EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

A hybrid LPMS-Transformer-ARFIMA framework for wind power forecasting with SPSO-based global hyperparameter optimization

Zhen Huang[1], Kaiyuan Hou[1], Deming Xia[1], Zhiyu Liao[2,*], Xihai Guo[1], Linsong Ge[1], Yu Sun[1]

[1] Northeast Branch of State Grid Corporation of China, Shenyang 110180, China
[2] School of Renewable Energy, Shenyang Institute of Engineering, Shenyang 110136, China

Abstract

The severe stochastic nature, multi-scale variations, and non-stationary characteristics typical of wind power temporal data creates substantial difficulties for reliable forecasting. To tackle these specific obstacles, this paper develops an advanced hybrid framework for wind power forecasting, VMD-PSO-LPMS-Transformer-ARFIMA, with global hyperparameter optimization based on SPSO. Firstly, the core hyperparameters of the variational mode decomposition (VMD) algorithm are adaptively optimized using particle swarm optimization (PSO), facilitating stable signal decomposition and robust noise suppression. Subsequently, a linear predictability median split (LPMS) method is proposed. This strategy quantifies the predictability of each decomposed component through the forecasting error of a linear proxy model on the validation set and uses this measure to partition the components into high-frequency and low-frequency groups. In the forecasting stage, the low-frequency components characterized by long memory trends are modeled using autoregressive fractionally integrated moving average (ARFIMA), whereas the high-frequency nonlinear fluctuations are captured using a Transformer architecture to model global dependencies. To alleviate the substantial computational expenses frequently encountered during hyperparameter optimization for deep learning frameworks, this study incorporates a surrogate assisted particle swarm optimization (SPSO) strategy. This approach relies fundamentally on a Gaussian Process surrogate model to efficiently streamline the tuning procedure. Experimental validation is performed utilizing wind power datasets collected from a wind farm located in Xinjiang, China, with April and October selected as representative seasons for comparative and ablation analysis. Based on the empirical evaluations, superior stability across varying seasons alongside the highest prediction precision is successfully realized through the developed hybrid architecture.

Keywords: Deep learning, Frequency characteristics, Wind power forecasting

Received on 05 February 2026, accepted on 10 April 2026, published on 16 September 2026

doi: 10.4108/ew.14216

1. Introduction

As power systems accelerate the widespread grid integration of renewable power, wind power is increasingly incorporated as a flexibility resource to support critical operational requirements, including frequency regulation, ramp tracking, and real-time balancing. According to statistics reported by the World Meteorological Organization, an escalation of roughly 30% in worldwide power requirements has been observed throughout the preceding decade. To limit global warming to within 1.5 °C, the annual improvement in global energy efficiency must increase from the current 2 % to more than 4 % by 2030. Furthermore, by 2050, global electricity demand is expected to be supplied entirely by renewable energy sources [1]. China has made substantial contributions toward these global objectives by continuously promoting the accelerated transformation of its energy structure and facilitating the large-scale utilization of renewable energy.

However, wind power generation exhibits pronounced stochasticity and uncertainty, which significantly increases

* Corresponding author. Email: liaozhiyu0123@163.com

the complexity of dispatch planning, reserve capacity allocation, and real-time balancing control. Consequently, to guarantee the operational security of utility grids and facilitate optimal scheduling strategies amidst the continuous penetration of fluctuating renewables, securing highly accurate wind power forecasts is now universally recognized as an essential requirement [2-4]. From the perspective of forecasting horizon, prediction tasks are typically categorized into four primary tiers: long-term, medium-term, short-term, and ultra-short-term forecasting. The ultra-short-term forecasting primarily supports rapid control and real-time regulation, while the others are mainly employed for unit commitment, maintenance scheduling, and strategic operational planning [5]. Among these forecasting horizons, ultra-short-term prediction is the most demanding due to its short prediction window and extremely low tolerance for error. Nevertheless, its outputs directly influence closed loop control and regulation processes. Once forecast deviations accumulate, they may lead to operational risks such as insufficient reserve capacity or regulation limit violations.

Wind power generation series frequently alternate between relatively smooth variations and abrupt ramp-like changes. These dynamics are typically accompanied by intermittency, non-stationarity, multi-scale oscillations, and strong stochastic disturbances [6]. These characteristics often coexist in a superimposed manner. Consequently, if a single predictive model is trained directly on the raw sequence, it becomes highly sensitive to disturbances, and its generalization capability across different seasons or operating conditions may deteriorate significantly [7]. To overcome the aforementioned obstacles, researchers have established a diverse array of prediction techniques, which broadly fall under physical models, learning-driven models, and statistical models based on their underlying mechanisms [8-9]. Physical models establish a mapping from wind fields to power output by integrating numerical weather prediction (NWP), turbine characteristics, and terrain information. Their primary advantage lies in strong physical interpretability. However, these models often entail significant computational overhead and are highly contingent upon the precision and reliability of meteorological inputs [10]. Statistical models are characterized by relatively simple workflows and low computational costs, and they often exhibit stable performance under near linear conditions or in cases of mild non-stationarity. Commonly used statistical approaches include the autoregressive moving average model (ARMA) [11], the autoregressive integrated moving average model (ARIMA) [12], and the Kalman filter (KF) [13]. However, in ultra-short-term scenarios characterized by high volatility, the underlying linear assumptions of these methods frequently deviate from actual system dynamics, which may lead to significant degradation in forecasting performance [14]. Data-driven approaches, encompassing both classical machine learning models and modern deep learning frameworks, have gradually become the mainstream choice, as they are capable of approximating complex nonlinear mappings and learning temporal dependencies from historical observations. Literature [15] developed a statistical wind power forecasting model that does not rely on NWP. The initial signal processing stage applies wavelet analysis to partition the raw wind velocity data. An adaptive wavelet neural network (AWNN) is then employed to regress each decomposed component. Subsequently, a feedforward neural network (FFNN) is utilized to establish the nonlinear mapping between wind speed and wind power output, thereby achieving the prediction. Literature [16] integrates the advantages of deep learning and ensemble learning. Within this system, a long short-term memory (LSTM) algorithm functions as the primary feature extractor, tasked with uncovering dynamic characteristics and tracking multi-range temporal correlations embedded within the historical data. Subsequently, the extreme gradient boosting algorithm (XGBoost) further enhances forecasting performance through nonlinear modeling and feature selection. Literature [17] combined a novel self-attention temporal convolutional network (SATCN) with a LSTM to forecast wind power generation and support continuous power supply. However, extensive practical applications have demonstrated that relying solely on learning-based predictors is insufficient to effectively address the non-stationarity and noise contamination of signals, thereby constraining the achievable upper bound of forecasting accuracy.

Consequently, hybrid frameworks that integrate signal processing techniques with learning-driven approaches have emerged as a prominent research direction, among which decomposition-integration strategies are particularly representative. The key principle of this strategy is to separate the original sequence into distinct modes that are more regular in pattern and relatively simpler to forecast. Each component is then forecast individually at the component level, followed by an integration and reconstruction process. In this way, the learning difficulty faced by a single predictive model when handling wind power series contaminated with substantial noise can be effectively reduced [18].

Early decomposition frameworks primarily relied on empirical mode decomposition (EMD) [19] and its noise-assisted variants to separate multi-scale fluctuations. Representative approaches include ensemble empirical mode decomposition (EEMD) [20] and complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) [21]. However, these recursive sifting methods lack a rigorous mathematical foundation and are highly susceptible to issues such as mode mixing, error accumulation, and endpoint effects during signal reconstruction. In contrast, variational mode decomposition (VMD) formulates the signal decomposition task as a non-recursive constrained variational optimization problem, thereby providing stronger mathematical rigor, well-defined intrinsic mode function (IMF), and enhanced robustness against noise [22]. Although VMD possesses clear theoretical advantages, its effectiveness is critically dependent on the choice of hyperparameters. Improper hyperparameter selection may lead to significant information loss or the emergence of spurious center frequencies. Therefore, heuristic optimization algorithms such as particle swarm optimization (PSO) [23], the whale optimization algorithm (WOA) [24], and the grey wolf optimization (GWO) algorithm [25] are frequently employed to adaptively determine these parameters, thereby ensuring that frequency-domain features are optimally extracted prior to the modeling stage.

After the signal decomposition process, the resulting subsequences typically exhibit different levels of dynamic complexity. To avoid computational redundancy and reduce the accumulation of forecasting errors, the subsequences are usually divided into high-frequency and low-frequency components based on their complexity. This classification facilitates subsequent targeted modeling tailored to the distinct characteristics of each component. At present, mainstream decomposition-integration frameworks primarily employ information entropy as the criterion for evaluating sequence complexity. In the literature [26], the sample entropy of each subsequence and the residual obtained from EMD is calculated to sort the components into two sets: high-frequency and low-frequency. Subsequently, an LSTM model and an ARIMA model are employed to perform forecasting for the respective components, resulting in improved predictive performance. However, entropy measures metrics essentially characterize the uncertainty or disorder of information contained in a time series. From a mathematical perspective, such disorder does not necessarily correspond to the actual predictability of the sequence for a forecasting model. Treating statistical randomness and algorithmic predictability as equivalent without rigorous justification may lead to inappropriate assignment of subsequences to forecasting models, thereby constraining the predictive performance.

