Energy Engineering

Tech Science Press

Doi:10.32604/ee.2025.065124

ARTICLE

Few-Short Photovoltaic Systems Predictions Algorithm in Cold-Wave Weather via WOA-CNN-LSTM Model

Ruiheng Pan*, Shuyan Wang, Yihan Huang and Gang Ma

School of Electrical and Automation Engineering, Nanjing Normal University, Nanjing, 210023, China

*Corresponding Author: Ruiheng Pan. Email: 21220231@njnu.edu.cn

Received: 04 March 2025; Accepted: 29 May 2025; Published: 24 July 2025

ABSTRACT: Contemporary power network planning faces critical challenges from intensifying climate variability, including greenhouse effect amplification, extreme precipitation anomalies, and persistent thermal extremes. These meteorological disruptions compromise the reliability of renewable energy generation forecasts, particularly in photovoltaic (PV) systems. However, current predictive methodologies exhibit notable deficiencies in extreme weather monitoring, systematic transient phenomena analysis, and preemptive operational strategies, especially for cold-wave weather. In order to address these limitations, we propose a dual-phase data enhancement protocol that takes advantage of Time-series Generative Adversarial Networks (TimeGAN) for temporal pattern expansion and the K-medoids clustering algorithm for synthetic data quality verification. In order to better extract the spatiotemporal features of the model input simultaneously, we develop a hybrid neural architecture integrating Convolutional Neural Networks with Long Short-Term Memory modules (CNN-LSTM). To avoid the problem of hyperparameters getting trapped in local optimal solutions, we use the Whale Optimization Algorithm (WOA) algorithm to obtain the global optimal solution by simulating the hunting of humpback whales, further enhancing the generalization ability of the model. Experimental validation demonstrates performance improvements, with the proposed model achieving 30% higher prediction accuracy compared to Genetic Algorithm-Backpropagation Neural Network (GA-BPNN) and Radial Basis Function-Support Vector Regression (RBF-SVR) benchmarks, promoting the renewable energy prediction in data-constrained extreme weather scenarios for future power networks.

KEYWORDS: Cold-wave weather; few-short; WOA-CNN-LSTM; photovoltaic systems

1 Introduction

Recent years have witnessed the emergence of sustainable energy technologies, which are playing an vital role in shaping the future power networks by reducing carbon emissions and supporting the transition toward a environmentally friendly energy infrastructure [1]. At present, photovoltaic power generation technology that uses solar energy for energy conversion is gradually maturing and improving. In the new power system with new energy as the main body, new energy power generation has gradually become the largest power source, has the ability to actively support, and has become the main source of power and electricity supply [2]. Due to the randomness and volatility of wind power and photovoltaic power, their large-scale integration into the power grid has had an impact on the safe and stable operation of the power system. Predicting the future output power of wind power and photovoltaic power in advance and reserving consumption space based on the prediction results are important technical means for China to improve the consumption level of new energy and ensure the safety of the power system [3].

The problems of strong volatility and high randomness in new energy power generation have increased the difficulty of power prediction for new energy power generation and raised the accuracy requirements for power prediction. This means that the requirements for power prediction systems are also increasing day by day [4]. At present, the basic data related to new energy power generation is insufficient, and the characteristics and patterns of changes have not been clearly understood. The relevant meteorological data is incomplete, making it difficult to grasp the characteristics of weather changes and the spatio-temporal regularity. In addition, new energy power stations are numerous and widespread. Coupled with the frequent occurrence of major weather conditions such as sandstorms and cold waves in recent years, it has made the medium and short-term prediction of new energy difficult, seriously affecting the accuracy of new energy power prediction. Therefore, studying a new energy power prediction model that is universal in both major weather and normal weather conditions is of great significance for maintaining the stable operation of the power grid and improving the consumption of new energy.

Firstly, the prediction of photovoltaic power output is beneficial to the stability and reliability of the power system. Under major weather conditions, the output of photovoltaic systems is affected by many uncertain factors, such as strong winds and strong light radiation. The prediction of photovoltaic power output can help the energy system better plan and adjust energy output, thereby enhancing the stability and reliability of the system. This is crucial for ensuring the normal operation of the power system under extreme conditions. Secondly, the prediction of photovoltaic power output can make operation and maintenance of power systems more intelligent. Photovoltaic power output prediction can provide valuable information for operation and maintenance personnel, helping them understand the possible challenges the system may face in advance. By predicting the performance of the system under different major weather conditions, the operation and maintenance team can take corresponding measures, such as timely maintenance and allocation of backup equipment, to ensure the long-term operation of the photovoltaic system. At the same time, it can also achieve the maximization of economic benefits: under major weather conditions, the power generation capacity of photovoltaic (PV) systems may fluctuate significantly, and normal operation may be affected to a certain extent. Through the prediction of photovoltaic power, the energy system can adjust the power production plan more flexibly to maximize the utilization of available solar energy resources, improve the efficiency of power generation, and reduce the cost of power production. Then it is more conducive to addressing the challenges of climate change. Major weather events caused by climate change are becoming more frequent and severe, which puts forward higher requirements for photovoltaic systems. Through the precise prediction of photovoltaic power, energy systems can better adapt to and respond to the challenges brought about by climate change, ensuring the sustainability of power supply [5]. Finally, the prediction of PV is convenient for the integration and optimization of renewable energy resources. PV prediction can not only be applied independently to solar energy systems, but also be integrated with other renewable energy systems, such as wind energy and energy storage, etc. [6]. By rationally coordinating the output of various renewable energy sources, a more efficient and sustainable energy supply system can be achieved. In addition, the prediction of PV has also received governmental support and social recognition. Accurate photovoltaic power prediction is conducive to formulating more scientific and reasonable energy policies [7]. At the same time, for the general public, the reliability of photovoltaic systems under major weather conditions is also an important factor in assessing the sustainability of renewable energy, which can enhance people's trust and acceptance of renewable energy.

