



Research
Next Gen-AI Enabled New Type of Power System—Article

Wind Power Forecasting Based on Large Time Series Model

Jie Yan ^{a,*}, Yujia Li ^{b,*}, Han Wang ^b, Shuang Han ^a, Wenlong Shang ^{c,*}, Yongqian Liu ^a

^a State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources (NCEPU), School of New Energy, North China Electric Power University, Beijing, 102206, China

^b School of New Energy, North China Electric Power University, Beijing 102206, China

^c Department of Civil and Environmental Engineering, Imperial College London, London SW7 2AZ, UK

ARTICLE INFO

Article history:

Received 1 May 2025

Revised 1 August 2025

Accepted 16 November 2025

Keywords:

Wind power forecasting

Moirai large time series model

Few-shot

Zero-shot

Large wind power forecasting foundation model

ABSTRACT

To address the insufficient forecasting accuracy in newly built wind farms due to a lack of historical data, we propose wind power forecasting methods based on the Moirai large time series model. For the few-shot scenario, we fine tune Moirai on limited local samples, leveraging its generic time series representations learned from the large scale LOSTA dataset and adapting them to the target wind farm to mitigate overfitting under data scarcity. In the zero-shot scenario, we further pre train Moirai on a large scale, heterogeneous wind power dataset to build a large wind power forecasting foundation model with adaptive mapping dimensions. Experiments demonstrate that our few-shot approach significantly outperforms conventional deep neural networks under data-scarce conditions, meanwhile, our zero shot foundation model for wind power forecasting achieves 86.57% accuracy for ultra short term and 77.59% accuracy for short term forecasting, confirming its strong generalization and robustness across diverse wind farm environments.

© 2025 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

With the acceleration of industrialization, conventional energy sources, characterized by environmental pollution, severe waste, and resource depletion, no longer satisfy the requirements for economic growth and sustainable development [1]. Clean, environmentally friendly renewable energies, such as wind and solar power, are receiving widespread global attention [2]. As a vital component of the modern energy system, wind power generation has experienced rapid growth in recent years due to supportive policies and technological advances [3]. As the installed capacity and the proportion of grid-connected continue to increase, the uncertainty of wind power output has an increasingly significant impact on grid operation. Wind power output is constrained by multiple factors such as wind speed, wind direction, terrain, and humidity, exhibiting randomness, volatility, and intermittency, which pose severe challenges to grid dispatch and operational stability. Therefore, accurate wind power forecasting is essential for effective power system deployment and for reducing generation costs for energy producers [4].

Existing wind power forecasting methods can be divided into three categories: statistical model-based methods, physical model-based methods, and deep neural network-based methods. Deep neural network models have been widely applied in wind power forecasting due to their strong nonlinear data-fitting and parameter-learning capabilities [5]. Their core idea is to train model parameters using historical wind power data and learn the relationship between inputs and outputs [6]. Common deep neural network architectures include convolutional neural networks (CNN) [7], long short-term memory networks (LSTM) [8], gated recurrent units (GRU) [9], and Transformer [10]. Transformer can capture long-range dependencies within sequences and fully extract relevant information from wind power data [11], and have attracted considerable attention in recent years. In Ref. [12], a short-term multi-step wind power forecasting method based on spatiotemporal correlations and Transformer was proposed, which incorporates spatial information from surrounding turbines and uses Transformer to learn distance-invariant sequential dependencies, demonstrating superior performance at large forecasting horizons.

The above deep neural networks typically require large amounts of historical data for data-driven and perform poorly in few-shot scenarios [13]. Each year, numerous wind farms are constructed, expanded, or retrofitted. Owing to their short operating

* Corresponding authors.

E-mail addresses: yanjie@ncepu.edu.cn (J. Yan), liyujia011025@163.com (Y. Li), wenlong.shang12@imperial.ac.uk (W. Shang).

<https://doi.org/10.1016/j.eng.2025.11.007>

2095-8099/© 2025 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

histories, these farms have not accumulated sufficient data to establish high-accuracy deep neural network forecasting models [14]. Currently, the most commonly used method for few-shot wind power forecasting is the “pretraining-fine-tuning” method of transfer learning [15], which still faces three major challenges: firstly, transfer learning assumes that the source wind farms and the target wind farms have a high degree of homogeneity, and if the two wind farms have significant differences in terms of geography, turbine types or climate conditions, negative transfer is likely to occur, which may weaken the model performance; secondly, due to the constraints of data privacy and commercial confidentiality policy [16], it is often difficult to obtain enough historical data of neighboring wind farms with qualified quality, which leads to limited selection of the pre-training sources; and thirdly, the traditional deep neural network fixes the input and output dimensions at the pre-training stage, and lacks of flexibility in fine-tuning to fit the downstream tasks with different forecasting spans and number of variables, which hinders the achievement of truly efficient and general predictions.

In recent years, the rapid development of large-model technologies has opened new avenues for modeling in complex scenarios. Large language models such as GPT-4 [17], PaLM [18], and DeepSeek [19] have demonstrated powerful contextual understanding and generative capabilities in natural language processing through pretraining on massive text corpora and scaling of parameter counts. Inspired by these advances, many researchers have begun to explore large models for time series analysis, including TimeGPT [20], Lag-Llama [21], Moirai [22], and Timer [23]. Ref. [24] investigates the application of TimeGPT to load forecasting by leveraging its universal patterns learned from cross-domain time series data and fine-tuning on limited historical load records. This approach achieves significant improvements in short-term forecasting accuracy under data-scarce conditions and offers practical guidance for few-shot load forecasting. However, since TimeGPT is not yet open source and is accessible only via a restricted application programming interface (API), its widespread industrial adoption remains limited. The large time series model (LSTM) Lag Llama was trained on the Monash dataset, which contains fewer than one billion data points, consequently, its ability to learn truly generalizable time-series features is limited, and it does not support downstream forecasting for an arbitrary number of variables. The Timer model, while achieving strong forecasting performance on specific downstream tasks, is constrained by its architecture and cannot accommodate arbitrary input or output lengths, limiting its overall versatility.

The Moirai LSTM is the first fully open-source foundation model for time series. It is a general-purpose transformer for forecasting that employs a masked-encoder architecture, multi patch size projection layers, and an any-variate attention mechanism to accommodate varying temporal resolutions, arbitrary input or output lengths and numbers of variables across tasks [22]. Pretrained on the LOSTA dataset, a large-scale open time series collection designed for forecasting, which includes over 27 billion observations from diverse domains, Moirai has exhibited excellent zero-shot and few-shot fine-tuning performance across multiple benchmarks, demonstrating its versatility and strong generalization ability in time series forecasting.

Based on the above background, we aim to explore the application potential of the Moirai LSTM in few-shot and zero-shot wind power forecasting scenarios. As shown in Fig. 1, our study is organized around two main aspects:

- (1) For wind farms with limited historical data, we investigate an efficient fine-tuning strategy for Moirai based on the few-shot samples. This approach both leverages the universal time-series pattern representations learned by Moirai

during large-scale pretraining and adapts the model to the local temporal structures and data-distribution characteristics of the target wind farm via fine-tuning, thereby improving forecasting performance in specific few-shot scenario.