To address the above methodological and computational limitations, this study proposes a VMD-PSO-LPMS-Transformer-ARFIMA hybrid wind power forecasting framework driven by SPSO-based global hyperparameter optimization. As illustrated in Fig. 1, the proposed approach improves the effectiveness of hyperparameter optimization while simultaneously enhancing forecasting accuracy. The main contributions of this study are summarized as follows:

An adaptive signal decomposition and denoising architecture is developed by integrating PSO with VMD. In this framework, the decomposition hyperparameters are objectively optimized. As a direct result, the severe multi-scale fluctuations and non-stationarity characterizing the initial wind power time series are substantially suppressed.

A frequency-aware hybrid forecasting framework is constructed to align the model representation capability with the dynamic characteristics of wind power series. To successfully model the global dependencies hidden within intricate high-frequency signals, the proposed framework integrates a Transformer network that relies heavily on its multi-head self-attention mechanism. For low-frequency components, which typically exhibit pronounced trends and long-range correlations, the autoregressive fractionally integrated moving average (ARFIMA) model is employed to rigorously characterize the fractional integration structure and the resulting long-memory persistence. Through this design, complementary modeling of high and low frequency information is achieved, enabling coordinated and more accurate forecasting.

An SPSO mechanism is introduced to enable efficient hyperparameter tuning. By employing a Gaussian Process-based surrogate model, the computational cost associated with optimizing complex deep learning architectures is substantially reduced, thereby ensuring effective operation even under limited computational resources.

2. Signal Decomposition and Sequence Partitioning Method

2.1. Variational Mode Decomposition

The dynamic profiles of wind power generation series are continuously shaped by the presence of atmospheric turbulence, intermittent variations in wind speed, and the nonlinear power conversion mechanism of wind turbines. Consequently, intense stochastic characteristics, alongside significant non-stationarity and complex multi-scale fluctuations, consistently characterize the temporal data of wind power generation. Such complex dynamics make it difficult for models trained directly on the raw time series to accurately capture the underlying patterns, thereby limiting both forecasting accuracy and model generalization capability. Therefore, it is necessary to perform structured preprocessing of wind power signals prior to predictive modeling. By decomposing the original series into several subsequences with clearer spectral characteristics and improved stationarity, the complexity of the forecasting task can be effectively reduced.

VMD decomposes complex signals into a set of IMFs with finite bandwidth within a rigorous variational optimization framework, where each mode is distributed around an adaptively determined center frequency. Compared with empirical mode decomposition-based approaches, VMD does not rely on empirical sifting criteria and exhibits well-defined mathematical properties and superior numerical stability. Consequently, it is more suitable for decomposing wind power signals characterized by high noise levels and strong non-stationarity. For the original wind power time series $P(t)$, VMD formulates the signal decomposition task as the following constrained variational optimization problem:

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

$$s.t. \sum_{k=1}^{K} u_k(t) = P(t) \tag{2}$$

where K denotes the number of decomposed modes, $\{u_k\}$ and $\{\omega_k\}$ represent the k-th decomposed mode component and its corresponding center frequency, respectively. The

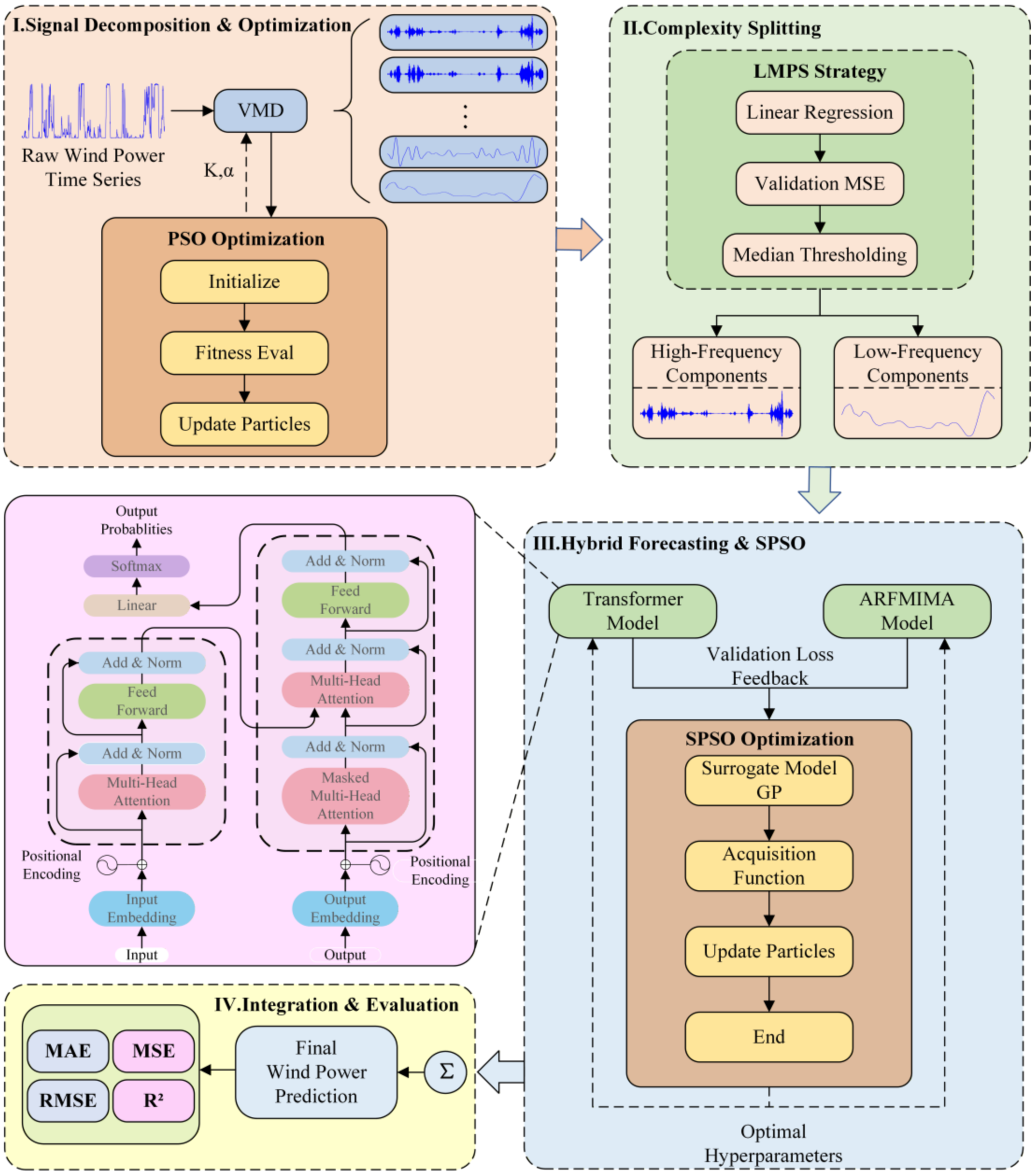

Figure 1. VMD-PSO-LPMS-Transformer-ARFIMA-SPSO frame

$\partial(t)$ denotes the partial derivative operator, $\delta(t)$ represents the Dirac function, and $*$ denotes the convolution operator. To transform the constrained variational problem into an unconstrained optimization problem, a Lagrange multiplier λ and a quadratic penalty factor α are introduced. The resulting augmented Lagrangian expression can be written as follows:

$$L(\{u_k\},\{\omega_k\},\lambda) = \alpha \sum_{k=1}^{K} \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 + \left\| P(t) - \sum_{k=1}^{K} u_k(t) \right\|_2^2 + \left\langle \lambda(t), P(t) - \sum_{k=1}^{K} u_k(t) \right\rangle \tag{3}$$

In the frequency domain, the alternating direction method of multipliers (ADMM) is employed to iteratively optimize u_k and ω_k. Based on the updated mode components and center frequencies, the Lagrange multiplier λ is subsequently updated. Through this alternating iterative procedure, the optimal solution is obtained. The detailed process is described as follows:

$$\hat{u}_k^{n+1}(\omega) = \frac{\hat{P}(\omega) - \sum_{i \neq k}^{K} \hat{u}_i^{n}(\omega) + \frac{\hat{\lambda}^{n}(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k^{n})^2} \tag{4}$$

$$\omega_k^{n+1} = \frac{\int_0^{\infty} \omega \left| \hat{u}_k^{n+1}(\omega) \right|^2 d\omega}{\int_0^{\infty} \left| \hat{u}_k^{n+1}(\omega) \right|^2 d\omega} \tag{5}$$

$$\hat{\lambda}^{n+1}(\omega)=\hat{\lambda}^{n}(\omega)+\gamma\left(\hat{P}(\omega)-\sum_{k=1}^{K}\hat{u}_k^{n+1}(\omega)\right) \tag{6}$$

2.2. Particle Swarm Optimization

PSO is a population-based global optimization algorithm inspired by swarm intelligence. The fundamental concept of PSO originates from the social behavior observed in bird flock foraging. By simulating the collective search process of birds seeking food sources within a search space, the algorithm performs optimization through cooperative information sharing among particles. The optimization procedure can be described as follows:

$$H_i^{(k+1)}=\omega V_i^k+\lambda_1 m_1\left(P_{best}-Y_i^k\right)+\lambda_2 m_2\left(G_{best}-Y_i^k\right) \tag{7}$$

$$Y_i^{(k+1)}=Y_i^k+H_i^{(k+1)} \tag{8}$$

where ω denotes the inertia weight coefficient, λ_1 and λ_2 represent the cognitive and social acceleration coefficients, respectively. P_{best} and G_{best} denote the individual best position and the global best position found by the swarm.