Overall, considering major weather conditions, conducting photovoltaic power output prediction not only has a profound impact on enhancing the stability and reliability of the energy system, but also provides strong support for sustainable development, climate change response, and other aspects. Through scientific prediction and reasonable planning, the stable operation of photovoltaic systems under various major

weather conditions can be better achieved, promoting a larger proportion of renewable energy in the global energy structure. PV output prediction is a research field that attracts extensive attention internationally. Many academic institutions, research teams, and the industrial sector are actively investing resources in related research. The following are some of the main aspects of current research at home and abroad:

Numerous studies have been conducted on AI-driven PV generation prediction, enhancing quality of meteorological data [8] and the adaptability of predictive models [9]. These solutions support the larger objectives of sustainability and energy efficiency in contemporary power systems by aiming to increase forecasting accuracy and guarantee financial gains for both energy suppliers and consumers [10].

There has been relevant research on the monitoring and characteristic analysis of major weather events internationally, but most studies focus on the overall considerations within conventional weather forecasting systems, with limited attention paid to specialized forecasting techniques for major weather events [11,12]. Literature [13] conducted feature selection and modeling analysis based on extreme weather data and proposed a machine learning-based prediction model framework. However, research on the spatiotemporal characteristics of major weather events, particularly the coupling characteristics of different major weather events with photovoltaic and wind power, is still lacking. In terms of the impact analysis of new energy output limitations under major weather conditions, the mechanisms limiting wind power output in low-temperature environments have been studied in depth. Literature [14] used meteorological factors, applied fuzzy recognition technology, and used the results as inputs to a neural network to achieve effective meteorological condition forecasting. Literature [15] proposed an application framework for power meteorological sensor monitoring technology and discussed in detail the development and application status of various sensing technologies. However, there has yet to be a model for analyzing new energy limitations under different major weather events, especially in terms of modeling the impact of future meteorological environmental uncertainties, where research is still insufficient.

In the study of new energy output prediction under major weather conditions, Literature [16] explored the relationship between meteorological factors and new energy output using distance analysis, and through SOM clustering, classified the samples to construct a photovoltaic output prediction model based on weather type clustering. Literature [17] mapped the proportional relationship between output means under different weather conditions to a weather type index, which was then used as input to a neural network, thereby accurately establishing an association model between meteorological conditions and new energy generation output. Literature [18] proposed a method for extracting power load characteristics for extreme weather based on historical load and meteorological data. Literature [19] proposed a correction method for wind power prediction under extreme weather conditions. However, the model feature extraction used in the above-mentioned methods is relatively simple, which may ignore the complex coupling relationship among meteorological features. Meanwhile, the model design is relatively traditional and does not make full use of the deep learning ability, resulting in relatively low prediction accuracy.

Additionally, the occurrence probability of cold-wave weather is relatively low, so the available sample data is limited. Therefore, expanding small sample data using models is considered, with many domestic precedents. Literature [20] employed a dual-generator and dual-discriminator structure of a recurrent generative adversarial network (GAN) to effectively expand the number of fault-labeled samples, solving the issue of fault recognition failure due to scarce labeled samples. Literature [21] used convolutional neural networks to construct generative and discriminative models separately and trained them through adversarial training of conditional GANs, generating prediction data based on key influencing factors. However, in photovoltaic power generation, the temporal variation characteristics are also crucial. Therefore, this paper focuses on generating sample models based on the temporal variation characteristics of cold-wave weather types.

Compared with published research articles in few-short PV prediction algorithms, the contributions and innovations of the proposed method are summarized as follows:

(1) A novel few-short data expansion method for cold-wave weather based on the TimeGAN temporal network is proposed in this paper to address errors and missing values within the dataset, thereby resolving the issue of insufficient few-short data in cold-wave scenarios.

(2) This paper proposes a neural network prediction model based on the WOA optimization algorithm, which optimizes hyperparameters on the basis of common models, which can effectively solve the problem that common models are prone to fall into local optimal solutions and improve the prediction accuracy.

(3) The effectiveness of the proposed framework is validated through a detailed case study in a cold-wave weather condition, achieving 30% higher prediction accuracy compared to Genetic Algorithm-Backpropagation Neural Network (GA-BPNN) and Radial Basis Function-Support Vector Regression (RBF-SVR) benchmarks.

The rest of this paper is organized as follows: The algorithm of this paper is introduced in detail in Section 2. Data testing and comparative analysis of this model are presented in Section 3. In the end, the paper is concluded in Section 4.

2 Proposed Algorithm

2.1 Research Strategy

This paper proposes a forecasting method for photovoltaic output under cold-wave weather, with the approach outlined in Fig. 1. The specific steps are as follows, (1) Define the criteria for determining cold-wave weather events and classify them into conventional and cold-wave weather categories; (2) Preprocess the sample data and augment the small sample data of cold-wave events using the TimeGAN; (3) Establish a prediction model based on WOA-CNN-LSTM using the augmented sample data and predict photovoltaic output, completing the model construction; (4) Compare the results of the prediction model with those of other similar models to verify the feasibility and accuracy of the proposed model.