- (2) For wind farms that completely lack historical observations, we construct a large-scale, heterogeneous wind power dataset covering multiple regions, meteorological conditions, and turbine configurations, and use it to re-pretrain Moirai. By training on this diverse collection of samples, we enhance the model's ability to capture both commonalities and differences across wind farms, thereby building a large wind power forecasting foundation model with strong cross-scenario generalization for direct zero-shot forecasting.

2. Wind power forecasting method based on the Moirai LSTM

2.1. Wind power forecasting architecture

The schematic diagram of the wind power forecasting approach discussed in this paper is shown in Fig. 2. The technical route for wind power forecasting is defined as extrapolating future wind power data based on historical wind power data. Specifically, the model takes the historical wind power data over the past T time steps, denoted as $\mathbf{X}$, as input, and predicts the wind power data for the next P time steps, denoted as $\mathbf{Y}$. The expressions for $\mathbf{X}$ and $\mathbf{Y}$ are given in Eqs. (1) and (2), respectively.

$$\mathbf{X} = [x_1 \ x_2 \ \cdots \ x_T] \in \mathbb{R}^T \quad (1)$$

$$\mathbf{Y} = [y_1 \ y_2 \ \cdots \ y_P] \in \mathbb{R}^P \quad (2)$$

where T represents the input sequence length, and P denotes the forecasting horizon, with $x_1, x_2, \dots, x_T$ representing the historical wind power data at time steps 1, 2, ..., T , and $y_1, y_2, \dots, y_P$ representing the future wind power data to be predicted at the subsequent time steps 1, 2, ..., P .

2.2. Architecture of the Moirai LSTM

Moirai, proposed by the Salesforce research team, is the first unified large model for time series forecasting. It is pretrained on the large-scale open time series dataset LOSTA and is available in three variants: Moirai-small, Moirai-base, and Moirai-large, containing approximately 140 million, 910 million, and 3.11 billion parameters, respectively [22]. In this paper, we base our subsequent experiments on Moirai-small. The core architecture of Moirai, shown in Fig. 3, is built upon an enhanced Transformer framework. Through multi patch size projection layers, any-variate attention mechanism, and a masked encoder design, Moirai addresses challenges faced by traditional time series models, such as variations in temporal frequency and the number of variables.

2.2.1. Multi patch size projection layers

To address the challenges posed by varying temporal frequencies in time series data, the model proposes a dynamic patching strategy. The input sequence is divided into patches with sizes selected according to the sampling frequency (seconds, minutes, hours, days, etc.): high-frequency data (such as second-level wind speed) is compressed using large patches of 128 points to reduce redundancy, while low-frequency data (such as annual temperature) is processed with small patches of 8 points to preserve detailed features. Each patch is mapped into a unified hidden space through an independent linear projection layer, as shown in Eq. (3):

$$\mathbf{X}_{\text{patch}}^{(p)} = \mathbf{W}^p \cdot \mathbf{X}_{\text{raw}}^{(p)} \quad (3)$$

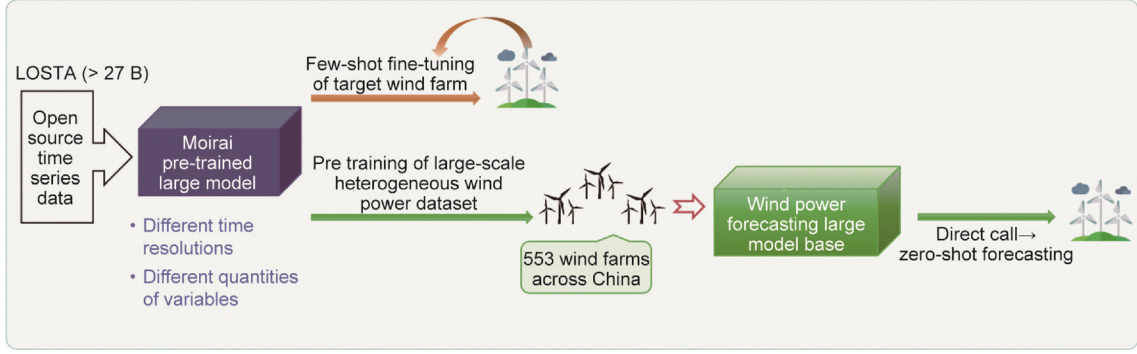


Fig. 1. Overall research roadmap. B: billion.

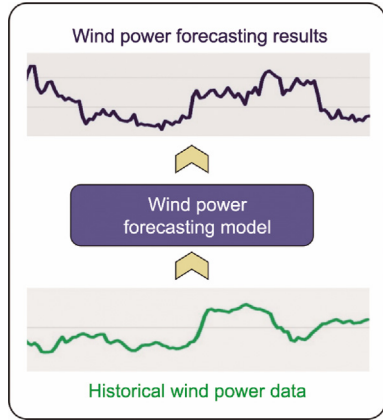


Fig. 2. Schematic diagram of wind power forecasting.

where $\mathbf{X}_{\text{raw}}^{(p)} \in \mathbb{R}^{d_p}$ denotes the original time-series patch vector; $\mathbf{W}^p \in \mathbb{R}^{d_h \times d_p}$ represents the linear projection matrix corresponding to the patch size; and $\mathbf{X}_{\text{patch}}^{(p)} \in \mathbb{R}^{d_h}$ is the resulting hidden feature vector after projection. Here, d_p is the dimension of the input patch and d_h is the dimension of the hidden representation.

2.2.2. Any-variate attention mechanism

In response to the limitations of traditional Transformer models in handling varying numbers of variables for time series forecasting, an any-variate attention mechanism is introduced. By flatten-

ing the multivariate time series into a single-sequence structure, a unified modeling framework that supports any number of variables is constructed. To enable the model to maintain clear variable awareness within the flattened sequence, a dual-encoding scheme is introduced into the attention mechanism.

Time encoding. The time dependencies within the variables are captured through rotational position encoding. For the query vector $\mathbf{q}_{i,m}$ at time step i and variate m and the key vector $\mathbf{k}_{j,n}$ at time step j and variate n , the relative position interaction calculation formula is shown in Eq. (4).

$$R_{i-j}(\mathbf{q}_{i,m}, \mathbf{k}_{j,n}) = \mathbf{q}_{i,m}^T \begin{bmatrix} \cos((i-j)\theta_{d'}) & -\sin((i-j)\theta_{d'}) \\ \sin((i-j)\theta_{d'}) & \cos((i-j)\theta_{d'}) \end{bmatrix} \mathbf{k}_{j,n} \quad (4)$$

where m and n are variate indices, denoting the m th and n th variate in the multivariate time series, respectively; d' is the dimension index ($d' = 1, 2, \dots, d_h$) of the hidden state vector, and $\theta_{d'}$ controls the frequency decay rate of each dimension. The time encoding enables the model to accurately model the time lag effects within the variables.