2.3. Linear Predictability Median Split

In decomposition integration forecasting frameworks, the decomposed components typically exhibit distinct frequency characteristics. Consequently, a principled component level complexity evaluation method is required to support frequency aware and complexity aware heterogeneous modeling. In this study, a linear predictability median split (LPMS) model is proposed. This approach introduces a lightweight linear proxy model to quantify the complexity of each modal component. Based on the resulting predictability scores, a median threshold partition is applied, thereby adaptively separating the IMFs components into two categories: high frequency and low frequency subsequences. As illustrated in Fig. 2.

For the k-th modal component u_k obtained from VMD decomposition, let the embedding dimension be L. The input vector at time t is given by $x_t^{(k)}=\left[u_k(t-L),u_k(t-L+1),\cdots,u_k(t-1)\right]^T$. The core strategy of LPMS assumes that each component can be temporarily approximated using a simple linear regression model, whose functional form is defined as follows:

$$\hat{u}_k(t)=W_k^T x_t^{(k)}+b_k \tag{9}$$

where W_k denotes the weight vector matrix and b_k represents the bias term. After the parameters are estimated on the training subset using the ordinary least squares (OLS) method, the model performs one step ahead forecasting for the corresponding component on the validation subset. At this stage, the linear predictability score of the k-th component, denoted as S_k, is defined as the mean squared error (MSE) on the validation set. This metric directly reflects the extent to which the component contains exploitable linear trend information:

$$S_k=\frac{1}{N_{val}}\sum_{t=1}^{N_{val}}\left(u_k(t)-\hat{u}_k(t)\right)^2 \tag{10}$$

where N_{val} denotes the number of samples in the validation set. A smaller value of S_k indicates that the temporal dependency of the component is relatively simple and is mainly characterized by stable linear trends or periodic patterns. In contrast, a larger value of S_k suggests that the component contains substantial abrupt variations, high frequency noise, or chaotic characteristics that cannot be effectively captured by linear models. To avoid the subjective bias introduced by manually specified thresholds, an adaptive partitioning strategy based on the statistical median is adopted. The median value $T_{med}=median\{S_1,S_2,\cdots,S_K\}$ is computed from the set of predictability scores corresponding to all K components and the residual term. Based on this critical threshold, the set of modal components $u=\{u_1,u_2,\cdots,u_k\}$ is partitioned into a high-frequency component subset u_h and a low-frequency component subset u_l:

$$\begin{aligned} u_l &= \{u_k \mid S_k \le T_{med}\} \\ u_h &= \{u_k \mid S_k > T_{med}\} \end{aligned} \tag{11}$$

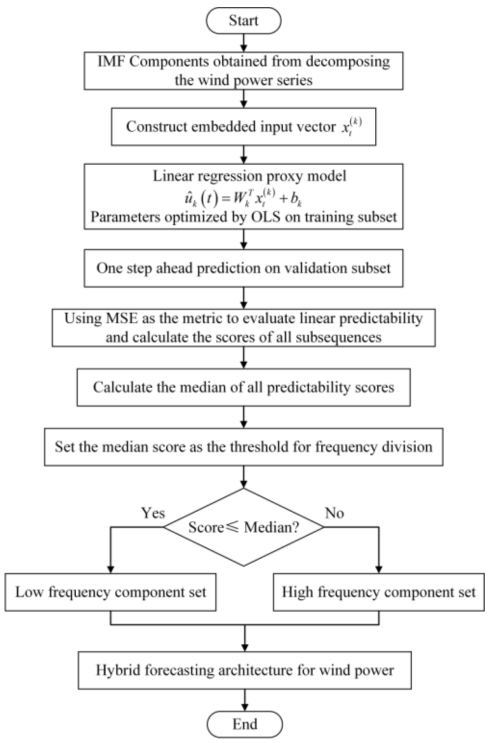

Figure 2. The framework of the LPMS Method

3. Transformer-ARFIMA-SPSO Hybrid Forecasting Framework

3.1. Transformer

The Transformer architecture was originally proposed by Vaswani et al. to address the long term dependency problem in sequence transduction tasks. It abandons the recursive structures of recurrent neural networks (RNN) and the convolution operations used in convolutional neural networks (CNN), and instead driven exclusively by the self-attention mechanism to capture global dependencies within the input sequence. In wind power forecasting tasks, generated power data alongside historical wind speed are fundamentally driven by long range temporal correlations, as well as extreme nonlinear fluctuations. Owing to its parallel computation capability and its advantage in modeling long distance dependencies, the Transformer architecture can effectively overcome the gradient vanishing and information forgetting limitations that often arise in models such as the LSTM network when processing long sequences. The architecture is illustrated in Fig. 3

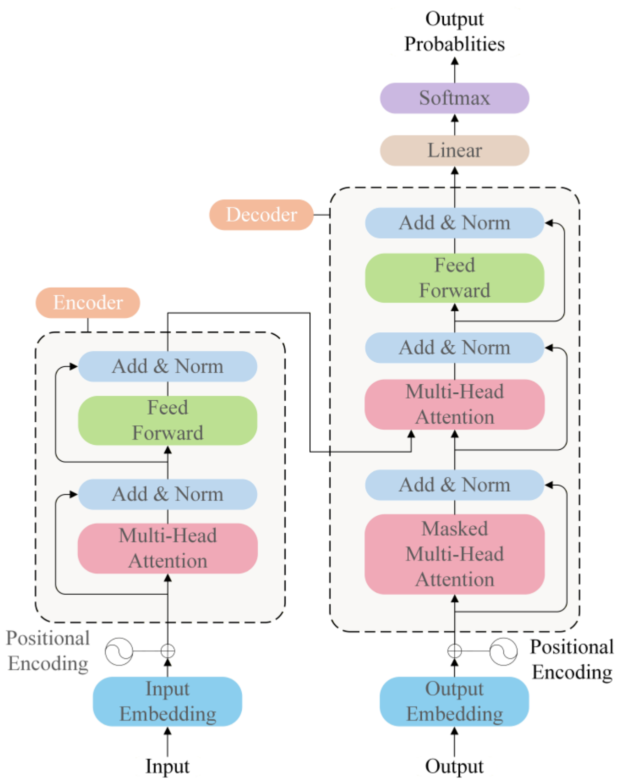

Figure 3. The framework of Transformer model

Embedding. Since the Transformer architecture does not contain recurrent structures, it inherently lacks the ability to capture the chronological arrangement of incoming data. Therefore, positional encoding is first applied to the normalized wind power time series input $X \in \mathbb{R}^{T\times d_{model}}$. Positional encoding incorporates sinusoidal and cosine functions with specific frequencies into the input embeddings, enabling the model to exploit both relative and absolute positional information. The encoding formulation is given as follows:

$$
\begin{aligned}
PE_{(pos,2i)} &= \sin\left(\frac{pos}{10000^{2i/d}}\right) \\
PE_{(pos,2i+1)} &= \cos\left(\frac{pos}{10000^{2i/d}}\right)
\end{aligned}
\tag{12}
$$

where pos denotes the absolute position index and i represents the embedding dimension index.

Encoder. The encoder consists of N identical layers stacked sequentially, aiming to map the input sequence into high dimensional contextual representations. Each layer contains two principal sublayers: a multi head self attention mechanism and a position wise feedforward network. Before performing the attention computation, the input vectors are first projected through learnable linear transformation matrices to generate three sequences, namely the query Q, key K, and value V:

$$Q = XW_Q,\ K = HW_K,\ V = HW_V \tag{13}$$

The output of the attention module can then be expressed as follows:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{14}$$

where d_k denotes the dimensionality of the key vectors. The architecture integrates a multi-head attention mechanism, which facilitates the extraction of diverse features distributed across various representation subspaces. The outputs of all attention heads are first concatenated and then projected through an output linear transformation matrix, thereby restoring the original model dimension and integrating features from multiple perspectives:

$$head_i = Attention\left(XW_i^Q, XW_i^K, XW_i^V\right) \tag{15}$$

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}\left(head_1, head_2, \cdots, head_h\right)W^O \tag{16}$$

where W_i^Q, W_i^k and W_i^v denote the weight matrices of the i-th attention head, and W^O represents the linear transformation matrix used to project the concatenated outputs into the output space. The operator $\mathrm{Concat}(\cdot)$ denotes the concatenation of the outputs from h attention heads along the feature dimension.

