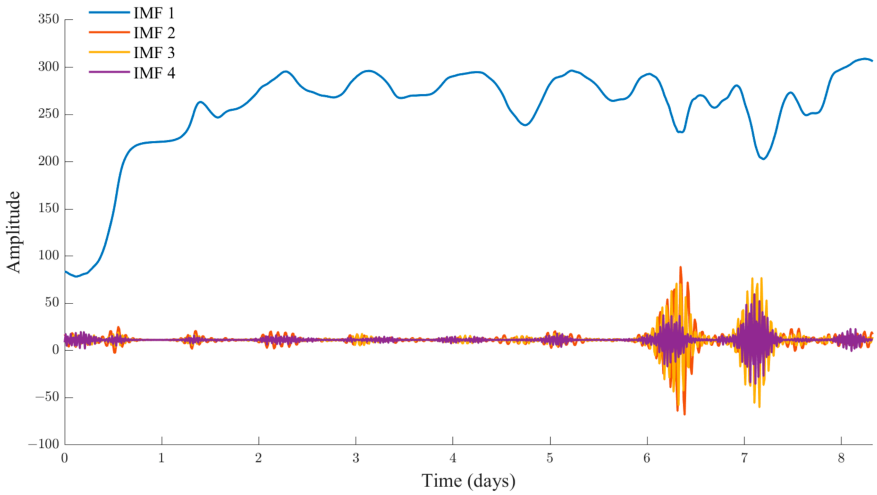


Figure 4. Temporal curves of four IMF components decomposed from wind direction time series.

Figure 5 presents the statistical results of VMD signal reconstruction error. The left subgraph plots the time-varying curve of reconstruction deviation between the superposed IMF components and the raw wind direction sequence, and the right subgraph illustrates the frequency distribution histogram of reconstruction error, with core statistical indicators marked. The reconstruction is performed by separately decomposing the sine and

cosine components to account for the circular nature of wind direction data, yielding a mean error of 0.31°, a standard deviation of 13.28°, and an RMSE of 19.94°. This level of reconstruction fidelity does not compromise downstream forecasting performance: the ablation study (Section 3.2.5) shows that removing VMD increases MAE by 67.0%, demonstrating that the IMF features carry substantial predictive information despite the moderate reconstruction deviation.

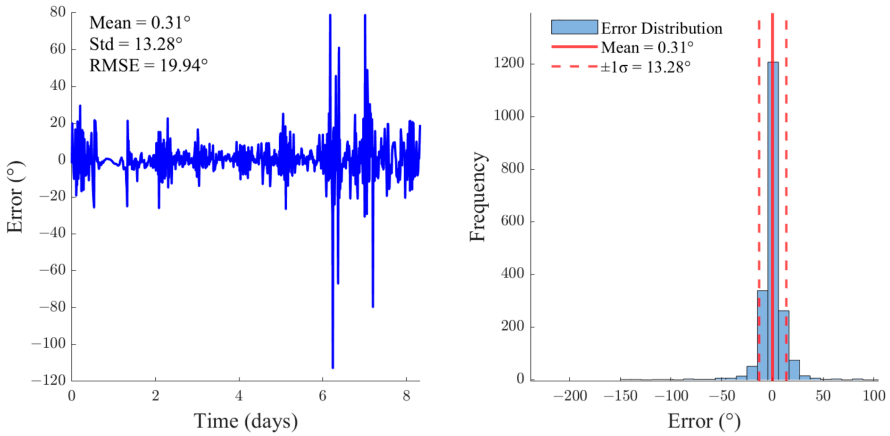
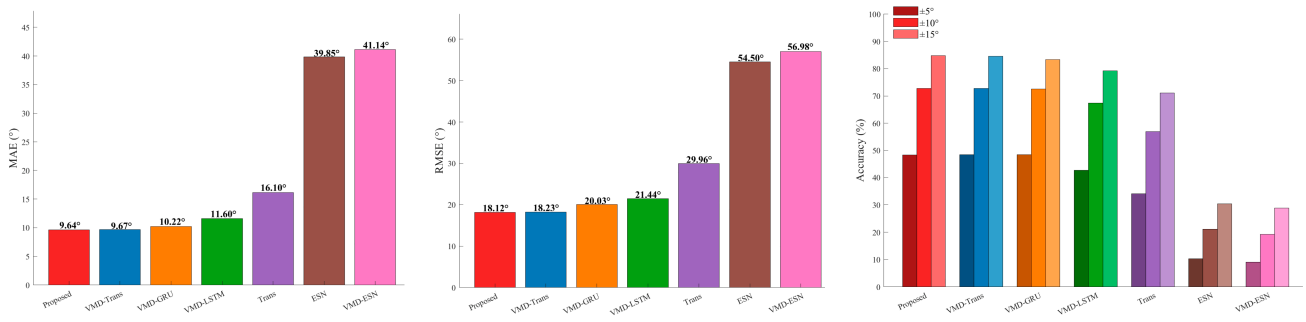