Variable encoding. Cross-variable interaction control is achieved through learnable binary bias parameters, and the calculation of its attention score is shown in Eq. (5).

$$E_{(i,m)(j,n)} = R_{i-j}(\mathbf{q}_{i,m}, \mathbf{k}_{j,n}) + u^{(1)}\delta_{mn} + u^{(2)}(1 - \delta_{mn}) \quad (5)$$

where δ_{mn} is the Kronecker delta function, which takes the value 1 when the variable index $m = n$, and 0 otherwise; $u^{(1)}$ strengthens the

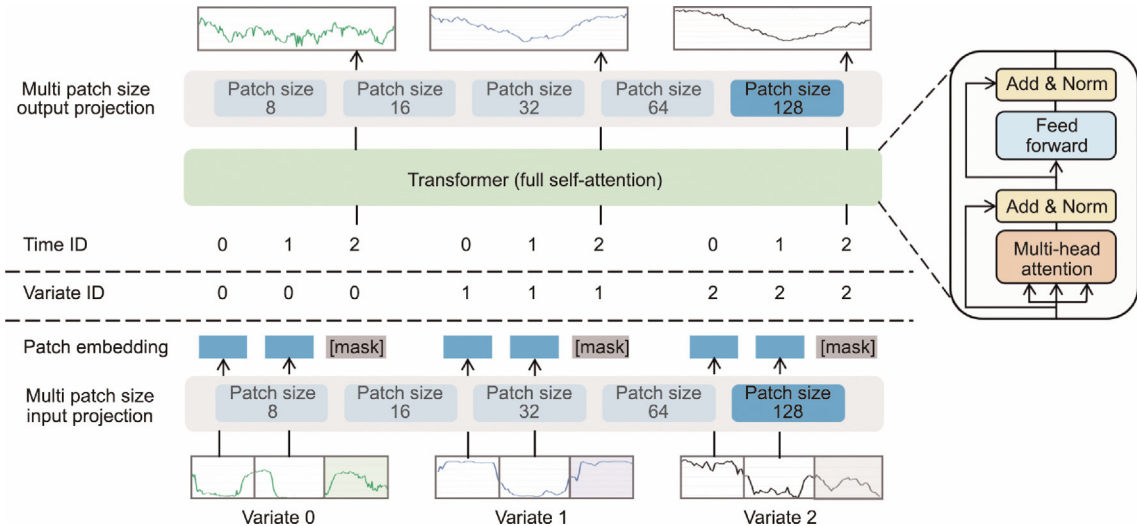


Fig. 3. Moirai structure diagram. ID: identity document.

temporal dependencies within the same variable; $u^{(2)}$ adjusts the strength of the associations between different variables.

2.2.3. Masking encoder

The core module of Moirai is the encoder part of the Transformer model [25], and it has been improved based on large pre-trained models as follows.

Pre-normalization mechanism [26]. RMSNorm normalization [27] is applied before the multi-head attention mechanism and the feed-forward network. Its calculation formula is shown in Eq. (6).

$$\hat{\mathbf{x}} = (\mathbf{x} / \sqrt{\frac{1}{d_h} \sum_{i=1}^{d_h} \mathbf{x}_i^2 + \varepsilon}) \odot \mathbf{g} \quad (6)$$

where $\mathbf{x}$ is the input vector, $\mathbf{g}$ is the learnable scaling parameter vector; ε is used to prevent numerical overflow.

Query-key normalization [28]. L2 constraints are applied to the query matrix $\mathbf{Q} \in \mathbb{R}^{T \times d_h}$ and the key matrix $\mathbf{K} \in \mathbb{R}^{T \times d_h}$ as shown in Eqs. (7) and (8).

$$\tilde{\mathbf{Q}} = \frac{\mathbf{Q}}{\max(\|\mathbf{Q}\|_2, 1)} \quad (7)$$

$$\tilde{\mathbf{K}} = \frac{\mathbf{K}}{\max(\|\mathbf{K}\|_2, 1)} \quad (8)$$

where $\tilde{\mathbf{Q}}$ and $\tilde{\mathbf{K}}$ are the normalized query and key matrices, respectively.

SwiGLU feedforward network [29]. The SwiGLU feedforward network is used to replace the nonlinear layer in the original fully connected layer, enhancing the nonlinear relationships and adjusting the hidden layer dimensions to maintain the same parameter count as the original fully connected layer. The calculation formula for the fully connected layer is shown in Eq. (9).

$$\text{FFN}(\mathbf{x}) = (\text{Swish}(\mathbf{W}_1 \mathbf{x}) \odot \mathbf{W}_2 \mathbf{x}) \mathbf{W}_3 \quad (9)$$

where FFN denotes the feedforward network module, $\mathbf{W}_1$ – $\mathbf{W}_3$ are the first, second, and output weight matrices, respectively.

2.3. Few-shot wind power forecasting based on Moirai fine-tuning

The Moirai large time series forecasting model has been pre-trained on the large-scale time series dataset LOSTA. This dataset integrates time-series data from nine major fields, including global energy, transportation, climate, and more, with a total scale of over 270 billion observations [22]. It not only covers time series features with different frequencies, periodicities, and seasonalities, but also exhibits variability in noise levels and the presence of outliers, providing a powerful training condition for the model [24]. Some datasets display orderly and predictable patterns, while others show significant noise spikes or unexpected events, offering a wide range of scenarios for pattern assimilation.

For wind farms with limited historical data, a few-shot wind power forecasting method based on Moirai fine-tuning is proposed. To fully leverage the rich general time series representations acquired by Moirai during its base training on the large-scale open time series dataset LOSTA, we initialize all layer parameters with those learned weights before performing task specific adaptation, thereby accelerating convergence and enhancing the model's generalization ability. During the fine-tuning process, the limited sample data from the target wind farm is used to train all layer parameters of the Moirai model to adjust the weights of each layer.

It is important to note that to prevent the model from drifting away from the general knowledge acquired during its base

training, a lower learning rate is employed during task specific adaptation, thereby focusing more on capturing the unique fluctuation patterns of the target wind farm. The lower learning rate allows for smoother parameter updates during fine-tuning, with only minor adjustments to the pre-training parameters to avoid destroying the generalized features learned by the model on the large-scale data LOSTA; it also slows down the rate of overfitting so that the model maintains a stronger generalization ability on test data. To improve training stability and accuracy, the same optimizer, AdamW, as used in the Moirai pre-training process is selected, and mean squared error (MSE) is used as the loss function. In the few-shot fine-tuning scenario, due to the limited training data, the model is prone to overfitting on the training set and the performance of the validation set starts to degrade. Therefore, we set a maximum number of training epochs and introduce an early stopping method. The validation set is used to monitor the model training process. Specifically, if the accuracy of the validation set decreases consecutively for E times, the fine-tuning training process will be terminated early, which prevents the model from continuing to remember noise and chance, thus maintaining its generalization ability on test data.

2.4. Zero-shot wind power forecasting based on Moirai re-pretraining

For wind farms with no historical data, a zero-shot wind power forecasting method based on a large-scale heterogeneous wind power dataset and Moirai re-pretraining is proposed. The core idea of this method is to use a large-scale dataset that covers multiple regions, meteorological conditions, and turbine configuration differences, and to re-pretrain the Moirai large model, which has already been pre-trained on the large-scale open time series dataset LOSTA. This builds a large wind power forecasting foundation model that has cross-wind farm generalization ability, allowing it to perform zero-shot forecasting on unseen wind farm tasks.