Functioning as an essential counterpart to the attention sublayer, a fully connected feedforward network (FFN) is embedded into every individual block comprising the Encoder and Decoder modules. This network processes each position of the sequence independently and consists of two linear transformations with a ReLU activation function in between, which enhances the nonlinear representation capability of the model:

$$FFN(X) = ReLU(XW_1 + B_1)W_2 + B_2 \tag{17}$$

where X denotes the input vector, W_1 and W_2 represent the weight matrices, B_1 and B_2 denote W_1 the bias terms.

To facilitate stable gradient propagation in deep networks, each sublayer in the Encoder is followed by a residual connection and layer normalization:

$$Output = LayerNorm(X + Sublayer(X)) \tag{18}$$

Decoder. The decoder also consists of N stacked layers and is responsible for predicting the next token based on the encoder output and the previously generated sequence. In terms of architectural design, the Decoder closely mirrors the foundational layout established by the Encoder. However, the self attention component must additionally handle the interaction between the target sequence and the encoder output. To prevent information leakage from future time steps, a masking mechanism is introduced in the self-attention layer, ensuring that the model only attends to previously generated elements during the decoding process.

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V \tag{19}$$

where M denotes the masking matrix.

In the encoder decoder attention sublayer of the Decoder, the sources of the attention inputs are different. The query Q is derived from the output of the previous Decoder layer H, whereas the key K and value V originate from the final output of the Encoder. This design enables the decoding process to dynamically retrieve relevant information from the source sequence. These intermediate representations proceed into a feedforward network. This stage is structurally reinforced by integrating layer normalization alongside residual connections. Ultimately, deriving the final predicted output sequence requires routing these refined characteristics into a linear projection layer, culminating in a softmax transformation.

3.2. Autoregressive Fractionally Integrated Moving Average

The autoregressive fractionally integrated moving average (ARFIMA) model represents a generalized extension of the conventional ARIMA framework and is specifically designed to capture the inherent long range dependence observed in complex wind power time series. Unlike the standard ARIMA model, which enforces integer order differencing and may therefore lead to over differencing or fail to represent slowly decaying autocorrelation structures, the ARFIMA model introduces a fractional differencing parameter d. This mechanism enables rigorous modeling of the hyperbolic decay behavior of the autocorrelation function. The overall architecture is illustrated in Fig. 4. For a discrete time series X_t, the ARFIMA process can be mathematically defined by the following difference equation:

$$\varphi(B)(1-B)^d(X_t - \mu) = \theta(B)e_t \tag{20}$$

where μ denotes the mean of the series, e_t represents a white noise sequence with zero mean and variance σ_e^2, B denotes the lag operator. The polynomials $\varphi(B)$ and $\theta(B)$ correspond to the autoregressive (AR) and moving average (MA) components of orders p and q, respectively, which characterize the short range dependence structure of the process. These polynomials can be expressed as follows:

$$\begin{aligned}\varphi(B) &= 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p \\ \theta(B) &= 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\end{aligned} \tag{21}$$

The core innovation of the ARFIMA model lies in the fractional differencing operator $(1-B)^d$, where d is a real number rather than an integer. When $d \in (-0.5, 0.5)$, the process remains stationary and invertible. By executing an infinite binomial expansion, this specific operator empowers the forecasting model to concurrently characterize macroscopic low frequency trends and microscopic high frequency fluctuations. Using the Gamma function $\Gamma(\cdot)$, the fractional differencing operator can be expanded into an infinite series of the lag operator as follows:This is the body text with indent. This is the body text with indent. This is the body text with indent. This is the body text with indent. This is the body text with indent. This is the body text with indent. This is the body text with indent. This is the body text with indent. This is the body text with indent. This is the body text with indent. This is the body text with indent. This is the body text with indent.

$$(1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k = \sum_{k=0}^{\infty} \frac{\Gamma(k-d)}{\Gamma(-d)\Gamma(k+1)} B^k \tag{22}$$

This expansion indicates that the current value X_t depends on an infinite sequence of past observations. As the lag k approaches infinity, the corresponding weights decay hyperbolically at a rate proportional to k^{d-1}. This behavior stands in sharp contrast to the exponential decay observed in ARMA processes and the non-decaying memory characteristic associated with integrated processes.

3.3. Surrogate Assisted Particle Swarm Optimization

Driven by its straightforward algorithmic framework and rapid convergence capabilities, PSO has seen extensive

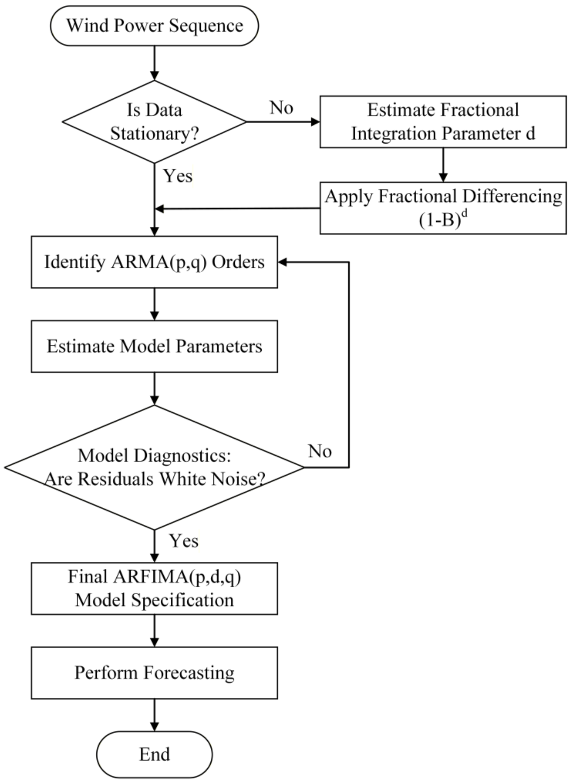

Figure 4. The framework of the ARFIMA model

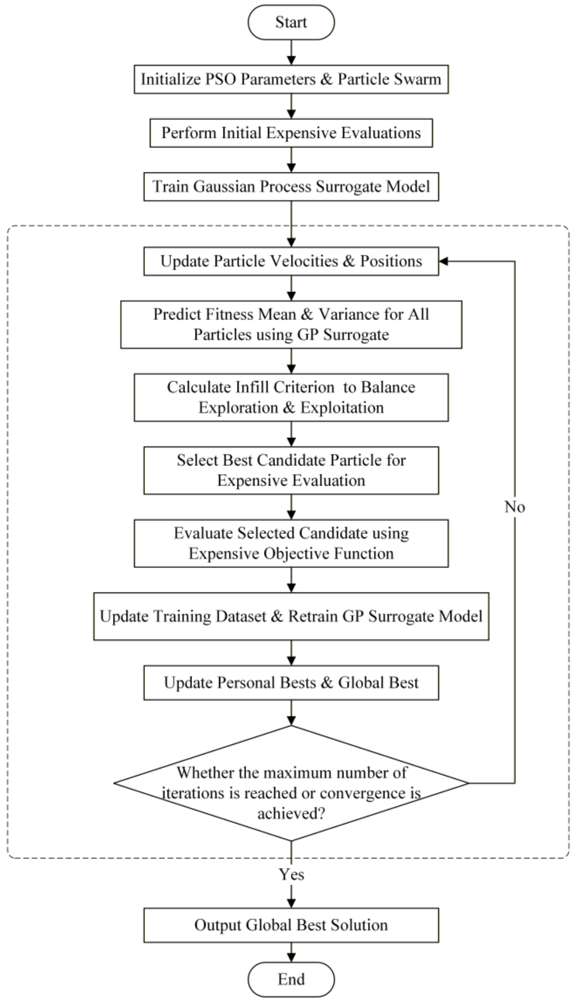

Figure 5. SPSO based hyperparameter optimization framework

utilization across global optimization scenarios. Fundamentally, this algorithm operates as a stochastic evolutionary computation method rooted in the principles of swarm intelligence. However, in the context of hyperparameter optimization for complex deep learning architectures, the evaluation of the fitness function typically requires training neural networks on large scale datasets, which results in extremely high computational costs. To alleviate this computational burden while preserving the exploration capability of the particle swarm, this study adopts a surrogate assisted particle swarm optimization (SPSO) framework, as illustrated in Fig. 5. To handle the high computational cost, the SPSO approach utilizes a surrogate model to estimate the objective function., thereby guiding the swarm toward the global optimum with significantly fewer evaluations of the true objective function.