Figure 5. Statistical analysis of VMD reconstruction errors, including time-series deviation and error distribution histogram.

3.2.2. Overall Prediction Performance Comparison

Quantitative comparisons of forecasting errors for all benchmark models are given in Figure 6, and Figure 7 shows the fitting effect between model-predicted sequences and ground-truth values for each scheme. Figure 6 consists of three bar charts for multi-dimensional error assessment. Subplot (a) shows the MAE values of all comparative schemes, and the proposed dual-branch framework yielded the minimum MAE of 9.64°. Subplot (b) lists the statistical RMSE results; the proposed method also achieved the lowest RMSE of 18.12°. Since RMSE assigns larger penalty weights to large prediction deviations, this indicator verified the reliable performance of the designed model under abrupt variations. Subplot (c) compares prediction accuracy under three error limits (±5°, ±10° and ±15°), which further validated the stable forecasting capacity of the dual-branch architecture.



(**a**) (**b**) (**c**)

Figure 6. Quantitative performance comparisons of all benchmark models. (**a**) MAE distribution of each forecasting model; (**b**) RMSE distribution of each forecasting model; (**c**) Prediction accuracy of different models under three error-tolerance thresholds. Each model group is plotted with a unique color system. Within each group, the darker bar corresponds to the accuracy within an error of ±5°, the medium-shade bar represents the accuracy within an error of ±10°, and the light-shade bar indicates the accuracy within an error of ±15°.