2.4.1. Large-scale heterogeneous wind power dataset

The constructed large-scale heterogeneous wind power dataset covers 553 wind farms from various regions in China, including Inner Mongolia, Gansu, Henan, Jilin, Jiangxi, Zhejiang, Fujian, among others. In addition to time series measurements, we collected static site attributes (latitude, longitude, rated capacity, and turbine make and model) to further validate the broad spatial coverage and representativeness of the assembled dataset. These wind farms are widely distributed in different municipalities of various provinces, covering a wide range of typical terrains, including plateau, grassland, plains, hills, mountains, and coasts, fully reflecting the diversity of geographic and topographic conditions. We selected wind farms based on the dual considerations of data quality and model diversity: wind farms with more than 10% missing data were excluded, the installed capacity of the selected wind farms was within the range of 750 kW to 3 MW, and there was a reasonable distribution of several mainstream wind turbine models, with each province accounting for less than 20% of the total data volume, so as to avoid over-concentration of a single model or region. The temporal resolution of the data for each wind farm is 15 min or 1 h, and the collection period is one full year to completely capture seasonal fluctuations and extreme weather events.

In order to ensure the data quality and statistical consistency, in the pre-processing stage, all the power series are firstly standardized according to the method of normalization of rated capacity, so as to eliminate the order of magnitude difference of different wind farm sizes, and the calculation formula is shown in Eq. (8). Subsequently, outliers in the normalized

sequence are removed based on the 3σ principle. For missing values, linear interpolation between two neighboring points was used if the consecutive missing time did not exceed 24 h; if the missing time exceeded 24 h, the values were filled in based on the historical average values of the same month at the same station in order to maintain the integrity of the seasonal characteristics. Finally, consistency checks are performed to ensure that the percentage of synthetic data in each sequence does not exceed 5%, and overrun sequences are flagged for backup manual review.

$$v_{\text{norm}}^e = v^e / v_c \quad (10)$$

where v_{norm}^e refers to the e th power value after normalization based on rated capacity; v^e refers to the power value at the e th point, and v_c refers to the rated capacity of the wind farm.

2.4.2. Re-pretraining strategy based on large-scale heterogeneous wind power dataset

Firstly, the network parameters of each layer of the Moirai large model, pre-trained on the large-scale open time series dataset LOSTA, are used as initialization weights before re-pretraining. Then, during the re-pretraining process, the Moirai model's entire layer parameters are trained again based on the samples from the large-scale heterogeneous wind power dataset, constructing the large wind power forecasting foundation model. At this point, the large wind power forecasting foundation model has fully learned and refined the universal features and differentiated patterns presented by different geographical environments, meteorological conditions, and wind turbine configurations in the heterogeneous dataset, resulting in strong generalization ability and robustness.

To enhance the cross-scenario generalization ability of the large wind power forecasting foundation model across different wind farms, during re-pretraining, we follow the cross-task mixed training mechanism [22] used in the original pre-training process of the Moirai large model. Specifically, at each training step, a wind farm scenario and a time-series segment are randomly sampled from the heterogeneous wind power dataset, and the prediction loss is computed for backpropagation. This mechanism avoids the bias caused by traditional sequential training by scenario or task, and effectively promotes the model's ability to extract universal time-series patterns across different wind farms.

In terms of sample construction, a dynamic sliding window strategy [22] is adopted to simulate different wind power forecasting demands. In each training epoch, a time window length $L \in [8, 960]$ is randomly selected and randomly divided into an input sequence T and a prediction length P where $P = \lambda L, \lambda \in [0.1, 0.5]$. For example, when $L = 192$, the prediction length P may range from 17 to 64 points. This design enables the model to simultaneously learn wind power forecasting tasks at different forecasting scales, thereby enhancing its adaptability to various downstream forecasting tasks.

Additionally, in terms of variable dimensions, the Moirai model samples the number of input variables r in each training epoch based on a Beta-Binomial distribution, where $r \sim \text{BetaBinomial}(\text{num} = 64, \alpha = 2, \beta = 5)$, where num is the maximum number of variates, and α and β are the parameters of the Beta distribution that control the shape of the sampling probability. If the actual number of variables in a sample is less than r , zero-padding and masking are applied; otherwise, redundant dimensions are randomly dropped. Similarly, during the re-pretraining process, we use the AdamW optimizer and adopt MSE as the loss function. To prevent model overfitting, a maximum number of training epochs is set, and the early stopping method is introduced.

3. Case study

3.1. Data description

The proposed method is validated using test data from 12 wind farms located across various regions of China. These 12 wind farms are distributed across Shaanxi Province, Gansu Province, Jiangxi Province, Chongqing Municipality, Hunan Province, Jiangsu Province, Guangxi Province, Zhejiang Province, Fujian Province, and the Xinjiang Uygur Autonomous Region. The wind farms are geographically distant from each other and feature diverse terrains, making them well-suited for verifying the model's generalization ability and robustness across diverse wind farm scenarios. Each wind farm dataset contains actual power data from January 1, 2021, to December 31, 2022, with a time resolution of 15 min. To facilitate the calculation of subsequent prediction accuracy metrics, the actual power data is normalized by the rated capacity, thereby masking the original characteristics of each wind farm and enabling direct comparison of forecasting results across different farms.

3.2. Evaluation metrics

To assess the reliability of the proposed method, in accordance with the technical regulations for wind farm integration into power systems, we select accuracy (Acc), pass rate (Pasr), where pass rate is the proportion of forecasts whose absolute error falls within a predefined threshold, and root mean square error (RMSE) as evaluation metrics for wind power forecasting performance. The calculation methods are shown in Eqs. (11)–(13).

$$\text{Acc} = \frac{\sum_{b=1}^B [1 - \sqrt{\frac{\sum_{a=1}^A (\hat{v}_a - v_a)^2}{A}}]}{B} \times 100\% \quad (11)$$

$$\text{Pasr} = \frac{C}{A \times B} \times 100\% \quad (12)$$

$$\text{RMSE} = \sqrt{\frac{1}{A \times B} \sum_{a=1}^{A \times B} (\hat{v}_a - v_a)^2} \quad (13)$$

where $\hat{v}_a$ denotes the forecasting value, v_a denotes the true value; A is the number of points output by the model each time; B is the number of prediction instances; and C is the number of output points where the absolute difference between the predicted value and the true value is less than 0.25, that is, $|\hat{v}_a - v_a| < 0.25$. It should be noted that, according to the engineering experience of actual wind power forecasting, the forecasting error within $\pm 25\%$ of the installed capacity is usually considered as acceptable, so we set the threshold of Pasr at 0.25 to ensure that the evaluation indexes are comparable in the industry and reflect the actual forecasting performance of the model.

3.3. Performance analysis of few-shot wind power forecasting based on Moirai fine-tuning

Based on our experimental experience, when the number of training samples is less than one month, the accuracy of ultra-short-term wind power forecasting is generally low, and it is difficult to meet the forecasting demand in actual operation. Therefore, to evaluate the performance of Moirai on few-shot datasets with different numbers of training samples, the original dataset is divided into five scenarios as shown in Table 1, with the number of training samples gradually increasing from 3 to 30 d. Ultra-short-term wind power forecasting is conducted, with the input

historical sequence length T set to 48 points (12 h) and the forecasting sequence length P set to 16 points (4 h).

The entire layer parameters of the Moirai LSTM are fine-tuned using the few-shot training samples from the five cases mentioned above, and testing is conducted on the 2022 data. To fully evaluate the effectiveness of the proposed method, the following comparison methods are set.

LSTM. Traditional supervised learning method, directly training the LSTM deep neural network using the limited training samples and tests it on the 2022 data.