Within the SPSO framework, a Gaussian Process (GP) is adopted as the surrogate model because it provides not only predictive estimates but also the predictive variance at unobserved points. Assume that the objective function $f(x)$ follows a Gaussian Process distribution with mean function μ and covariance function $k(x,x')$. The correlation between any two points x_i and x_j is characterized by the kernel function $R(x_i, x_j, \theta)$, for which the Gaussian kernel is commonly employed. For an unobserved candidate solution x^*, the predictive mean $\hat{f}(x^*)$ and predictive variance $\hat{s}^2(x^*)$ can be derived as follows:

$$\begin{aligned} \hat{f}(x^*) &= \hat{\mu} + \mathbf{d}^T \mathbf{L}^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu}) \\ \hat{s}^2(x^*) &= \hat{\sigma}^2 \left[1 - \mathbf{d}^T \mathbf{L}^{-1} \mathbf{d} + \frac{\left(1 - \mathbf{1}^T \mathbf{L}^{-1} \mathbf{d}\right)^2}{\mathbf{1}^T \mathbf{L}^{-1} \mathbf{1}} \right] \end{aligned} \tag{23}$$

where $\mathbf{y}$ denotes the vector of observed fitness values, $\mathbf{L}$ represents the correlation matrix of the observed samples, $\mathbf{d}$ denotes the correlation vector between the observed samples and the candidate point x^*, and $\mathbf{1}$ is the unit column vector. To balance exploration and exploitation during the optimization process, an infill sampling criterion based on the lower confidence bound (LCB) is adopted to identification of candidate particles with high potential for exact evaluation. The LCB criterion constructs an acquisition function by subtracting a weighted uncertainty term from the predictive mean, which can be expressed as follows:

$$LCB(x) = \hat{f}(x) - \alpha \hat{s}(x) \tag{24}$$

where α denotes a tuning parameter that controls the trade off between exploration and exploitation. Accordingly, the particle associated with the lowest LCB value is exclusively selected for evaluation using the true objective function. The obtained result is then utilized to update the global best solution while correcting the surrogate model. The remaining particles are instead evaluated using the Gaussian Process approximation, thereby enabling efficient optimization under limited computational resources.

4. Case Analysis

4.1. Datasets Description

The dataset employed in this study consists of wind power measurements obtained from a wind farm situated in Xinjiang, China. As illustrated in Fig. 6, the dataset contains the operational records of the wind farm for the entire year of 2019. The sampling interval is 15 min, and the total installed capacity of the wind farm is 199.5 MW. To assess the robustness and predictive performance of the proposed model across varying meteorological conditions, two datasets with pronounced seasonal differences were selected for comparative analysis and ablation experiments, namely the measured wind power data from Spring (April) and Autumn (October).

4.2. Evaluation Metrics and Signal Decomposition

In this study, four standard performance indicators are employed to assess model accuracy, including the mean absolute error (MAE), mean squared error (MSE), root mean square error (RMSE), and the coefficient of determination (R^2):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{25}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 \tag{26}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2} \tag{27}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{n}\left(\bar{y} - y_i\right)^2} \tag{28}$$

where y_i denotes the observed value, $\bar{y}$ represents the mean of the observed values, $\hat{y}_i$ denotes the predicted value, and n is the total number of data samples.

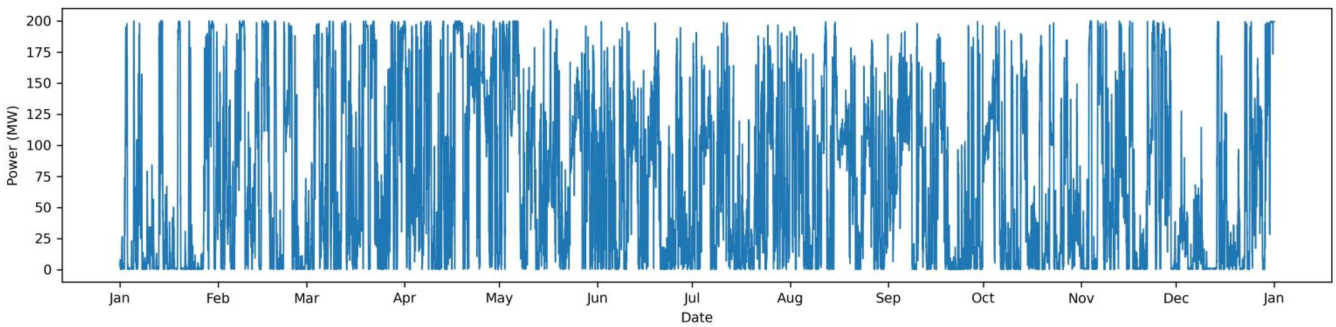

Figure 6. Original sequence

Wind power output is characterized by large ramp amplitudes and strong volatility. Consequently, relying on expert experience to manually configure the VMD hyperparameters cannot ensure sufficient decomposition and denoising of the original power series. Moreover, empirically fixed parameters generally lack cross seasonal transferability. In addition, the proposed LPMS frequency partition strategy allocates IMF components according to a linear predictability proxy. If the VMD decomposition is inadequate, component misallocation may occur during the LPMS partition process. To address this issue, PSO is introduced in the VMD stage to perform a global search for the hyperparameters K and α . This strategy transforms the decomposition procedure from manual trial and error into a data driven adaptive selection process. As a result, the stability of the decomposition and its cross seasonal robustness are improved at the source, this process establishes a more dependable signal basis, which facilitates the stable convergence of the error distributions in the subsequent forecasting stage. The VMD decomposed signals and the corresponding hyperparameter selections for

April and October are presented in Figs. 7-8 and Table 1, respectively.

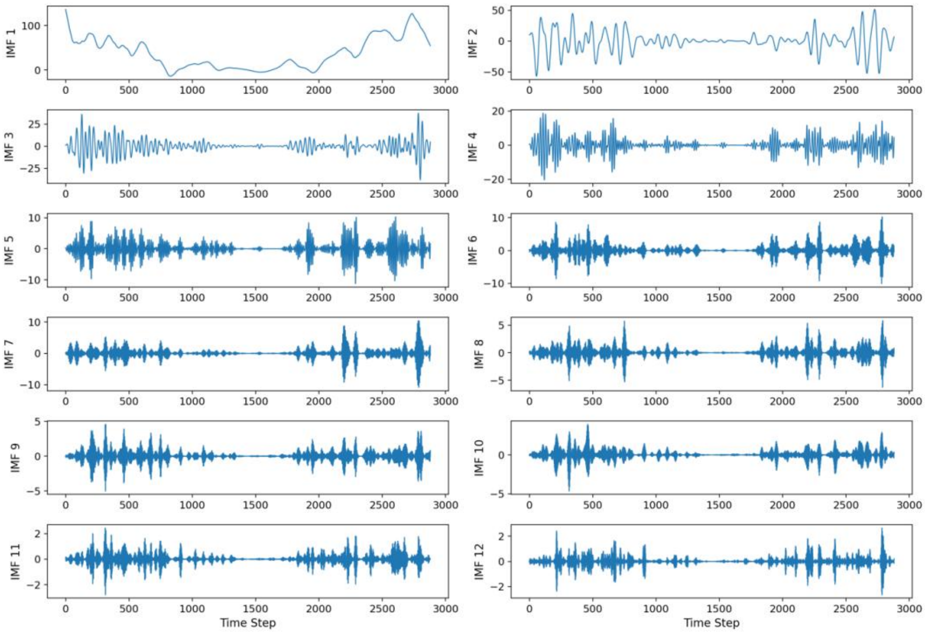

Figure 7. IMF components of the April wind power series after VMD-PSO decomposition

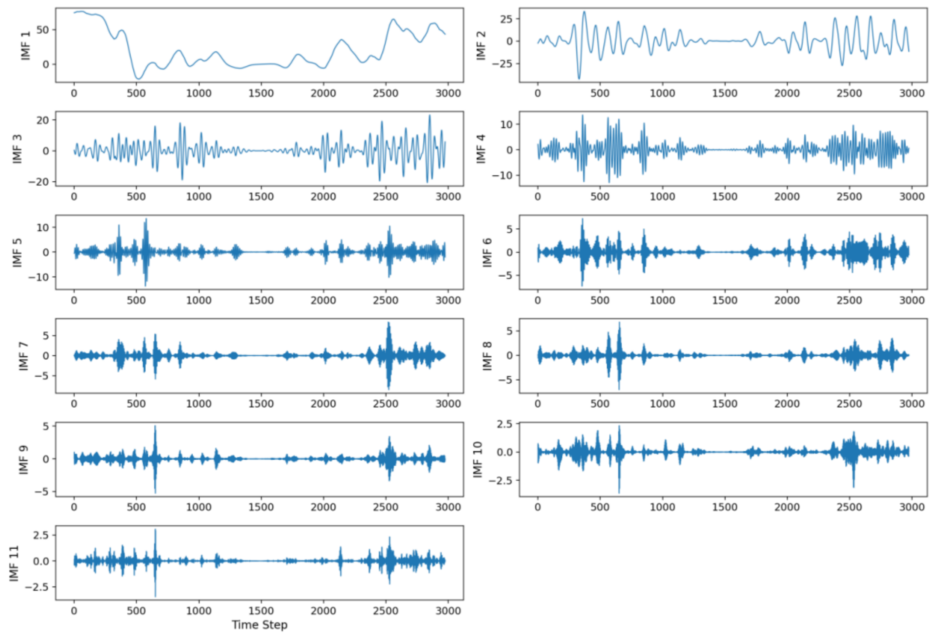

Figure 8. IMF components of the October wind power series after VMD-PSO decomposition

Table 1. PSO optimized VMD hyperparameters

Month	K	α
April	12	300
October	11	340.2815

4.3. Comparative Study Analysis

To verify the superiority of the proposed model, several benchmark methods were selected for rigorous comparison. These include the single model XGBoost, and LSTM, the conventional hybrid architecture CNN-LSTM, and two decomposition forecasting frameworks, namely EEMD-Transformer and CEEMDAN-CNN-Transformer.