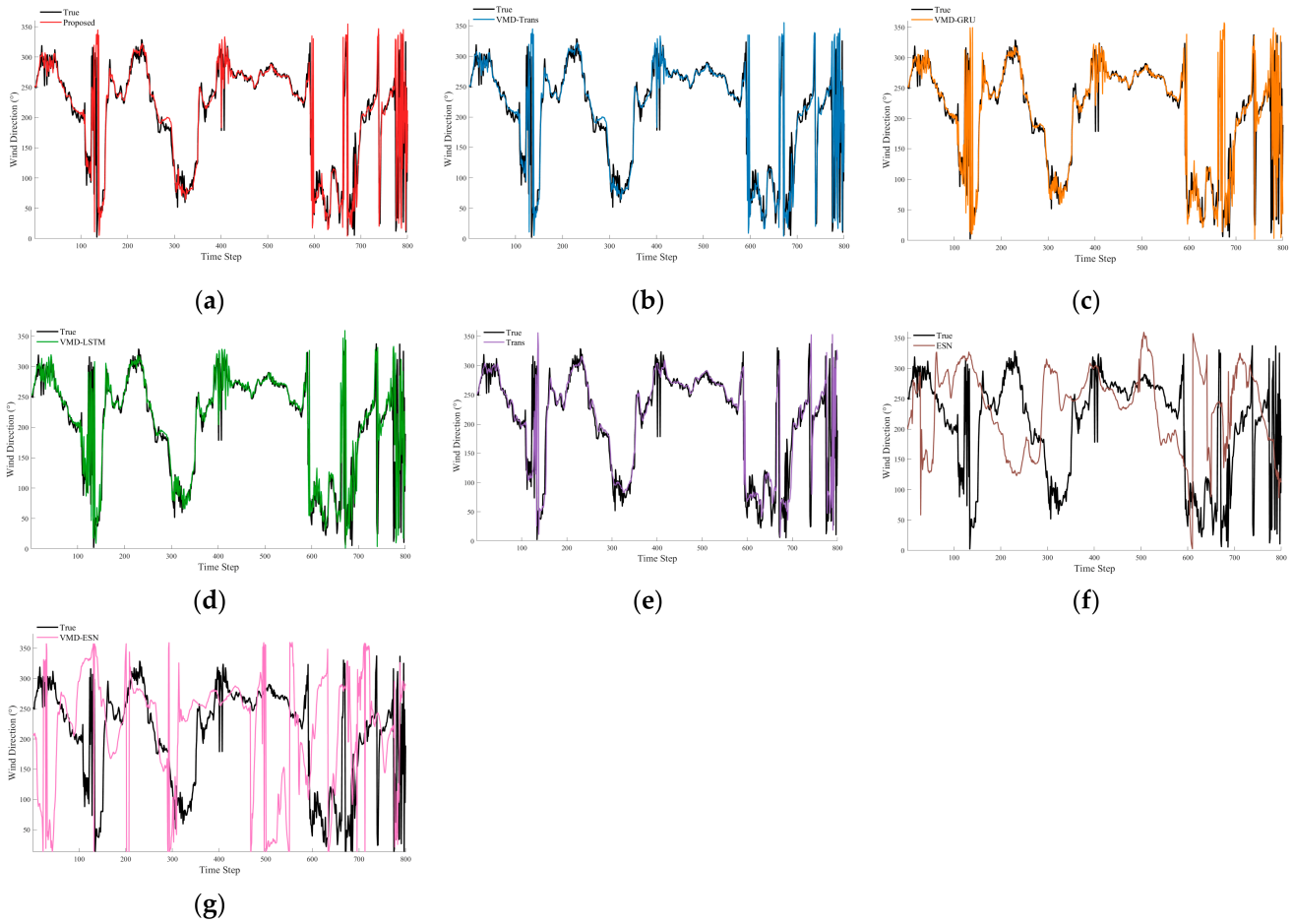


Figure 7. Temporal fitting curves of different models against ground-truth wind direction values. (**a**) Proposed model; (**b**) VMD-Transformer; (**c**) VMD-GRU; (**d**) VMD-LSTM; (**e**) Pure Transformer; (**f**) Pure ESN; (**g**) VMD-ESN.

Figure 7 displays time-series fitting curves of the seven models following the unified sequence defined in Section 3.1.1. Subplot (a) corresponds to the proposed model; its predicted curve maintained high coincidence with ground-truth values across the whole time span. Even during abrupt wind direction fluctuations, the predicted values deviated only slightly from the ground-truth values, highlighting the outstanding tracking performance for sudden wind shifts. By comparison, the other six benchmark models suffered from obvious response lag and considerable deviations at all mutation points. This curve difference accounted for their distinctly higher overall MAE and RMSE relative to the proposed framework. These results were consistent with recent findings that Transformer-based models achieve competitive performance across multiple forecasting horizons in wind energy applications [28].

3.2.3. Performance on Wind Direction Mutation Events

Figure 8 presents the distribution of error improvement across all 192 triggered samples, with green markers denoting normal triggers (angular change $\leq 70°$) and red markers denoting genuine mutation triggers (angular change $>70°$). The overall mean improvement across all triggered samples is 0.32°, indicating that the compensation branch produces limited gains for most routine triggers. However, for the improved genuine mutation samples with angular change exceeding 70°, the mean improvement reaches 6.02°, demonstrating that the compensation mechanism is specifically effective under extreme wind direction shifts.