2.2 Meteorological Feature Analysis of Cold-Wave Weather

2.2.1 Cold-Wave Weather Process and Meteorological Features

A cold-wave is a typical phenomenon characterized by rapid temperature drops, strong winds, and cold weather. Its formation process involves multiple factors such as the source of cold air, airflow transport, and terrain. The most significant feature of a cold-wave is the sharp drop in temperature, which typically occurs at the front of the cold-wave and may be accompanied by rain or snow. During the cooling process, the wind chill effect from strong winds further lowers the perceived temperature. Even if the actual temperature is not extremely low, it can significantly enhance the feeling of coldness. The meteorological features of a cold-wave are mainly manifested as rain or snow, strong winds, and rapid cooling. These characteristics can have a large impact on photovoltaic systems. In rain or snow conditions, due to regional differences, snow accumulation is minimal, but rainfall significantly reduces photovoltaic output. Before and after rainfall, cloud cover, and increased humidity may cause light scattering, further lowering the power generation efficiency of photovoltaic panels. In strong wind conditions, increased air turbidity reduces the transmission of solar radiation, and in extreme cases, it may even damage the photovoltaic support structure. Under extreme cooling conditions, ice or frost may form on the surface of the panels, affecting photovoltaic power generation efficiency. In summary, the impact of cold-wave weather on photovoltaic systems' performance and output mainly manifests as a decrease in efficiency, snow coverage, and potential equipment failure. Therefore, early forecasting and protective measures are crucial.

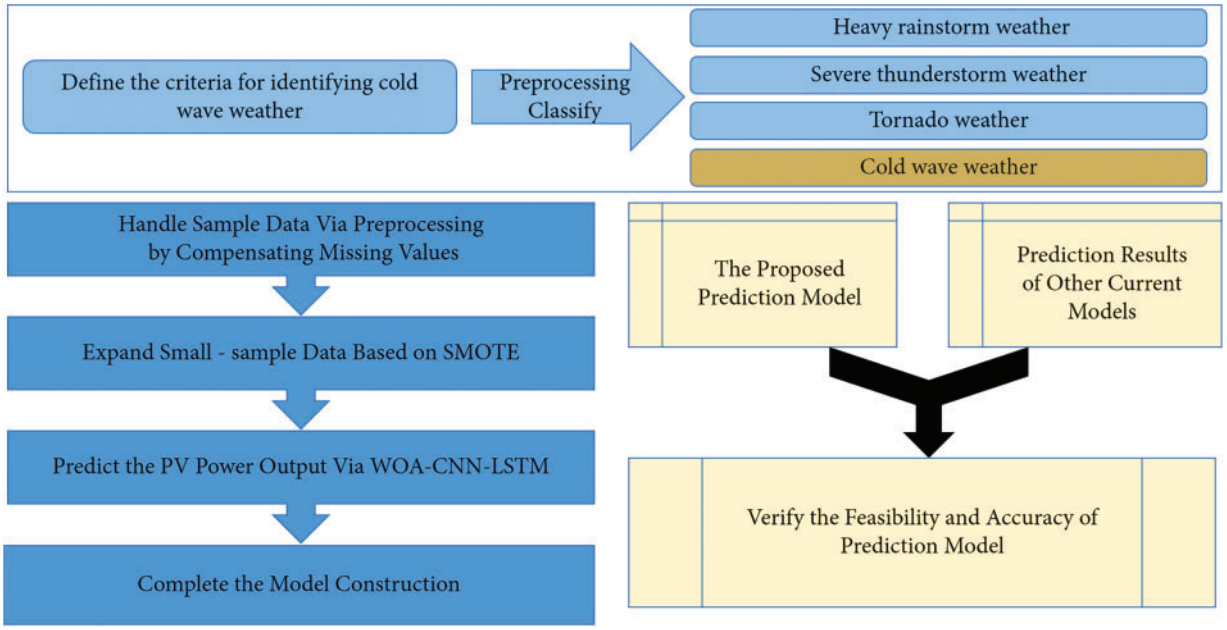

Figure 1: A method for small sample photovoltaic power prediction under cold-wave weather

According to the standards provided on the National Meteorological Administration website, a cold-wave event occurs after the passage of cold air, with a temperature drop of more than 8°C within 24 h or more than 10°C within 48 h, accompanied by temperatures dropping below 4°C. Based on comparative research and empirical analysis of cold-wave events in Jiangsu Province, this paper defines a cold-wave event as starting at the highest temperature when the rapid cooling begins and ending when the temperature rises above a low-temperature threshold of 5°C.

2.2.2 Data Sources

The data in this paper is sourced from a key scientific research project in Jiangsu Province, focusing on a photovoltaic power station located in a specific area of Jiangsu. The data used includes NWP (Numerical Weather Prediction) forecast data for major meteorological factors in 2022 and corresponding photovoltaic power output data for the same time period. The data processing flow of the proposed method are illustrated in Fig. 2, we take advantage of grubbs interpolation test to eliminate outliers and piecewise cubic Hermite interpolation to complete missing data. The few-shot dataset is expanded with TimeGAN, then processed with K-medoids clustering. The time resolution of the meteorological data is 15 min, while the time resolution of the photovoltaic output power data is 1 min. The meteorological data includes key indicators such as irradiance, relative humidity, temperature, and atmospheric pressure. Some of the original data and their corresponding time-series graphs are shown in Fig. 3.