The comprehensive performance evaluation metrics of all models for April and October are reported in Tables 2-3, respectively, the corresponding prediction curves for model comparison are presented in Figs. 9-10.

Table 2. Performance evaluation metrics of the compared models in April

Model	MAE (MW)	MSE (MW^2)	RMSE (MW)	R^2
XGBoost	6.2617	93.8280	9.6865	0.9775
LSTM	5.6759	85.4419	9.2435	0.9795
CNN-LSTM	5.6063	83.3708	9.1308	0.9710
EEMD-Transformer	5.5107	46.6671	6.8313	0.9888
CEEMDAN-CNN-Transformer	7.1059	103.6302	10.1799	0.9751
Proposed Model	3.5806	31.2753	5.5924	0.99245

Based on the data in the table, the proposed model consistently yields the smallest prediction errors while attaining the highest R^2 values across all evaluated scenarios. This demonstrates its enhanced predictive accuracy as well as its robustness under varying seasonal conditions.

For the April dataset, the proposed model obtains an MAE of 3.5806 MW, an MSE of 31.2753 MW^2, and an RMSE of 5.5924 MW. Compared with the second best performing EEMD-Transformer mode0l, the proposed method reduces the RMSE by approximately 18.1 %, demonstrating a substantial improvement in error suppression capability. Meanwhile, the CEEMDAN-CNN-Transformer framework yields an RMSE as high as 10.1799 MW on the same dataset. Relative to this model, the proposed approach reduces the RMSE by approximately 45.1 %. These results indicate that relying solely on empirical mode decomposition is insufficient to fully mitigate the complex mode mixing phenomenon inherent in wind power time series. In contrast, the proposed VMD-PSO decomposition combined with the LPMS high frequency and low frequency partition strategy enables more targeted modeling of different modal components according to their predictability characteristics. Consequently, the feature extraction efficiency of the subsequent forecasting networks is improved and the prediction error can be more effectively controlled.

Table 3. Performance evaluation metrics of the compared models in October

Model	MAE (MW)	MSE (MW^2)	RMSE (MW)	R^2
XGBoost	5.4716	70.0165	8.3676	0.9682
LSTM	5.5257	69.4060	8.3310	0.9684
CNN-LSTM	5.8954	75.8357	8.7084	0.9655
EEMD-Transformer	4.3022	33.4322	5.7820	0.9848
CEEMDAN-CNN-Transformer	5.9221	56.3227	7.5048	0.9744
Proposed Model	3.0843	20.2098	4.4955	0.9908

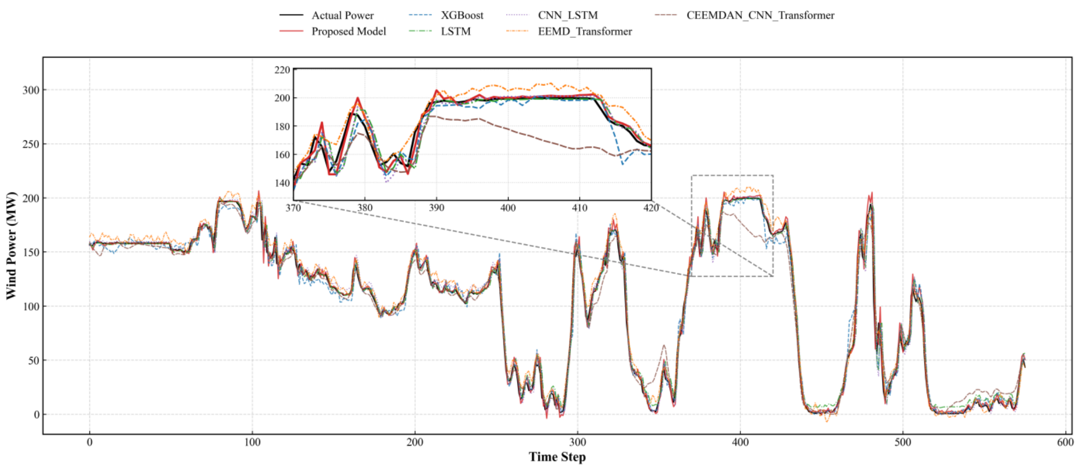

Figure 9. Prediction curves of the compared models in April

For the October dataset, where wind conditions exhibit more pronounced fluctuations, the proposed model continues to exhibit a high level of fitting accuracy, achieving an R^2 value of 0.9908. In contrast, the single models XGBoost and LSTM show evident performance degradation when dealing with high frequency non stationary sequences, with RMSE values reaching 8.36767 MW and 8.3310 MW, respectively. In contrast, the proposed model reduces the RMSE by approximately 46.3 % and 46.0 %, respectively. These results indicate that, in modeling wind power series, a single network architecture or conventional machine learning model struggles to concurrently capture long term trends and local abrupt variations. By contrast, the proposed hybrid architecture is able to more effectively characterize the complex nonlinear dynamics inherent in wind power generation processes.

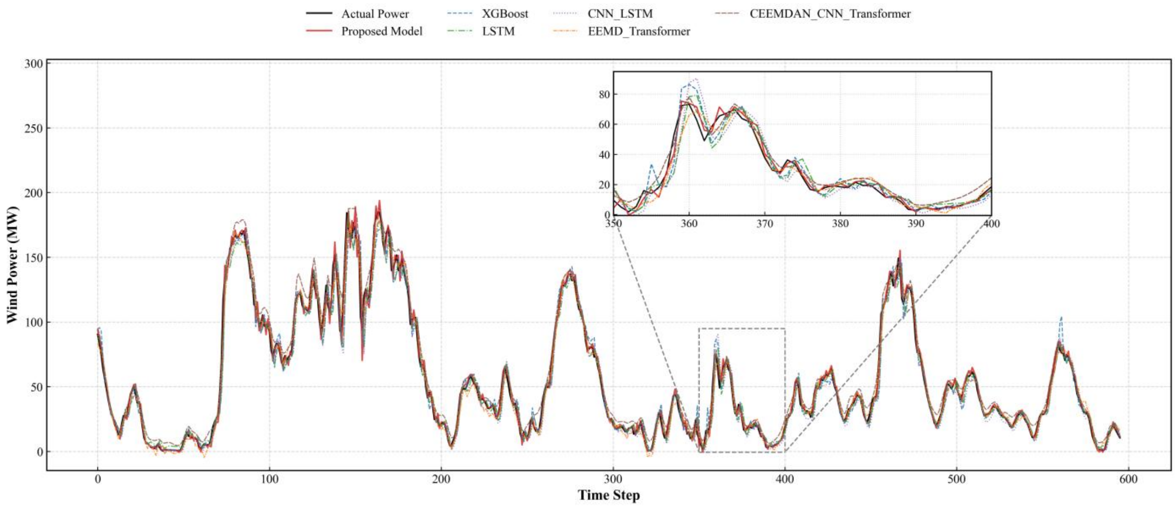

Figure 10. Prediction curves of the compared models in October

4.4. Ablation Study Analysis

To systematically assess the individual contributions of each key component within the proposed hybrid framework, four ablation variants were constructed by removing individual components: VMD-LPMS-Transformer-ARFIMA-SPSO, VMD-PSO-LPMS-ARFIMA-SPSO, VMD-PSO-LPMS-Transformer-SPSO and VMD-PSO-LPMS-Transformer-ARFIMA. The evaluation metrics and prediction curves for April and October are presented in Tables 4-5 and Figs. 11-12, respectively.

As shown in Tables 4-5, the model performance deteriorates noticeably when the PSO optimized VMD decomposition is removed. In the April dataset, the RMSE of VMD- LPMS-Transformer-ARFIMA-SPSO reaches 7.0074 MW, whereas the proposed model reduces this value by approximately 20.2 %. In the October dataset, the RMSE of VMD- LPMS-Transformer-ARFIMA-SPSO is 5.8123 MW, and the proposed model further decreases the error by about 22.7 %. These results indicate that PSO optimized VMD parameters play a critical role in improving modal separation quality. Without parameter optimization, VMD cannot adaptively perform refined decomposition according to the time varying characteristics of wind power series across different months, thereby weakening the effectiveness of the subsequent high frequency and low frequency heterogeneous modeling.