Transformer. Traditional supervised learning method, directly training the Transformer deep neural network using the limited training samples and tests it on the 2022 data.

It is worth noting that in the case of limited sample size, we avoid setting too many deep network layers that may lead to overfitting. Therefore, the LSTM model we used consisted of one layer of recurrent units: First, historical power data was raised to the dimension of the hyperparameter vector (d_{model}) through linear mapping, then input into a single-layer LSTM, and finally returned to one-dimensional power forecasting through linear projection. The Transformer model used consists of one encoder layer and one decoder layer: each layer is composed of multi-head self-attention and a feedforward network. The historical power sequence is linearly mapped to d_{model} dimensions and position encoding is added before being fed into the encoder and decoder. The decoder output is mapped back to a single-dimensional power value through a linear layer and Tanh activation. Both models were evaluated using a grid search method for hyperparameter optimization. The hyperparameter settings are shown in Appendix Table S1.

To make the forecasting results more intuitive and avoid randomness, the average forecasting results of each method are obtained by independently running each method 20 times on 12 wind farms, as shown in Table 2.

From Table 2, it can be seen that as the training sample size increases from 3 to 30 d, the forecasting performance of the proposed Moirai fine-tuned few-shot wind power forecasting method generally shows a gradual upward trend, while the traditional supervised learning methods, LSTM and Transformer, exhibit a more significant performance improvement. It is worth noting that the proposed method has a significant advantage with very small sample sizes: when the training sample sizes are 3, 5, and 10 d, the proposed method's forecasting Acc is 6.88%, 4.45%, and 1.95% higher than the LSTM method, respectively, and 8.76%, 4.34%, and 0.63% higher than the Transformer method, respectively. When the sample size exceeds 15 d, the Transformer model, by fully learning the local time-series mapping relationships, significantly improves in forecasting accuracy and surpasses the proposed method. However, the proposed Moirai fine-tuned few-shot wind power forecasting method, constrained by the general features learned from pretraining, shows a more gradual increase in accuracy.

This trend indicates that the Moirai model, through pretraining on the large-scale time series dataset LOSTA, has acquired general time-series representation prior knowledge. When the sample size is below 15 d, it effectively mitigates the overfitting issue caused

by data scarcity through fine-tuning. However, as the sample size increases, its general features are gradually surpassed by the extensive training of traditional supervised learning models, which better capture the target wind farm's forecasting patterns. The experimental results validate the effectiveness of the proposed Moirai fine-tuned few-shot wind power forecasting method in extremely small sample scenarios. Through the synergy of transferring general time-series knowledge from the large-scale dataset and targeted fine-tuning for the specific wind farm, it achieves high-accuracy wind power forecasting under limited sample conditions.

To further demonstrate the generalization performance of the proposed method across different target wind farms, Figs. 4 and 5 present the forecasting accuracy comparison of 12 wind farms under three different methods, using the Acc results from Case 1 and Case 2 as examples. As shown in Figs. 4 and 5, when the amount of training data is limited to 3 and 5 d, traditional supervised learning methods based on LSTM and Transformer struggle to perform well on each wind farm due to their heavy reliance on large amounts of historical data. Their forecasting performance is poor and highly unstable across different wind farms. Although the proposed method exhibits relatively lower performance on Wf1, Wf3, and Wf11 in Case 2, it achieves significantly better results on most wind farms compared to the other methods and maintains more consistent performance across different wind farms. These findings further demonstrate the strong generalization capability of the proposed few-shot wind power forecasting method based on Moirai fine-tuning in modeling multi-source temporal features, offering a universal solution for power forecasting in newly built or data-scarce wind farms.

3.4. Performance analysis of zero-shot wind power forecasting based on Moirai re-pretraining

To evaluate the zero-shot forecasting capability of the large wind power forecasting foundation model built by re-pretraining Moirai on a large-scale heterogeneous wind power dataset, we use the 2022 data from 12 wind farms as the test set and directly perform ultra-short-term and short-term wind power forecasting on the pretrained model foundation. Specifically, for ultra-short-term wind power forecasting, the input historical sequence length T is 48 points (12 h), and the forecasting sequence length P is 16 points (4 h); for short-term wind power forecasting, the input historical sequence length T is 480 points (5 d), and the forecasting sequence length P is 16 points (1 d).

To validate the effectiveness of the proposed zero-shot wind power forecasting method based on the large wind power forecasting foundation model, we set the baseline method as follows: directly using the Moirai LSTM pretrained only on the large-scale open time series dataset LOSTA to perform zero-shot wind power forecasting, with the 2022 data from the 12 wind farms serving as the test set. The ultra-short-term and short-term wind power forecasting results of the two methods are presented in Table 3.

As shown in Table 3, the proposed zero-shot wind power forecasting method based on Moirai re-pretraining significantly outperforms the baseline model pretrained only on the large-scale open time series dataset LOSTA across both ultra-short-term and short-term forecasting horizons. For ultra-short-term zero-shot forecasting, the proposed method achieves an average Acc of 86.57%, Pasr of 91.66%, and an RMSE of 0.14. Compared to the baseline, this represents improvements of 1.47% and 1.66% in Acc and Pasr, respectively, and a reduction of 0.01 in RMSE. Notably, the forecasting performance even surpasses that of traditional Transformer-based supervised learning methods trained with 15 d of historical data. For short-term zero-shot forecasting, the proposed method achieves an average Acc of 77.59%, a Pasr of