According to the cold-wave warnings issued by the China Meteorological Network and local meteorological departments, and in accordance with the data standards outlined in Section 2.2.1, we have compiled the relevant data from three cold-wave events. A portion of the temperature variation trends is shown in Fig. 4, which includes one windstorm event and two rainfall events. A total of 1179 meteorological data samples from the cold-wave events were collected, which will be directly used for small sample augmentation processing.

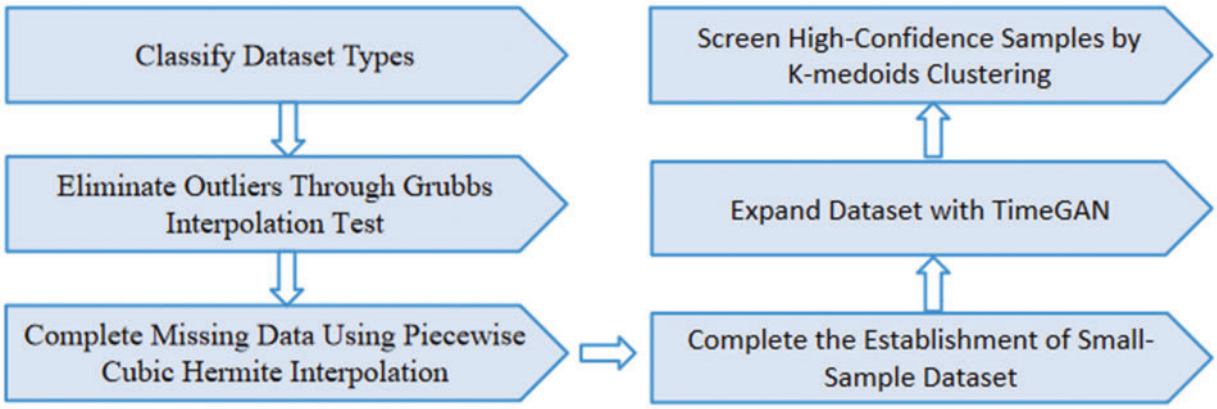

Figure 2: Data processing flow

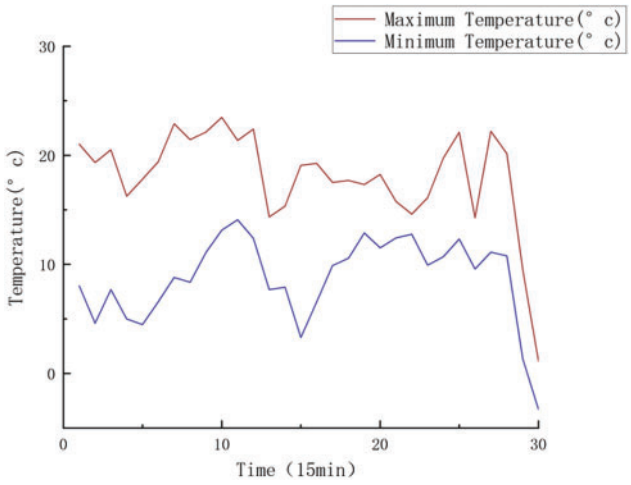

Figure 3: Partial display of the original data

2.3 Meteorological Data Processing

2.3.1 Outlier Detection and Handling

In the historical data of PV power stations, abnormally high or low values are commonly encountered, which are typically caused by measurement errors in the equipment, leading to power or irradiance data deviating from the actual levels. In addition, power data may be lower due to curtailment or limitations imposed on power generation by the grid. Curtailment is usually associated with grid overload or other technical issues that force the power station to reduce generation or even suspend operations. During routine inspections or scheduled maintenance periods, the station may shut down, resulting in power data being zero or significantly reduced for certain time intervals. These factors together contribute to the generation of anomalous data. In this paper, we use the Grubbs test for outlier detection, assuming that the data follows a normal distribution. The test compares the deviation of each data point from the mean to determine whether it is an outlier. The formula for the Grubbs test is as follows:

$$G = \frac{\left|X_i - \frac{1}{N}\sum_{i=1}^{N} X_i\right|}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{1}\left(X_i - \frac{1}{N}\sum_{i=1}^{N} X_i\right)^2}}, \tag{1}$$

$$G_{crit} = \frac{N-1}{\sqrt{N}} \times \sqrt{\frac{t^2}{N-2+t^2}}, \tag{2}$$

where t is the critical value based on the t-distribution with $N-2$ degrees of freedom, and G_{crit} is the critical value for the Grubbs test method.