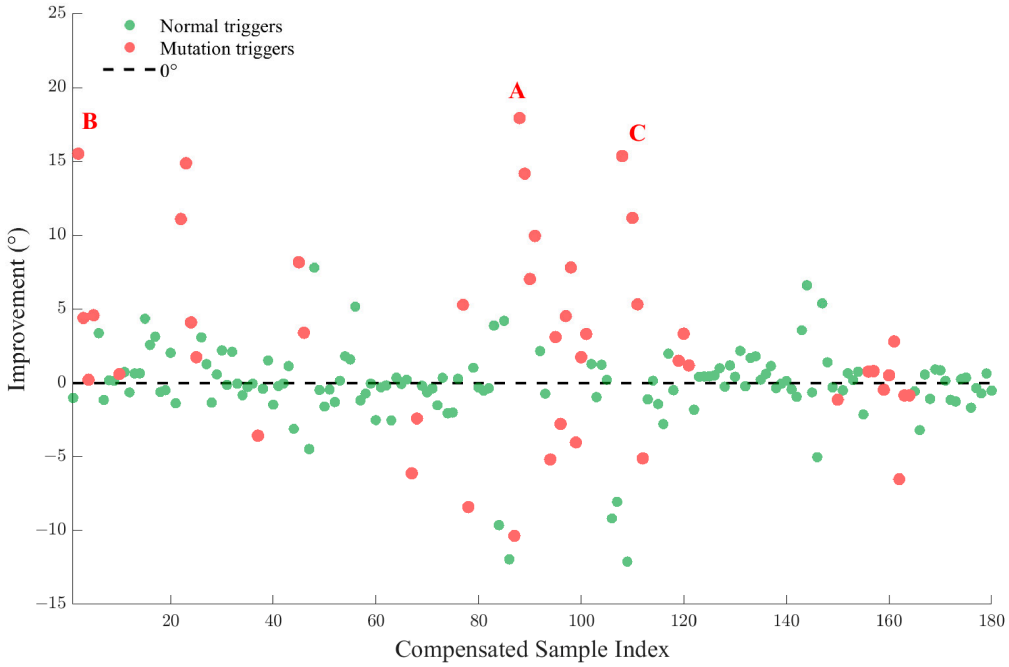


Figure 8. Distribution of error improvement across all 192 triggered samples. Green: normal triggers (angular change ≤70°), red: genuine mutation triggers (angular change >70°). A, B, and C denote the three most improved extreme cases.

Three extreme cases (A, B, C) achieve improvements of 17.94°, 15.53°, and 14.89°, respectively, highlighting the maximum error correction potential of the compensation mechanism under severe wind shifts. These results confirm that the designed compensation module delivers targeted error correction for genuine mutation events, while avoiding unnecessary compensation for routine fluctuations.

3.2.4. Prediction Uncertainty and Stability Assessment

The temporal evolution of prediction uncertainty and long-term forecasting stability is assessed, as presented in Figure 9. The black solid line represents ground-truth wind direction measurements, the red curve corresponds to predictions generated by the proposed dual-branch model, and the gray shaded region indicates the 95% confidence interval that quantifies the fluctuation scope of prediction uncertainty. Across the whole time span, the confidence band remained narrow and only widened slightly during wind direction mutations. The predicted sequence persistently followed the changing trend of actual wind direction; the uncertainty band only underwent slight broadening when drastic wind direction mutations occurred. Such characteristics demonstrated that the proposed model featured limited uncertainty oscillation and satisfactory long-term stability for continuous wind direction forecasting.

Figure 10 quantifies the overall statistical characteristics of forecasting errors from a static perspective. The blue histogram shows the probability density distribution of all prediction deviations. The solid red line denotes the zero-error baseline, and dashed lines mark the MAE value as well as the ±10° and ±15° error thresholds introduced in the previous section. Prediction errors clustered tightly near zero without evident systematic bias, indicating the absence of consistent offset in model outputs. Most deviations fell within the ±15° allowable error range, a feature that verified the steady and reliable forecasting capacity of the proposed dual-branch model.

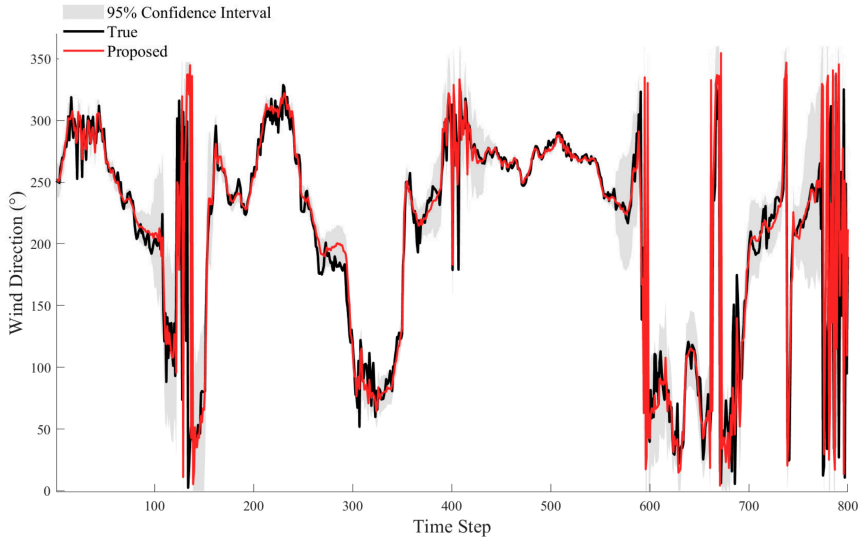


Figure 9. Time-series forecasting results with 95% confidence interval.