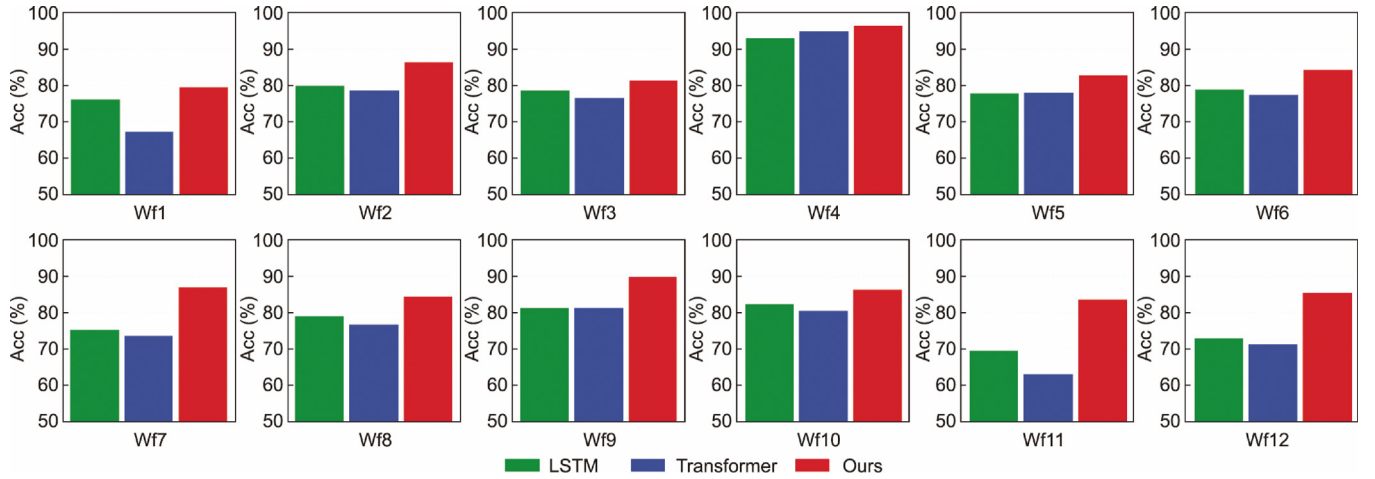
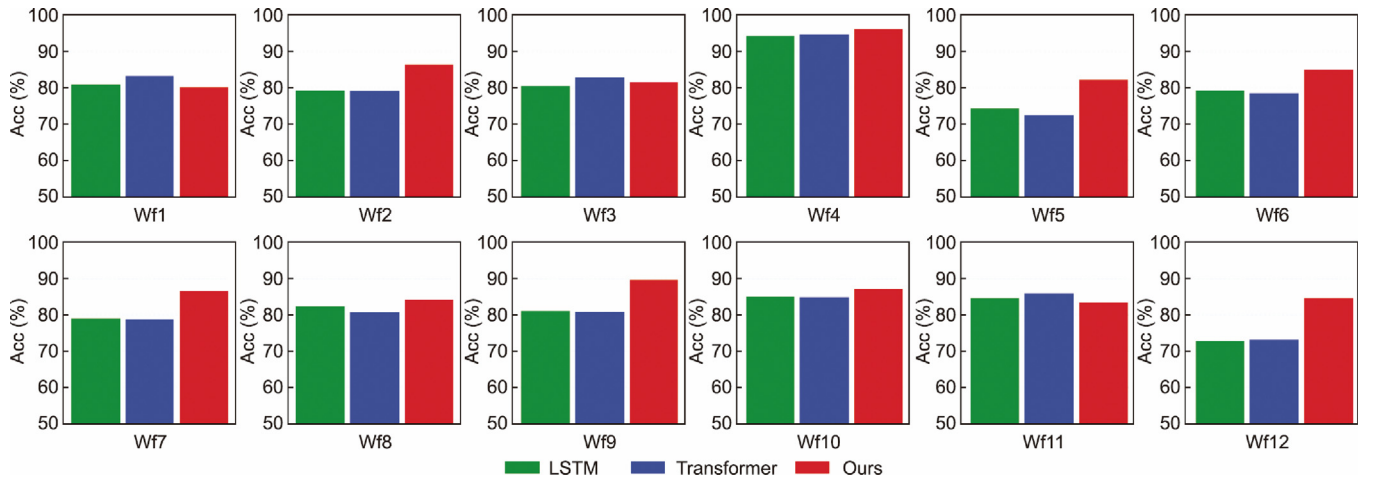
Table 1
Five case settings.

Case	Training dataset	Testing dataset
1	2021.12.29–2021.12.31 (3 d)	2022.1.1–2022.12.31
2	2021.12.27–2021.12.31 (5 d)	2022.1.1–2022.12.31
3	2021.12.22–2021.12.31 (10 d)	2022.1.1–2022.12.31
4	2021.12.17–2021.12.31 (15 d)	2022.1.1–2022.12.31
5	2021.12.2–2021.12.31 (30 d)	2022.1.1–2022.12.31

Table 2

Average forecasting results on 12 wind farms of each method.

Case	Acc (%)			Pasr (%)			RMSE		
	Ours	LSTM	Transformer	Ours	LSTM	Transformer	Ours	LSTM	Transformer
1	85.50	78.62	76.47	90.52	70.69	66.98	0.15	0.25	0.27
2	85.46	81.01	81.12	90.30	75.80	75.77	0.15	0.23	0.23
3	86.21	84.26	85.85	91.34	80.78	83.81	0.14	0.19	0.17
4	85.97	84.39	86.12	91.10	81.53	84.35	0.14	0.18	0.17
5	86.49	85.85	89.21	91.71	84.60	90.12	0.14	0.17	0.13

**Fig. 4.** Comparison chart of forecasting accuracy for 12 wind farms under three methods in Case 1. Wf: wind farm.**Fig. 5.** Comparison chart of forecasting accuracy for 12 wind farms under three methods in Case 2.

78.62%, and an RMSE of 0.22, improving upon the baseline by 1.29% and 1.97%, respectively, with the RMSE reduced by 0.02. Except for Wf12 in the short-term forecasting, which performs slightly worse than the baseline model, the proposed method achieves significant improvement in the remaining 11 wind farms, with the Acc of most of the farms improved by more than 1% compared with the baseline method. In addition, the performance improvements under various terrains and turbine types are relatively consistent, and no obvious trend of superiority or inferiority is seen under a certain scenario, indicating that the constructed large wind power forecasting foundation model is able to cross different terrains and turbine types, and achieve a good generalization ability.

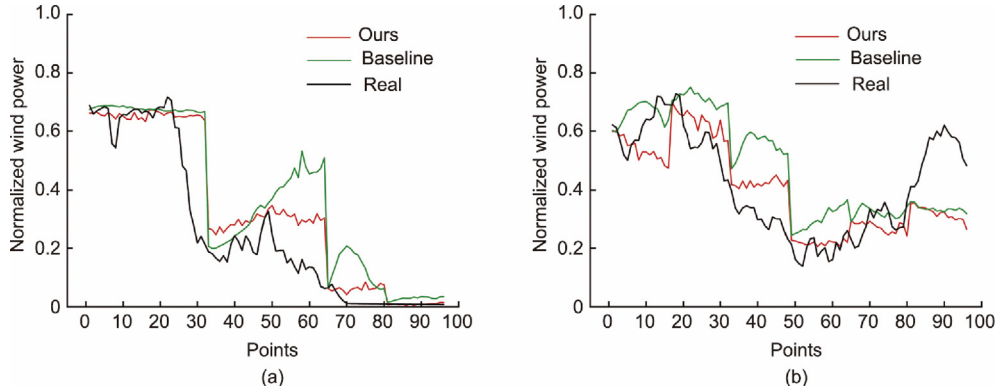
To make the forecasting results more intuitive, we plot both ultra-short-term and short-term wind power forecasting curves

for Wf1 and Wf2, each showing 96 points from a randomly selected day. From the Figs. 6(a)–(b), it is clear that the baseline method's curves show reverse jumps that deviate from the true power downtrend, and although our method occasionally has small reverse jitters, the magnitude is significantly smaller. In Fig. 7(b), the real power decreases rapidly in the first half of the curve, and our method is able to fit this trend faster, while the baseline method curve has a decreasing direction, but the curve decreases more slowly and with a significant lag, leading to a larger deviation. It's all because of the baseline model—pretrained on the large-scale open time series dataset LOSTA—was not optimized for wind farm scenarios and cannot accommodate task-specific features such as sudden wind-speed jumps and turbine start-stop events. Furthermore, the baseline model is highly sensitive to local

Table 3

Zero-shot ultra-short-term and short-term wind power forecasting performance under two methods.