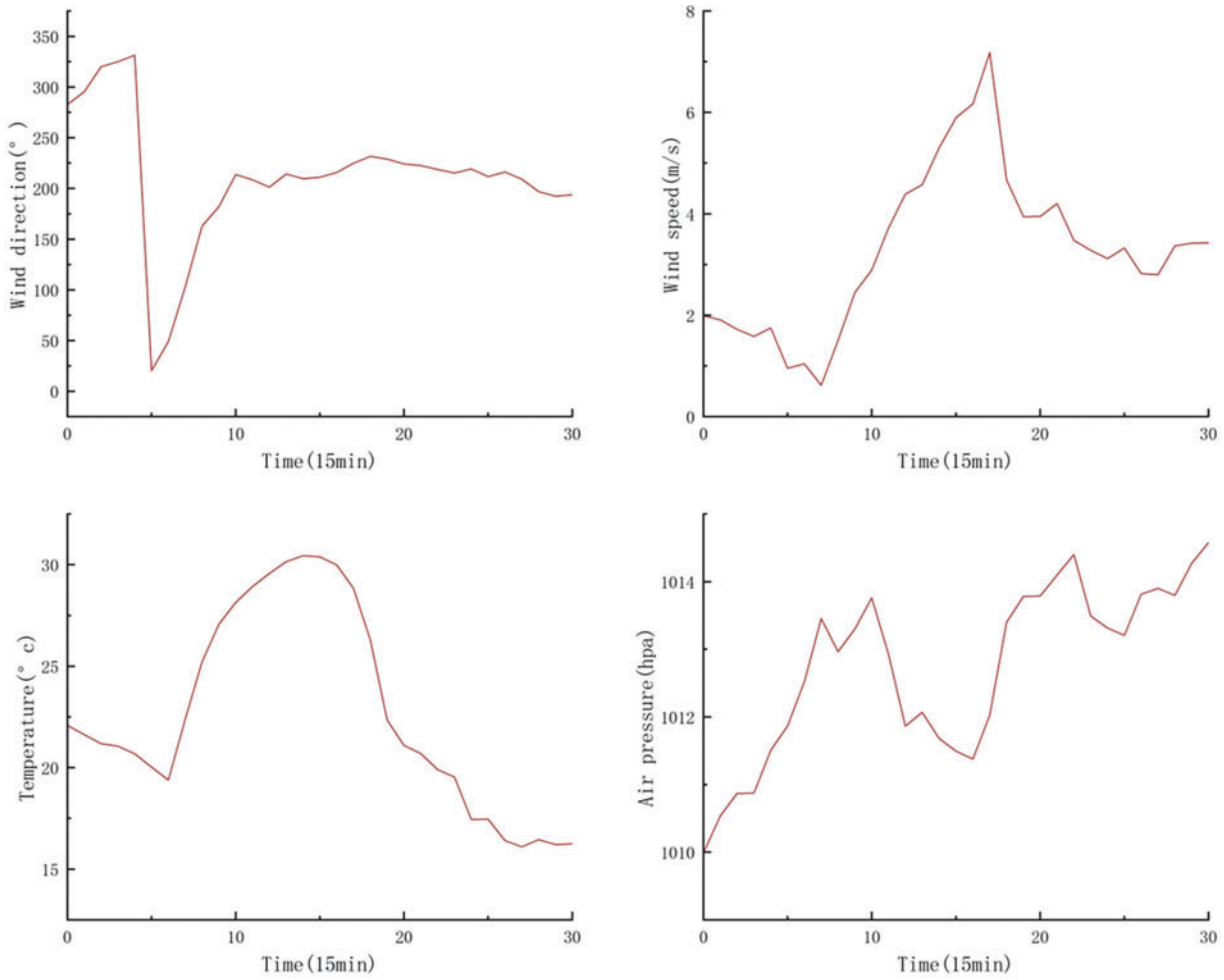

Figure 4: Temperature variation trend

2.3.2 Missing Data Handling

After using the G_{crit} method to detect and remove outliers, the amount of missing data increases, with missing values appearing in the form of "0", "null", or "0.00". When the amount of missing data is not large, it is generally preferred to impute these missing values. In general, when the number of missing data points does not exceed 50, this paper uses Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) for data imputation. The interpolation formula is as follows,

Let the observed time series be $T = \{t_1, t_2, \cdots, t_n\}$, and the corresponding photovoltaic characteristic sequence be $P = \{P_1, P_2, \cdots, P_n\}$, where some data points in the photovoltaic characteristic sequence are

missing. Let (t_i, P_i) represent the known data points, and the interpolation function for each segment $[t_i, t_{i+1}]$ is given by,

$$P_i(t) = \frac{m_{i+1} + m_i - 2\Delta_i}{(t_{i+1} - t_i)^2}(t_{i+1} - t_i)^3 + \frac{3\Delta_i - 2m_i - m_{i+1}}{t_{i+1} - t_i}(t_{i+1} - t_i)^2 + m_i(t_{i+1} - t_i) + P_i \tag{3}$$

where

$$m_i = 0, \Delta_i \times \Delta_{i-1} \leq 0 \tag{4}$$

$$m_i = \frac{3}{\frac{1}{\Delta_{i-1}} + \frac{1}{\Delta_i}}, \Delta_i \times \Delta_{i-1} > 0 \tag{5}$$

$$\Delta_i = \frac{P_{i+1} - P_i}{t_{i+1} - t_i}, i = 1, 2, \cdots, n-1 \tag{6}$$

2.4 The TimeGAN-Based Small Sample Data Augmentation for Cold-Wave Weather

2.4.1 TimeGAN-Based Few-Short Augmentation

To expand the meteorological NWP data severe weather conditions and the output data $X_{1\sim T} = \left\{\left(NWP_{gf}, P_{gf}\right) | NWP_{gf}, P_{gf} \in \Omega_Z . gf = 1, \cdots, T\right\}$ from T photovoltaic power stations, the overall sample space is set as Ω_Z, with the three loss functions including autoencoder loss, unsupervised loss, and supervised loss are defined as,

$$L_R = E_{X_{1T-P}}\left[\sum_t \left\|x_t - \widehat{x_t}\right\|_2\right], \tag{7}$$

$$L_U = E_{X_{1l-p}}\left[\sum_t \left\|\log y_t\right\|\right] + E_{X_{1l-\hat{p}}}\left[\sum_t \left\|\log\left(1 - \widehat{y_t}\right)\right\|\right], \tag{8}$$

$$L_s = E_{X_{1T-p}}\left[\sum_t \left\|h_t - g_X\left(h_{t-1}, Z_t\right)\right\|_2\right], \tag{9}$$

where E represents the Euclidean distance between two vectors, Z_t is a randomly generated vector, x_t is the NWP data with temporal characteristics for the t-th photovoltaic station, $\widehat{x_t}$ is the temporal data generated after learning, g_X represents the recurrent network in the generator, h_t is the hidden layer vector, y_t represents the actual sample class, and $\widehat{y_t}$ represents the generated sample class.