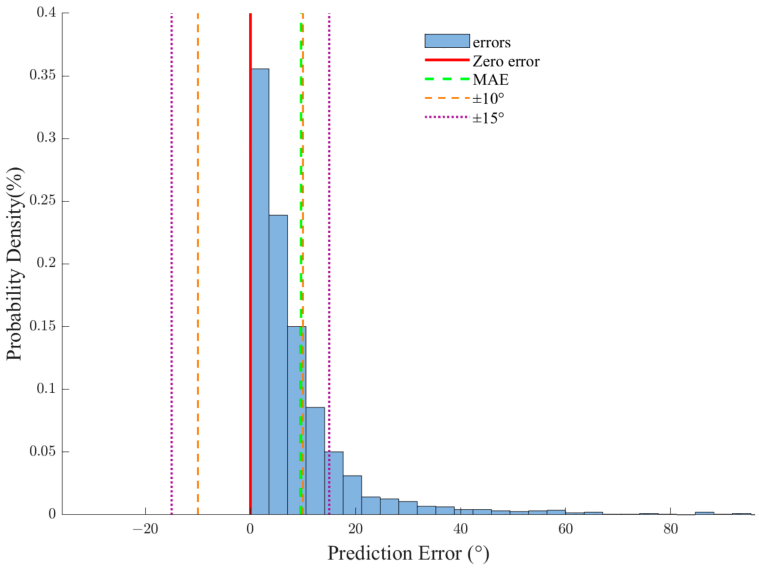


Figure 10. Probability density distribution histogram of prediction errors. The red solid line represents zero error; the green dashed line denotes the MAE; the orange dashed line indicates the ±10° error threshold; the pink dotted-dashed line represents the ±15° error threshold.

3.2.5. Ablation Study and Module Contribution Analysis

Table 2 reports the test-set performance of the proposed model alongside six benchmark models. To quantify the isolated contribution of each core module, ablation analysis focuses on three configurations: the full model, the VMD-Transformer baseline, and the Pure Transformer.

From the table, eliminating VMD led to the most obvious performance drop. In detail, after removing VMD preprocessing, MAE climbed from 9.64° to 16.10°, corresponding to a 67.0% relative rise; RMSE grew from 18.12° to 29.96°, with a 65.3% increase, while prediction accuracy within ±10° fell by 15.8 percentage points. Such obvious performance regression verified that VMD served as the fundamental component of the dual-branch framework. The raw non-stationary wind direction time series is split into several band-limited IMF components via VMD, which greatly lowers the difficulty of subsequent modeling. Without this preprocessing step, the Transformer encoder cannot adequately extract multi-scale frequency features hidden in the mixed signal.

Table 2. Performance comparison of all benchmark models.

Model Configuration	MAE (°)	RMSE (°)	±10° Accuracy (%)
Proposed	9.64	18.12	72.73
VMD-Transformer	9.67	18.23	72.73
VMD-GRU	10.22	20.03	72.57
VMD-LSTM	11.60	21.44	67.39
Pure Transformer	16.10	29.96	56.92
Pure ESN	39.85	54.50	21.15
VMD-ESN	41.14	56.98	19.25

In addition, removing the compensation branch with hysteresis gating caused mild performance degradation. MAE rose slightly from 9.64° to 9.67°, and RMSE increased marginally from 18.12° to 18.23°, while the ±10° forecasting accuracy remained unchanged at 72.73%. The limited improvement can be explained by the fact that, as analyzed in Section 3.2.3, the compensation branch achieved prominent error correction at abrupt wind mutation locations, with error reductions of 17.94°, 15.53°, and 14.89° at Point A, Point B, and Point C. Nevertheless, mutation samples only occupied 10.1% of the entire test dataset, so their optimization effect cannot be fully reflected in aggregate evaluation metrics.

4. Discussions

The experimental results presented in Section 3 consistently validated the working hypothesis that combining VMD frequency separation, independent residual compensation, and adaptive hysteresis gating can simultaneously improve overall wind direction prediction accuracy and strengthen tracking capability under abrupt wind shifts. The core findings are interpreted below in light of existing literature and industrial engineering demands as follows:

(1) Foundational role of VMD preprocessing. Eliminating VMD raised MAE by 67.0%, proving VMD is foundational to the framework. Raw wind direction series mix slow seasonal trends and fast turbulent fluctuations; these intertwined frequency components hinder the Transformer from extracting long-term temporal correlations. VMD splits the non-stationary sequence into multiple band-limited IMF components and greatly lowers modeling difficulty. Without this decomposition step, the model cannot achieve reliable wind direction prediction.