Item	Ultra-short-term			Short-term		
	Acc (%)	Passr (%)	RMSE	Acc (%)	Passr (%)	RMSE
Wf1						
Ours	81.18	84.98	0.19	70.32	61.14	0.30
Baseline	78.79	82.40	0.21	68.96	61.98	0.31
Wf2						
Ours	87.21	92.82	0.13	78.97	83.55	0.21
Baseline	85.65	90.73	0.14	77.35	79.61	0.23
Wf3						
Ours	82.37	86.75	0.18	69.75	67.68	0.30
Baseline	80.71	84.92	0.19	68.04	63.92	0.32
Wf4						
Ours	96.40	99.93	0.04	93.08	99.10	0.07
Baseline	96.09	99.93	0.04	92.77	98.87	0.07
Wf5						
Ours	83.50	89.34	0.17	70.92	71.16	0.29
Baseline	82.24	87.75	0.18	69.45	68.34	0.31
Wf6						
Ours	85.50	90.76	0.15	77.34	79.47	0.23
Baseline	83.68	88.05	0.16	75.71	75.07	0.24
Wf7						
Ours	87.73	93.44	0.12	78.17	79.53	0.22
Baseline	86.63	92.20	0.13	76.91	77.61	0.23
Wf8						
Ours	85.30	90.63	0.15	77.23	79.28	0.23
Baseline	83.93	88.67	0.16	76.73	76.54	0.23
Wf9						
Ours	90.43	96.45	0.10	84.11	88.07	0.16
Baseline	89.42	95.26	0.11	82.02	85.96	0.18
Wf10						
Ours	87.96	93.88	0.12	82.35	86.65	0.18
Baseline	85.87	92.17	0.14	81.58	85.39	0.18
Wf11						
Ours	84.45	89.12	0.16	76.01	76.36	0.24
Baseline	83.21	87.24	0.17	73.14	73.10	0.27
Wf12						
Ours	86.17	91.86	0.14	72.86	71.49	0.27
Baseline	85.02	90.63	0.15	72.93	73.45	0.27
Mean						
Ours	86.57	91.66	0.14	77.59	78.62	0.22
Baseline	85.10	90.00	0.15	76.30	76.65	0.24

**Fig. 6.** Zero-shot ultra-short-term wind power forecasting curves with different methods: (a) Wf1 and (b) Wf2.

noise anomalies, causing the baseline forecasts to exhibit high-frequency, irregular oscillations, with some forecasted values even exceeding the normalized range of 0–1. In contrast, our proposed method closely follows the true power trends in both ultra-short-term and short-term forecasts, with markedly reduced fluctuation amplitudes. This demonstrates that by re-pretraining Moirai on a large scale heterogeneous wind power dataset, the model acquires prior knowledge of the nonlinear temporal dynamics of power output, equipment capacity constraints, and seasonal

cycles, thereby improving its robustness when forecasting for downstream, previously unseen wind farms.

The above results not only demonstrate the overall improvement in zero-shot forecasting accuracy brought by the proposed Moirai re-pretraining approach, but also verify the strong adaptability and robustness of the constructed large wind power forecasting foundation model across downstream wind farms with varying geographical locations, meteorological conditions, and turbine configurations. Thanks to the cross-task mixed training strat-

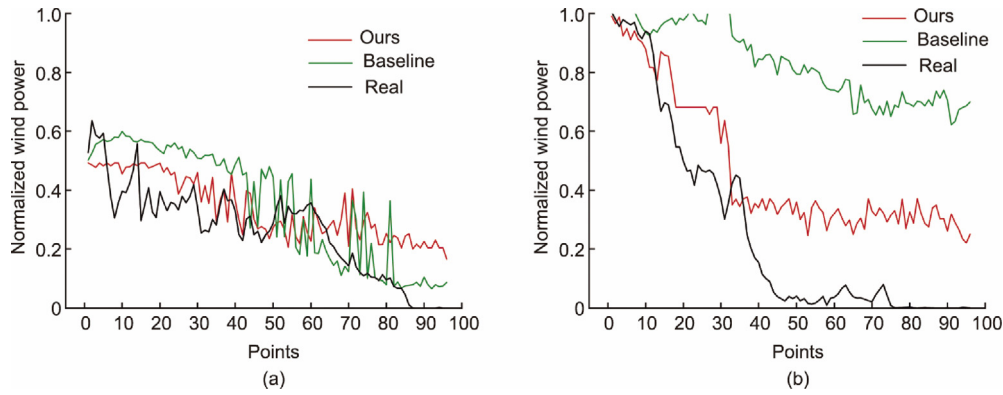


Fig. 7. Zero-shot short-term wind power forecasting curves with different methods: (a) Wf1 and (b) Wf2.

egy and dynamic sliding window mechanism employed during re-pretraining, as well as Moirai's architectural designs such as multi patch size projection layers and any-variate attention, the resulting model foundation can flexibly adapt to different time frequencies, input-output sequence lengths, and numbers of forecasting variables. This endows the model with strong engineering applicability, providing a reliable solution for zero-shot wind power forecasting in practical engineering scenarios.

3.5. Computational resources analysis

All experiments were conducted on two platforms. The few-shot fine-tuning of Moirai and its re-pretraining ran on a CentOS 7.8 high performance computing (HPC) cluster with Intel Xeon Platinum central processing units (CPUs), 320 GB random access memory (RAM) and four NVIDIA Tesla V100 graphics processing units (GPUs). Both the Moirai model testing and baseline training/testing were performed on a Windows 10 workstation equipped with an Intel Core i5 CPU, NVIDIA RTX 4060 GPU and 16 GB RAM.

The computational cost of neural networks can be divided into training and testing phases. Once a network is well trained, its inference cost is minimal and typically negligible. Under optimal parameter settings, the processing times for the Moirai LSTM-based wind power forecasting method during the fine-tuning and re-pretraining stages are shown in Table 4. In the fine-tuning stage, the training times for the 12 wind farms across Cases 1–5 exhibit no significant variation; for clarity and brevity, only the average time is reported.

4. Conclusions

As wind power is integrated at scale and new power system architectures advance, the importance of high-accuracy wind power forecasting for grid dispatch and energy absorption grows ever more pronounced. To address the inability of traditional deep neural network models to handle newly built or expanded wind

farms with limited or no historical power data, we propose a wind power forecasting method based on the Moirai LSTM, exploring both few-shot and zero-shot scenarios. For wind farms with few historical data points, we introduce a few-shot wind power forecasting method via Moirai fine-tuning, using the Moirai model parameters pretrained on LOSTA as initialization and fine-tuning them on the target farm's limited samples. For target farms completely lacking historical observations, we propose a zero-shot wind power forecasting method via Moirai re-pretraining, likewise initializing all layers with the original Moirai parameters and re-training them on our constructed large-scale heterogeneous wind power dataset to build a large wind power forecasting foundation model. Experimental results show that:

Efficient forecasting in the few-shot scenario. By leveraging Moirai's pretraining on the large-scale time series dataset LOSTA and applying targeted fine-tuning on the target wind farm, the proposed method significantly mitigates overfitting under few-shot conditions and achieves high-accuracy few-shot wind power forecasting. When the training sample size falls below 15 d, this Moirai-based few-shot forecasting method markedly outperforms traditional supervised learning approaches built on deep neural network models.

Substantial improvement in the zero-shot scenario. By re-pretraining Moirai on a heterogeneous dataset covering 553 wind farms, the resulting large wind power forecasting foundation model demonstrates exceptional cross-scenario generalization under zero-shot conditions. In the ultra-short-term forecasting task, it achieves an average Acc of 86.57%, a 1.47% improvement over the baseline model; in the short-term forecasting task, it attains an average Acc of 77.59%, representing a 1.29% increase over the baseline.