The parameter optimization for the embedding function and recovery function is defined by,

$$\min_{\theta_e} \theta_r\left(\lambda L_S + L_R\right). \tag{10}$$

Optimization of parameters in the generator and discriminator is denoted by,

$$\min_{\theta_g}\left(\eta L_S + \max_{\theta_d} L_U\right). \tag{11}$$

where θ_e, θ_r, θ_g and θ_d represent the parameters of the autoencoder network (embedding function and reconstruction function) and the adversarial generative network (generator and discriminator), where both λ and η are positive numbers.

TimeGAN, through the joint training of the autoencoder and adversarial network, is capable of capturing both the trends of sequences under real extreme weather conditions and ensuring that the feature distribution at each time slice aligns with the real sequence. To further enhance the stability and reliability of power data, this study employs the box plot method to handle and remove individual outlier data. The processed data is then used as the foundation for neural network modeling. This step ensures that the generated time series align with the distribution characteristics of real weather data, making the generated data more credible.

2.4.2 Data Quality Detection

Clustering algorithms partition a dataset into multiple clusters based on similarity metrics, such that the similarity within a cluster is as high as possible, while difference between clusters is as large as possible.

In this study, such algorithms are used to validate the quality of the data generated by TimeGAN. Unlike the commonly used K-means clustering algorithm, this paper employs K-medoids clustering, which uses the most similar actual sample in each cluster as the cluster center. This approach improves upon the former method, which is prone to getting stuck in local optima.

1) Data Preparation, the original data samples are combined with the synthetic samples generated by TimeGAN into a single dataset X.

2) Choosing the Number of Clusters k, the appropriate number of clusters is selected using the **Elbow Method** by calculating the total sum of squared errors (SSE) for different values of k. As the number of clusters increases, SSE will continually decrease, but when the number of clusters reaches a certain point, the rate of decrease in SSE slows down significantly. The point where this slowdown occurs is considered the optimal number of clusters. The specific calculation formula is as follows,

$$\mathrm{SSE} = \sum_{k=1}^{K} \sum_{i \in C_k} \left\| X_i - \mu_k \right\|^2 . \tag{12}$$

3) After determining the number of clusters, the K-medoids algorithm performs the following steps,

① Initialize the cluster center points,

② Assign each data point to the cluster center that is closest to it,

③ Update the cluster center points,

④ Repeat the above steps until the cluster centers stabilize.

4) To assess the quality of the clustering, the **Silhouette Score** is used to measure the cohesion of a sample point with its assigned cluster as well as the distance to the nearest cluster. This score indirectly reflects the accuracy of the TimeGAN-generated samples. The specific calculation formula is as follows, for a sample point i,

① The average distance from sample i to other samples in the same cluster, denoted as $\mathrm{a}(i)$, is calculated as follows,

$$\mathrm{a}(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} \left\| X_i - X_j \right\| , \tag{13}$$

where C_i is the cluster that sample i belongs to, $|C_i|$ represents the number of samples in the cluster, and $\left\| X_i - X_j \right\|$ denotes the Euclidean distance between sample points i and j.

② The average distance from sample i to the nearest cluster, denoted as $\mathrm{b}(i)$, is calculated as follows,

$$\mathrm{b}(i) = \min_{C_i \neq C_j} \frac{1}{|C_j|} \sum_{j \in C_j} \left\| X_i - X_j \right\|, \tag{14}$$

③ The **Silhouette Score** for sample i, denoted as $\mathrm{s}(i)$, is calculated as follows,

$$\mathrm{s}(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}, \tag{15}$$

After these preparations, the TimeGAN-based small sample augmentation and data sample quality evaluation method are described in Algorithm 1.