(2) Targeted optimization for mutation events. Combining the compensation branch with hysteresis gating and zoning reduced errors by a maximum of 17.94° at wind mutation points, while producing only a slight improvement in overall MAE. Mutation samples accounted for only 10.1% of test data, in line with the low frequency of sudden wind changes in actual wind farms, hence the modest overall improvement. The asymmetric hysteresis rule prevents threshold jitter, and zoning adjusts compensation strength based on mutation intensity. The dual regulation guaranteed stable prediction under regular wind conditions and rapid response to abrupt fluctuations, an effect impossible for traditional single-branch structures.

(3) Superiority over existing methods. Recent studies have emphasized the effectiveness of VMD-enhanced deep learning models for wind energy time-series forecasting [25]. The dual-branch Transformer surpassed six mainstream benchmark models on all evaluation indicators, achieving the lowest test MAE of 9.64° and RMSE of 18.12° (Figure 6). Through parallel self-attention calculation, the model captured global temporal correlations and bypassed the sequential computation limitation of recurrent networks. Unlike conventional VMD-Transformer models that only optimize overall fitting precision, this

work added an independent compensation module specifically for abrupt wind variations, addressing the lack of targeted error correction in prior research.

(4) Uncertainty and stability assessment. Narrow 95% confidence intervals and zero-centered error distribution indicated negligible systematic bias, enabling stable continuous multi-step prediction. This stability is essential for wind farm yaw adjustment and power scheduling; quantified uncertainty avoids overconfident decisions and enhances operational safety.

(5) VMD reconstruction sensitivity across data segments. The reconstruction fidelity of VMD varies across data segments: when the analysis covers a longer time span with more diverse weather conditions, the reconstruction RMSE tends to increase. This is because VMD is a global decomposition algorithm with fixed parameters that cannot adapt to time-varying signal characteristics across different months or weather regimes. Nevertheless, as the Transformer operates directly on the decomposed IMF components rather than the reconstructed signal, this reconstruction variability does not materially compromise downstream forecasting performance—a finding consistent with the ablation results in Section 3.2.5.

This work has four main limitations. First, all data were collected from one wind farm, so the model could not be directly applied to coastal, offshore, or mountain wind fields with varied turbulence features. Second, hysteresis thresholds were determined empirically and lack dynamic adaptive adjustment under extreme weather. Third, only single-step prediction was realized, and cumulative errors degraded performance during multi-step forecasting. Fourth, VMD employs fixed global parameters and cannot adapt to time-varying wind characteristics across seasons. These limitations echo the challenges identified in recent studies on wind direction forecasting [30].

Future work will adopt multi-region datasets to improve cross-scenario generalization, use reinforcement learning to optimize gating thresholds dynamically, and compress the model for embedded deployment on wind turbine terminals. Besides wind direction prediction, the proposed decomposition–compensation workflow can be extended to other non-stationary time-series forecasting tasks, such as short-term power load [16].

Supplementary Materials: The following supporting information can be downloaded at: https://zenodo.org/record/21543995, including the processed feature files, source code, and additional experimental comparison results. The raw dataset can be accessed via reference [12]. The model training code is not publicly shared but can be provided by the corresponding author for verification purposes upon reasonable request.

Author Contributions: Conceptualization, Y.L. and Z.L.; methodology, Y.L. and Z.L.; software, Y.L.; validation, Y.L.; formal analysis, Y.L.; investigation, Y.L. and Z.L.; resources, X.J.; data curation, Y.L.; writing—original draft preparation, Z.L.; writing—review and editing, Z.L. and X.J.; visualization, Y.L. and Z.L.; supervision, X.J.; project administration, X.J.; funding acquisition, X.J., T.Y. and D.X. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by the National Natural Science Foundation of China (Grant No. 42405142), the Chengdu Science and Technology Program (Grant No. 2025-YF05-00257-SN), and the Open Research Fund of Key Laboratory of Digital Innovation of Tianfu Culture, Sichuan Provincial Department of Culture and Tourism, Chengdu University (Grant No. TFWH-2025-03).

Institutional Review Board Statement: Not applicable. The study did not involve human participants or animal experiments.

Informed Consent Statement: Not applicable. No human participants were involved in this study.