Table 4. Performance evaluation metrics of the ablation models in April

Model	MAE (MW)	MSE (MW^2)	RMSE (MW)	R^2
VMD-LPMS-Transformer-ARFIMA-SPSO	4.3523	49.1030	7.0074	0.9882
VMD-PSO-LPMS-ARFIMA-SPSO	4.2855	40.6516	6.3759	0.9902
VMD-PSO-LPMS-Transformer-SPSO	8.0207	132.6115	11.5157	0.9681
VMD-PSO-LPMS-Transformer-ARFIMA	4.1247	35.2810	5.9398	0.9915
Proposed Model	3.5806	31.2753	5.5924	0.9925

The ablation of the ARFIMA module leads to a substantial deterioration in prediction accuracy. For the April dataset, once ARFIMA is removed, the model MSE increases sharply from 31.2753 MW2 to 132.6115 MW2, while the RMSE rises from 5.5924 MW to 11.5157 MW, representing the worst performance among all ablation configurations. In comparison, the proposed model reduces these two metrics by approximately 76.4 % and 51.4 %, respectively. In the October dataset, the VMD-PSO-LPMS-ARFIMA-SPSO configuration yields an MAE of 5.2655 MW and an RMSE of 7.0920 MW. These results strongly indicate that wind power time series are largely governed by long memory characteristics and fractionally integrated linear trends across temporal scales. Within the proposed architecture, the ARFIMA module plays a critical role in modeling the global linear trend structure. Without this component, the deep learning network loses robustness when handling long period quasi stationary fluctuations.

Table 5. Performance evaluation metrics of the ablation models in October

Model	MAE (MW)	MSE (MW^2)	RMSE (MW)	R^2
VMD-LPMS-Transformer-ARFIMA-SPSO	4.0329	33.7826	5.8123	0.9846
VMD-PSO-LPMS-ARFIMA-SPSO	3.8706	30.0011	5.4773	0.9863
VMD-PSO-LPMS-Transformer-SPSO	5.2655	50.2964	7.0920	0.9771
VMD-PSO-LPMS-Transformer-ARFIMA	3.5059	23.3209	4.8292	0.9894
Proposed Model	3.0843	20.2098	4.4955	0.9908

Eliminating the Transformer component leads to a marked degradation in model performance. In the October dataset, the MAE increases from 3.0843 MW to 3.8706 MW, while in the April dataset the RMSE rises from 5.5924 MW to 6.3759 MW. These results indicate that a single statistical model is inadequate for precisely modeling the complex nonlinear variations caused by physical factors such as wind turbulence and turbine yaw dynamics. In contrast, the Transformer framework enables effective representation of complex short-term fluctuations driven by wind turbulence, meteorological disturbances, and changes in turbine operating states, making it a critical component for improving forecasting accuracy.

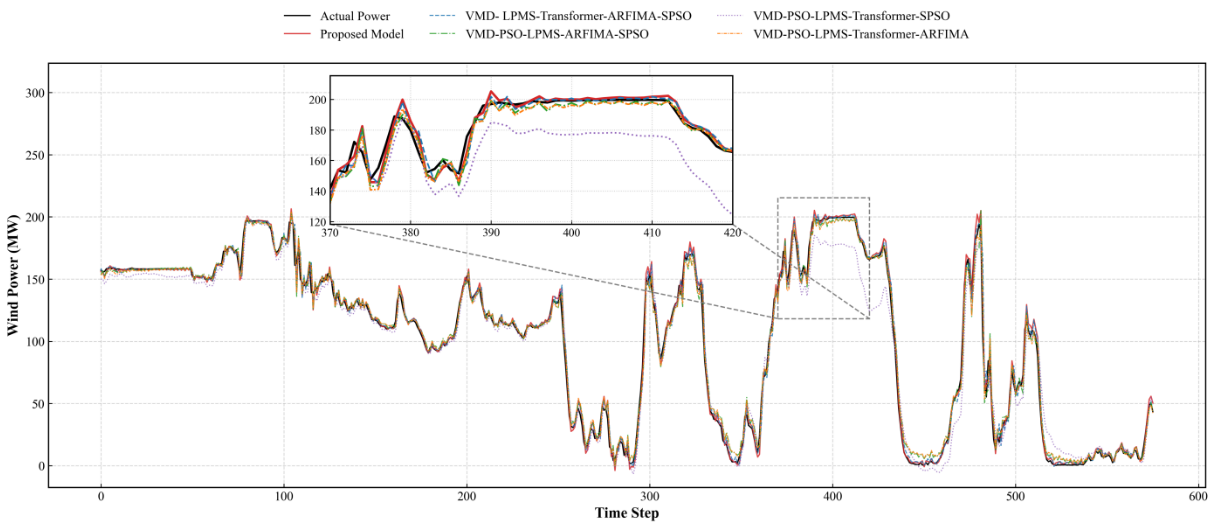

Figure 11. Forecasting trajectories of various ablation models on the April dataset

Finally, the SPSO based hyperparameter optimization module also contributes consistently to the improvement of model performance. Compared with VMD-PSO-LPMS-Transformer-ARFIMA, the proposed model reduces the MAE and RMSE on the April dataset by approximately 13.2 % and 5.8 %, respectively, while on the October dataset the reductions reach approximately 12.0 % and 6.9 %. Although the performance gains brought by this component are not as pronounced as those achieved by the Transformer module or the optimized VMD decomposition, its contribution remains non negligible. These results indicate that, within a complex hybrid forecasting architecture, the hyperparameter space often contains numerous local optima. Relying solely on empirical parameter settings makes it difficult to fully exploit the collaborative capability of different submodules. The introduction of SPSO therefore improves the rationality of parameter configuration and raises the overall performance ceiling of the model.

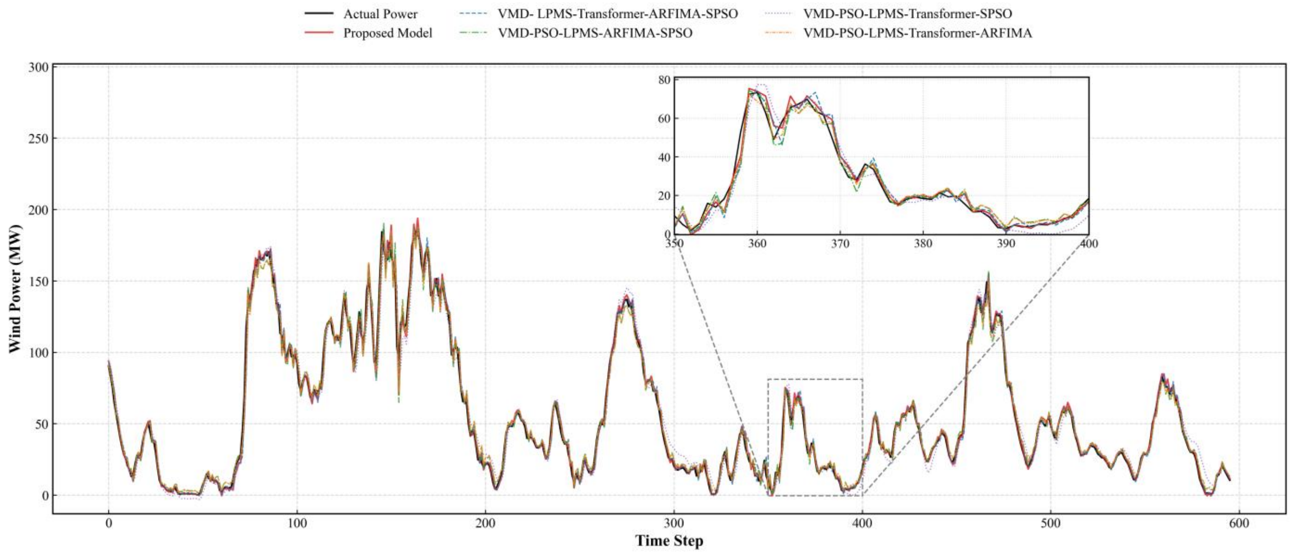

Figure 12. Forecasting trajectories of various ablation models on the October dataset

4.5 Computational Cost Analysis of SPSO

To further verify the computational efficiency of the proposed SPSO method, this study compares the iteration time of PSO and SPSO during hyperparameter optimization process under the same settings of hyperparameter search space, population size, and maximum number of iterations. As reported in Table 6, the conventional PSO required 1450.21 s to complete the iterative optimization, whereas SPSO reduced the iteration time to 456.94 s. This corresponds to an iteration speedup of approximately 3.17 times and a reduction of 68.49% in iteration time relative to PSO.

These results indicate that SPSO can effectively alleviate the computational burden associated with iterative hyperparameter optimization in the proposed LPMS-Transformer-ARFIMA hybrid forecasting framework.