Algorithm 1: Integrated TimeGAN-KMedoids analytical framework

Input: Original dataset $\mathcal{D}_{orig} \in R^{T \times N}$
Output: Analysis result
Initialize $\mathcal{D}_{syn} \leftarrow \varnothing$
for $\mathrm{t} \leftarrow 1$ to M do
  $z_t \leftarrow N(0, \mathbf{I})$ $x_{\mathrm{t}} \leftarrow G_\theta(z_t, h_{t-1})$ $\mathcal{D}_{syn} \leftarrow \mathcal{D}_{syn} \cup \{x_{\mathrm{t}}\}$
end
$\mathcal{D}_{exp} \leftarrow \mathcal{D}_{orig} \cup \mathcal{D}_{syn}$
Initialize K cluster medoids $\left\{m_k^0\right\}_{k=1}^{K}$ randomly from $\mathcal{D}_{exp}$ repeat
  for $x_n \in \mathcal{D}_{exp}$ **do**
    $k_n \leftarrow 1 \le k \le K \left\| x_n - m_k^{(i)} \right\|_2$ $C_k^{(i)} \leftarrow C_k^{(i)} \cup \{x_n\}$
  end
  for cluster $\mathrm{k} \in \{1, \cdots, \mathrm{K}\}$ do
    $m_k^{(i+1)} \leftarrow x \in C_k^{(i)} \sum_{x_j \in C_k^{(i)}} \left\| x_n - m_k^{(i)} \right\|_2^2$
  end
  $i \leftarrow i + 1$
Until $\max_k \left\| x_n - m_k^{(i)} \right\|_2 < \varepsilon$;
Compute silhouette score s
If $\mathrm{s} > 0.7$ **then**
  $\mathrm{p} \leftarrow \mathrm{STATISTICALTEST}\left(\{C_k\}_{k=1}^{K}\right)$ if p<0.05 then
  Significant cluster structure detected
  else
  Non-significant cluster structure
else
If 0 < s ≤ 0.7 then
  While **s ≤ 0.7 and iter < max _iter do**
  Re-initialize medoids randomly
  Update $\{C_k\}, \{m_k\}, s$
  iter ← iter + 1
end
If s > 0.7then
  Improved clustering

(Continued)

Algorithm 1 (continued)

28 else
29 Maximum iterations reached
30 end
31 end

2.5 A Whale Optimization Algorithm-Based Neural Network Prediction Model

2.5.1 Convolutional Neural Network-Based Long Short-Term Memory Algorithm (CNN-LSTM)

Convolutional Neural Networks (CNNs), with their superior feature extraction capabilities, are widely used in time series analysis, especially in revealing the spatial relationships between multidimensional time series data. This paper proposes a hybrid neural network model combining CNN and LSTM for predicting the power output of photovoltaic fields. The model first extracts features from the input data using the CNN module, and then these features are passed as inputs to the LSTM module, ultimately achieving high-accuracy predictions for photovoltaic power output. The structure of the model is shown in Fig. 5.

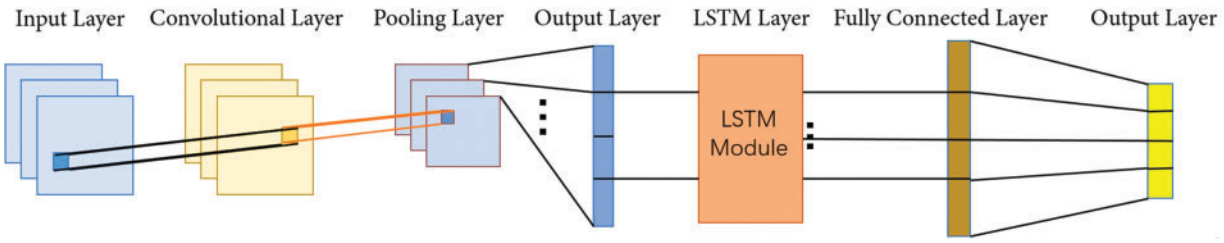

Figure 5: CNN-LSTM model structure diagram

The computation process of the model includes the following steps:

1) The data from the input layer is fed into the convolutional layer for feature extraction. A set of convolutional kernel functions is defined and applied to the input data for convolution operations. Then, by adding a bias value and using an activation function, the feature map of this layer is obtained. The convolution process can be represented by the following equation,

$$M_i^l = f\left(\sum_{i\in N_t} M_i^{l-1} \otimes w_{i,j}^l + b_j^l\right), \tag{16}$$

where M_i^l and M_i^{l-1} represent the i-th feature map of the l-th and $(l-1)$-th convolutional layers, respectively, b_j^l and $w_{i,j}^l$ represent the bias and weight between the i-th feature map of the l-th convolutional layer and the j-th feature map of the $(l-1)$-th convolutional layer, N_t represents the feature map set of the input layer, and $f(\cdot)$ denotes the activation function, with the Relu function being selected.

2) The data enters the pooling layer to reduce the feature dimension. Pooling operations help reduce sensitivity to input deformations. Average pooling is used to replace all values in the pooling region with the mean value of the region. The calculation can be represented as follows,

$$P_j^l = f\left(\beta_j^l down\left(M_i^{l-1}\right) + b_j^l\right), \tag{17}$$

where P_j^l represents the j-th feature map of the l-th pooling layer, β_j^l and b_j^l represent the multiplicative and additive biases of the feature map, respectively, $down(\cdot)$ represents the pooling function.

3) The data enters the flattening layer, which is used to "flatten" the input, i.e., transform the multidimensional input into a one-dimensional vector. This step is commonly used as a transition from the convolutional layer to the fully connected layer. After the features are extracted, the data is then fed into the LSTM module for prediction training.

4) The neurons in the fully connected layer are fully connected to the neurons in the LSTM hidden layer. The purpose is to integrate these features and generate the final output. The calculation formula is as follows,

$$F_j^l = f\left(\sum_{i\in l-1} w_i^l N_i^{l-1} + b_j^l\right), \tag{18}$$

where F_j^l represents the j-th neuron of the l-th layer.

2.5.2 Whale Optimization Algorithm-Based CNN-LSTM

The Whale Optimization Algorithm (WOA) is a metaheuristic optimization algorithm that simulates the hunting behavior of humpback whales. In WOA, each humpback whale's position represents a potential solution for photovoltaic power output prediction. By continuously updating the whale's position in the solution space, the algorithm ultimately finds the global optimum. The WOA algorithm is used to iteratively update and search for the optimal network parameters. The algorithm flow is shown in Fig. 6.