Data Availability Statement: The wind field data used in this paper are from the open dataset of the State Grid Renewable Energy Forecasting Competition, available at https://doi.org/10.1038/s41597-022-01696-6 [12]. The processed feature files are included in the Supplementary Materials. The

custom analysis code is not publicly available due to ongoing research considerations but may be obtained from the corresponding author upon reasonable request.

Acknowledgments: The authors thank State Grid for providing open wind farm measurement data. During manuscript revision, ChatGPT 4.0 was used for English grammar polishing; all text content has been fully reviewed and revised by the authors, who take full responsibility for the paper.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

RF	Random Forest
VMD	Variational Mode Decomposition
IMF	Intrinsic Mode Function
MLP	Multilayer Perceptron
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
ESN	Echo State Network
MHSA	Multi-Head Self-Attention
FFN	Feed-Forward Network
ADMM	Alternating Direction Method of Multipliers
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
Trans	Transformer
VMD-ESN	VMD combined with ESN
VMD-GRU	VMD combined with GRU
VMD-LSTM	VMD combined with LSTM
VMD-Transformer	VMD combined with Transformer
Proposed	VMD-Transformer with Gated Residual Compensation

References

1. Hu, C.C. Brief Discussion on the Wind Power Generation Development Status and Industry Development. *Sci-Tech Innov. Product.* **2023**, *3*, 55–57. [CrossRef]
2. Alazhare, A.; Howland, M.F.; Fitzgerald, B. Short-Horizon Data-Driven Joint Forecasting of Wind Speed and Direction for Future Aware Wind Farm Control. *J. Phys. Conf. Ser.* **2026**, *3224*, 022024. [CrossRef]
3. Al Qahtani, N.; Abolouz, M.; Mohandes, M.; Rehman, S. Comparative evaluation of recurrent and attention-based deep learning models for wind speed forecasting. *FME Trans.* **2025**, *53*, 693–704. [CrossRef]
4. Kishore, B.; Kabilan, K.; Rahul, L.S.; Priyadarshan, G.P.; Vishnutheerth, E.P.; Satheesh, R.; Kolhe, M.L. Advancing short-term wind power forecasting by AI-driven models for improved accuracy. *Electr. Eng.* **2025**, *107*, 14795–14810. [CrossRef]
5. Alves, D.; Mendonça, F.; Mostafa, S.S.; Morgado-Dias, F. The potential of machine learning for wind speed and direction short-term forecasting: A systematic review. *Computers* **2023**, *12*, 206. [CrossRef]
6. Dragomiretskiy, K.; Zosso, D. Variational Mode Decomposition. *IEEE Trans. Signal Process.* **2014**, *62*, 531–544. [CrossRef]
7. Lin, T.; Wang, J.J.; Zhang, D. Short-Term Wind Direction Prediction Research Based on VMD-BA-LSTM. *Chin. High Technol. Lett.* **2021**, *31*, 653–659. [CrossRef]
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010. [CrossRef]
9. Zhang, K.; Li, X.; Su, J. Variable Support Segment-Based Short-Term Wind Speed Forecasting. *Energies* **2022**, *15*, 4067. [CrossRef]
10. Nascimento, E.G.S.; de Melo, T.A.C.; Moreira, D.M. A transformer-based deep neural network with wavelet transform for forecasting wind speed and wind energy. *Energy* **2023**, *278*, 127678. [CrossRef]
11. Wu, J.; Chang, Z.; Zhang, L.; Chen, M.; Li, S.; Qiu, F. A Dual-Branch Transformer Network with Multi-Scale Attention Mechanism for Microgrid Wind Turbine Power Forecasting. *Electronics* **2025**, *14*, 2566. [CrossRef]