Universality and scalability of the technical approach. The proposed framework integrates architectural innovations such as multi patch size projection layers and any-variate attention mechanism, along with training strategies including dynamic sliding windows and cross-task mixed training, to flexibly handle varying temporal resolutions, input sequence lengths, forecasting horizons, and numbers of variables. Beyond wind power forecasting, it can also be extended to photovoltaic and load forecasting tasks, providing a versatile foundation for renewable energy time-series prediction.

To further enhance wind power forecasting performance based on the Moirai LSTM and reinforce its practical significance, future research may proceed along two directions: ① Incorporate numerical weather prediction data as dynamic covariates alongside historical wind power data when re-pretraining, fine-tuning and performing final forecasting with the Moirai LSTM. ② In the re-pretraining stage, rather than limiting the data to wind power alone,

Table 4
Computational time for Moirai fine-tuning and re-pretraining stages.

Stage	Case	Computational time
Fine-tune	Case 1 (Section 3.3)	2:48
	Case 2 (Section 3.3)	2:29
	Case 3 (Section 3.3)	2:24
	Case 4 (Section 3.3)	3:24
	Case 5 (Section 3.3)	3:43
Re-pretraining	Case in Section 3.4	128:34:21

incorporate solar and load measurements to construct a large-scale, heterogeneous renewable energy–electric load dataset. Apply the same cross task mixed training and dynamic sliding window strategies to build a general purpose foundation model for renewable energy and load forecasting with adaptive mapping dimensions.

CRediT authorship contribution statement

Jie Yan: Methodology, Funding acquisition, Conceptualization. **Yujia Li:** Writing – original draft, Software. **Han Wang:** Validation, Formal analysis, Data curation. **Shuang Han:** Supervision, Resources, Project administration. **Wenlong Shang:** Writing – review & editing, Validation, Supervision, Methodology. **Yongqian Liu:** Visualization, Supervision, Resources.

Declaration of competing interest

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

Acknowledgments

This work was supported by National Natural Science Foundation of China (52277099), Beijing Nova Program (20230484355), and the Fundamental Research Funds for the Central Universities (2024JC007).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.eng.2025.11.007>.

References

- [1] Zhang J, Hao Y, Fan R, Wang Z. An ultra-short-term PV power forecasting method for changeable weather based on clustering and signal decomposition. *Energies* 2023;16(7):3092.
- [2] Wang Y, Zou R, Liu F, Zhang L, Liu Q. A review of wind speed and wind power forecasting with deep neural networks. *Appl Energy* 2021;304:117766.
- [3] Feng S, Wang W, Wang Z, Song Z, Yang Q, Wang B. Global wind-power generation capacity in the context of climate change. *Engineering* 2025;51:86–97.
- [4] Meng A, Chen S, Ou Z, Xiao J, Zhang J, Chen S, et al. A novel few-shot learning approach for wind power prediction applying secondary evolutionary generative adversarial network. *Energy* 2022;261:125276.
- [5] Wang L, Tao R, Hu H, Zeng Y. Effective wind power prediction using novel deep learning network: stacked independently recurrent autoencoder. *Renew Energy* 2021;164:642–55.
- [6] Leslie M. Promising results predict role for artificial intelligence in weather forecasting. *Engineering* 2024;39:10–2.
- [7] Lim S, Huh J, Hong S, Park C, Kim J. Solar power forecasting using CNN-LSTM hybrid model. *Energies* 2022;15(21):8233.
- [8] Wei J, Wu X, Yang T, Jiao R. Ultra-short-term forecasting of wind power based on multi-task learning and LSTM. *Int J Electr Power Energy Syst* 2023;149:109073.
- [9] Xiao Y, Zou C, Chi H, Fang R. Boosted GRU model for short-term forecasting of wind power with feature-weighted principal component analysis. *Energy* 2023;267:126503.
- [10] Niu Z, Han X, Zhang D, Wu Y, Lan S. Interpretable wind power forecasting combining seasonal-trend representations learning with temporal fusion transformers architecture. *Energy* 2024;306:132482.
- [11] Qu K, Si G, Shan Z, Kong X, Yang X. Short-term forecasting for multiple wind farms based on transformer model. *Energy Rep* 2022;8(S5):483–90.
- [12] 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 Manage* 2023;283:116916.
- [13] Chen F, Yan J, Liu Y, Yan Y, Tjernberg LB. A novel meta-learning approach for few-shot short-term wind power forecasting. *Appl Energy* 2024;362:122838.
- [14] Ren X, Wang Y, Cao Z, Chen F, Li Y, Yan J. Feature transfer and rapid adaptation for few-shot solar power forecasting. *Energies* 2023;16(17):6211.
- [15] Hu J, Hu W, Cao D, Huang Y, Chen J, Li Y, et al. Bayesian averaging-enabled transfer learning method for probabilistic wind power forecasting of newly built wind farms. *Appl Energy* 2024;355:122185.
- [16] Fan H, Fan X, Wei W, Hao T, Chen K, Wang G, et al. Privacy preserving ultra-short-term prediction in clustered wind farms with encrypted data sharing: a secure multi-party computation approach. *Expert Syst Appl* 2025;278:127218.
- [17] Open AI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. 2024. arXiv:2303.088774.
- [18] Chowdhery A, Narang S, Narang J, Bosma M, Mishra G, Roberts A, et al. PaLM: scaling language modeling with pathways. 2022. arXiv:2204.02311.
- [19] Lu H, Liu W, Zhang B, Wang B, Dong K, Liu B, et al. DeepSeek-VL: towards real-world vision-language understanding. 2024. arXiv:2403.05525.
- [20] Garza A, Challu C, Mergenthaler-Canseco M. TimeGPT-1. 2024. arXiv:2010.03589.
- [21] Rasul K, Ashok A, Williams AR, Ghonia H, Bhagwatkar R, Khorasani A, et al. Lag-Llama: towards foundation models for probabilistic time series forecasting. 2024. arXiv:2310.08278.
- [22] Woo G, Liu C, Kumar A, Xiong C, Savarese S, Sahoo D. Unified training of universal time series forecasting transformers. 2024. arXiv:2402.02592.
- [23] Liu Y, Zhang H, Li C, Huang X, Wang J, Long M. Timer: generative pre-trained transformers are large time series models. 2024. arXiv:2402.02368.
- [24] Liao W, Wang S, Yang D, Yang Z, Fang J, Rehtanz C, et al. TimeGPT in load forecasting: a large time series model perspective. *Appl Energy* 2025;379:124973.
- [25] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN. Attention is all you need. 2023. arXiv:1706.03762.
- [26] Xiong R, Yang Y, He D, Zheng K, Liu T. On layer normalization in the transformer architecture. In: Proceedings of International Conference on Machine Learning; 2020 Jul 12–18; Vienna, Austria. PMLR; 2020.
- [27] Zhang B, Sennrich R. Root mean square layer normalization. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R. CoRR 2019; Advances in neural information processing systems; NeurIPS; 2019.
- [28] Henry A, Dachapally PR, Pawar SS, Chen Y. Query-key normalization for transformers. In: Proceedings of findings of the association for computational linguistics; 2020 Nov 16–20; online. Association for Computational Linguistics; 2022.
- [29] Shazeer N. GLU variants improve transformer. 2020. arXiv:2002.05202.