Table 6.Optimization performance comparison between PSO and SPSO

Method	Iteration time (s)
PSO	1450.21
SPSO	456.94

5. Conclusion

This study develops a hybrid framework aimed at ultra-short-term wind power forecasting, namely VMD-PSO-LPMS-Transformer-ARFIMA, and introduces the SPSO algorithm to achieve global hyperparameter optimization. Initially, the original non-stationary wind power time series is decomposed into multiple components using VMD, thereby effectively suppressing mode mixing and reducing the multi scale complexity of the sequence. Subsequently, considering the inherent differences in complexity and predictability among the decomposed subsequences, an LPMS strategy is designed and introduced. According to the predictability indicator of each subsequence, the components are divided into low-frequency components representing macroscopic trends and high-frequency components characterized by microscopic fluctuations, thereby forming a dual channel forecasting architecture with explicit functional roles.

During the forecasting stage, the low-frequency and high-frequency components are then respectively modeled using the ARFIMA model and the Transformer architecture, respectively. Meanwhile, SPSO is utilized to improve the global search capability while maintaining a manageable computational overhead cost, mitigating the likelihood of premature convergence to local optima and yielding more appropriate hyperparameter configurations. Finally, the outputs generated by the Transformer and ARFIMA branches are integrated to yield the final forecasting result, which enhancing the overall predictive performance and robustness of the framework.

To assess the effectiveness and cross-seasonal generalization capability of the proposed framework, measured wind power data collected from a wind farm in Xinjiang were adopted as the dataset for this study. Two months with significantly different meteorological conditions, April and October were selected for analysis. The forecasting performance was systematically evaluated using RMSE, MAE, MSE, and R^2, and the proposed model was compared with several baseline models and corresponding ablation configurations. The results demonstrate that the proposed method consistently achieves superior overall performance across all test scenarios. Compared with representative advanced models, it exhibits a stable performance advantage, while the R^2 value remains above 0.99 across different datasets.

The difficulty of wind power forecasting exhibits a pronounced seasonal dependence. In summer, extreme high temperatures affect forecasting accuracy through two primary pathways: thermal turbulence intensifies the short-term nonlinear fluctuations of the power series, while changes in air density alter the fundamental conversion relationship between wind speed and power output. In winter, the impact of blade icing is even more complex, as it not only modifies the aerodynamic performance of the blades but also introduces meteorological measurement errors, ultimately leading to several adverse effects, including abrupt power drops, switching of operating states, and aggravated non-stationarity. Although the forecasting framework proposed in this study demonstrates excellent robustness during transitional seasons, substantial further research is still required before it can be effectively extended to extreme summer and winter operating conditions.

Acknowledgements

This work was supported by Technological Upgrading Project from Northeast Branch of State Grid Corporation of China (SGDB0000DKMM2500404).

References

[1] Cui, X., Yu, X., & Niu, D.: The ultra-short-term wind power point-interval forecasting model based on improved variational mode decomposition and bidirectional gated recurrent unit improved by improved sparrow search algorithm and attention mechanism. Energy 288, 129714 (2024).

[2] Farah, S., Humaira, N., Aneela, Z., & Steffen, E.: Short-term multi-hour ahead country-wide wind power prediction for Germany using gated recurrent unit deep learning. Renewable and Sustainable Energy Reviews 167, 112700 (2022).

[3] Yang, M., Guo, Y., Huang, T., & Zhang, W.: Power prediction considering NWP wind speed error tolerability: A strategy to improve the accuracy of short-term wind power prediction under wind speed offset scenarios. Applied Energy 377, 124720 (2025).

[4] Yu, G., Liu, C., Tang, B., Chen, R., Lu, L., Cui, C., ... & Muyeen, S. M.: Short term wind power prediction for regional wind farms based on spatial-temporal characteristic distribution. Renewable Energy 199, 599-612 (2022).

[5] D.Y. Hong, T.Y. Ji, M.S. Li, Q.H. Wu.: Ultra-short-term forecast of wind speed and wind power based on morphological high frequency filter and double similarity search algorithm. International Journal of Electrical Power & Energy Systems 104, 868-879 (2019).

[6] Zhang, Y., Sun, H., Tan, J., Li, Z., Hou, W., & Guo, Y.: Capacity configuration optimization of multi-energy system integrating wind turbine/photovoltaic/hydrogen/battery. Energy 252, 124046 (2022).

[7] Ding, Y., Chen, Z., Zhang, H., Wang, X., & Guo, Y.: A short-term wind power prediction model based on CEEMD and WOA-KELM. Renewable Energy 189, 188-198 (2022).

[8] Tsai, W. C., Hong, C. M., Tu, C. S., Lin, W. M., & Chen, C. H.: A Review of Modern Wind Power Generation Forecasting Technologies. Sustainability 15, 10757 (2023).

[9] Akhtar, I., Kirmani, S., Ahmad, M., & Ahmad, S.: Average monthly wind power forecasting using fuzzy approach. IEEE Access 9, 30426-30440 (2021).

[10] Cuevas-Figueroa, G., Stansby, P. K., & Stallard, T.: Accuracy of WRF for prediction of operational wind farm data and assessment of influence of upwind farms on power production. Energy 254, 124362 (2022).

[11] Zhang, Y., Zhao, Y., Kong, C., & Chen, B.: A new prediction method based on VMD-PRBF-ARMA-E model considering wind speed characteristic. Energy Conversion and Management 203, 112254 (2020).

[12] Wang, J., Zhu, H., Cheng, F., Zhou, C., Zhang, Y., Xu, H., & Liu, M.: A novel wind power prediction model improved with feature enhancement and autoregressive error compensation. Journal of Cleaner Production 420, 138386 (2023).

[13] Liu, Z. H., Wang, C. T., Wei, H. L., Zeng, B., Li, M., & Song, X. P.: A wavelet-LSTM model for short-term wind power forecasting using wind farm SCADA data. Expert Systems with Applications 247, 123237 (2024).

[14] Meng, A., Xie, Z., Luo, J., Zeng, Y., Xu, X., Li, Y., ... & Yin, H.: An adaptive variational mode decomposition for wind power prediction using convolutional block attention deep learning network. Energy 282, 128945 (2023).

[15] Kanna, B., & Singh, S. N.: AWNN-assisted wind power forecasting using feed-forward neural network. IEEE Trans. Sustain. Energy 3(2), 306-315 (2012).

[16] Zhu, C., Tu, Y., Wei, Q., Zang, Y., Zhou, P., Yang, L., ... & Yan, W.: Short-term power prediction of photovoltaic power station based on LSTM-XGBoost model. Solar Energy 300, 113819 (2025).

[17] Xiang, L., Liu, J., Yang, X., Hu, A., & Su, H.: Ultra-short term wind power prediction applying a novel model named SATCN-LSTM. Energy Conversion and Management 252, 115036 (2022).

[18] Sun, H., Cui, Q., Wen, J., Kou, L., & Ke, W.: Short-term wind power prediction method based on CEEMDAN-GWO-Bi-LSTM. Energy Reports 11, 1487-1502 (2024).

[19] Li, N., Dong, J., Liu, L., Li, H., & Yan, J.: A novel EMD and causal convolutional network integrated with Transformer for ultra short-term wind power forecasting. International Journal of Electrical Power & Energy Systems 154, 109470 (2023).

[20] Wu, H., Meng, K., Fan, D., Zhang, Z., & Liu, Q.: Multistep short-term wind speed forecasting using transformer. Energy 261, 125231 (2022).

[21] Zhang, C., Hua, L., Ji, C., Nazir, M. S., & Peng, T.: An evolutionary robust solar radiation prediction model based on WT-CEEMDAN and IASO-optimized outlier robust extreme learning machine. Applied energy 322, 119518 (2022).

[22] Yu, M., Niu, D., Wang, K., Du, R., Yu, X., Sun, L., & Wang, F.: Short-term photovoltaic power point-interval forecasting based on double-layer decomposition and WOA-BiLSTM-Attention and considering weather classification. Energy 275, 127348 (2023).

[23] Nguyen, T. H. T., & Phan, Q. B.: Hourly day ahead wind speed forecasting based on a hybrid model of EEMD, CNN-Bi-LSTM embedded with GA optimization. Energy Reports 8, 53-60 (2022).

[24] Yu, M., Niu, D., Gao, T., Wang, K., Sun, L., Li, M., & Xu, X.: A novel framework for ultra-short-term interval wind power prediction based on RF-WOA-VMD and BiGRU optimized by the attention mechanism. Energy 269, 126738 (2023).

[25] Mudassar, S., Zameer, A., & Raja, M. A. Z.: GWO-DAGRU: A hybrid deep learning framework with metaheuristic feature selection and self-weighted context GRU for short-term wind

power forecast. Expert Systems with Applications 310, 131279 (2026).

[26] Liu, M. D., Ding, L., & Bai, Y. L.: Application of hybrid model based on empirical mode decomposition, novel recurrent neural networks and the ARIMA to wind speed prediction. Energy Conversion and Management 233, 113917 (2021).