12. Gao, R.; Hu, M.; Li, R.; Luo, X.; Suganthan, P.N.; Tanveer, M. DualNet: A dual-path network with adaptive compensation for multivariate time series forecasting. *Neurocomputing* **2025**, *649*, 130720. [CrossRef]
13. Xu, R.H.; Fang, H.Y.; Zeng, H.Z.; Wu, B.R. A novel interpretable wind speed forecasting based on the multivariate variational mode decomposition and temporal fusion transformer. *Energy* **2025**, *331*, 136497. [CrossRef]
14. Fan, Z.; Wang, Z.; Wang, N.; Wu, R. Research on Virtual LiDAR Wind Measurement Technology for Wind Farms Based on CFD and Deep Learning. In Proceedings of the 2024 International Conference on Electrical Power Systems and Intelligent Control (EPSIC), Changsha, China, 24–27 May 2024; pp. 153–157. [CrossRef]
15. Chen, Y.B.; Xu, J. Solar and Wind Power Dataset from State Grid Renewable Energy Forecasting Competition. *Sci. Data* **2022**, *9*, 577. [CrossRef] [PubMed]
16. Li, K.; Yuan, L.; Qian, F.; Song, L.; Wu, X.; Wang, L.; Dai, J.; Shen, L. Short-Term Load Forecasting for Electricity Spot Markets Across Different Seasons Based on a Hybrid VMD-LSTM-Random Forest Model. *Energies* **2025**, *18*, 6097. [CrossRef]
17. Chen, W.G.; Zhang, H.L. Application of RF-Light GBM Algorithm in Early Warning of Fan Blade Cracking. *Electron. Meas. Technol.* **2020**, *43*, 167–173. [CrossRef]
18. Tang, Z.H.; Zhao, G.N.; Cao, S.X. A Data-Analysis-Based Hybrid Wind Direction Forecasting Algorithm. *Acta Energ. Sol. Sin.* **2021**, *42*, 349–356. [CrossRef]
19. Xiong, B.; Lou, L.; Meng, X.; Wang, X.; Ma, H.; Wang, Z. Short-term wind power forecasting based on Attention Mechanism and Deep Learning. *Electr. Power Syst. Res.* **2022**, *206*, 107776. [CrossRef]
20. Sun, S.; Liu, Y.; Li, Q.; Wang, T.; Chu, F. Short-term multi-step wind power forecasting based on spatio-temporal correlations and transformer neural networks. *Energy Convers. Manag.* **2023**, *283*, 116916. [CrossRef]
21. Xu, W.; Dai, W.; Li, D.; Wu, Q. Short-term wind power prediction based on a variational mode decomposition–BiTCN–Psformer hybrid model. *Energies* **2024**, *17*, 4089. [CrossRef]
22. Liu, H.; Han, H.; Sun, Y.; Shi, G.; Su, M.; Liu, Z.; Wang, H.; Deng, X. Short-term wind power interval prediction method using VMD-RFG and Att-GRU. *Energy* **2022**, *251*, 123807. [CrossRef]
23. Zhang, D.; Chen, B.; Zhu, H.; Goh, H.H.; Dong, Y.; Wu, T. Short-term wind power prediction based on two-layer decomposition and BiTCN-BiLSTM-attention model. *Energy* **2023**, *285*, 128762. [CrossRef]
24. Guo, T.; Liu, Q.; Fang, X.; Fang, M.; Wu, Y.; Tao, R.; Xiao, R. Offshore wind power prediction based on improved Hankel-DMD-LSTM hybrid model: Strengthening mode selection and residual compensation mechanism. *Ocean Eng.* **2026**, *352*, 124607. [CrossRef]
25. Xu, H.; Zhu, Z.; Liu, H.; Zio, E. Spectral dynamic aggregation transformer and fitted swing-door algorithm for wind power monitoring. *Energy* **2025**, *341*, 139396. [CrossRef]
26. Graves, A. Long Short-Term Memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 37–45. [CrossRef]
27. Liu, W.; Bai, Y.; Yue, X.; Wang, R.; Song, Q. Wind Speed Forecasting Model Based on Optimized VMD and Multi-Head Self-Attention LSTM. *Energy* **2024**, *294*, 130726. [CrossRef]
28. Blazakis, K.; Perrakis, S.; Karapidakis, E.; Mavromatakis, F.; Katsigiannis, Y. ML-Based Multi-Horizon Wind Speed and Wind Direction Forecasting for Aviation and Energy Applications in Coastal Crete. *Electronics* **2025**, *14*, 4856. [CrossRef]
29. Wang, S.; Zhang, N.; Wu, L.; Wang, Y. Wind Speed Forecasting Based on the Hybrid Ensemble Empirical Mode Decomposition and GA-BP Neural Network Method. *Renew. Energy* **2016**, *94*, 629–636. [CrossRef]
30. Liu, H.; Yang, R. Artificial intelligence for wind speed forecasting: A review on multi-scale decomposition and intelligent fusion strategies. *Adv. Wind Eng.* **2025**, *2*, 100055. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.