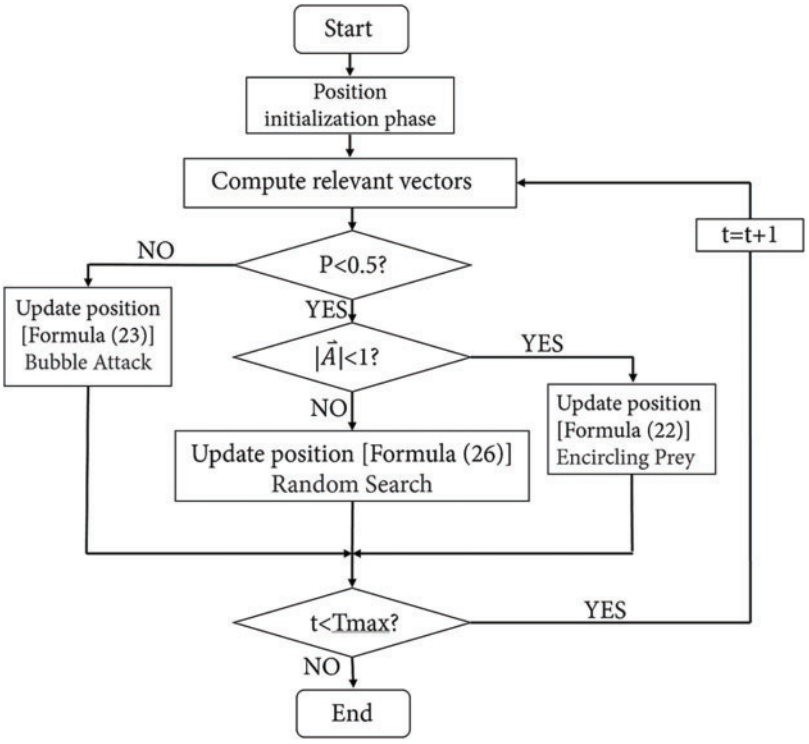

Figure 6: WOA algorithm flowchart

(1) We first initialize the population N, update the maximum number of iterations T_{max}, and compute the relevant vectors with the following equations,

$$\mathrm{a} = 2 \times \left(1 - \frac{t}{T_{max}}\right), \tag{19}$$

$$\vec{A} = 2a \cdot \vec{r}_1 - a, \tag{20}$$

$$\vec{C} = 2 \cdot \vec{r}_2, \tag{21}$$

where $\vec{r}_1$ and $\vec{r}_2$ are randomly generated vectors for each dimension independently, with values ranging from [0, 1].

(2) Then we introduce the encircling prey phase, the distance calculation formula and position update are defined as follows,

$$\vec{D} = \left|\vec{C} \cdot \vec{X} * (t) - \vec{X}(t)\right|, \tag{22}$$

$$\vec{X}(t+1) = \vec{X} * (t) - \vec{A} \cdot \vec{D}, \tag{23}$$

(3) Bubble attack phase (spiral update) is proposed to determine the spiral path and conduct the selection mechanism with the following equations,

$$\vec{X}(t+1) = \vec{D} \cdot e^{bl} \cdot \cos(2\pi l) + \vec{X} * (t), \tag{24}$$

$$\vec{X}(t+1) = Eq.(23), if\ (p < 0.5) \tag{25}$$

$$\vec{X}(t+1) = Eq.(24), if\ (p \geq 0.5) \tag{26}$$

where b is the spiral shape constant, set to 1 in this paper; l and p are both random numbers, where l follows a uniform distribution in [−1, 1]; p follows a uniform distribution in [0, 1], determining the position update method.

(4) When $\left|\vec{A}\right| \geq 1$, randomly select a whale individual for global search, the random individual distance and position update are calculated by,

$$\underset{rand}{\vec{D}} = \left|\vec{C} \cdot \underset{rand}{\vec{X}} - \vec{X}(t)\right|, \tag{27}$$

$$\vec{X}(t+1) = \underset{rand}{\vec{X}} - \vec{A} \cdot \underset{rand}{\vec{D}}. \tag{28}$$

where $\underset{rand}{\vec{X}}$ is the position vector of the randomly selected whale.

Based on the above method, we encode the key parameters in the CNN-LSTM model as position vectors $\vec{X}$ for parameter optimization, and the specific settings will be discussed in Section 3.

2.6 Small Sample Photovoltaic Power Prediction Method under Cold-Wave Weather

2.6.1 Prediction Method Framework

This paper presents a model to predict the power output of photovoltaic power plants under cold-wave weather conditions. The specific steps are as follows, 1) Preprocess the input data. First, unify the time scale of the NWP meteorological data and power data. Then, classify the data into normal weather samples and cold-wave weather samples based on the meteorological data. After that, use the Grubbs test to identify abnormal photovoltaic power output data under normal weather conditions, and perform normalization after supplementing or removing the data, 2) Use TimeGAN to augment the meteorological and power data under cold-wave weather conditions. Then, validate the quality of the augmented meteorological and power data using K-medoids clustering, 3) Build and train the prediction model, and predict the results. First, select the two most relevant factors—photovoltaic output and total irradiance—as input variables, forming a hard parameter sharing model for MTL, where the lower-level weights are shared. Then, connect the shared layer to the CNN module for feature extraction. Next, use the LSTM part for prediction. Finally, input the test dataset, obtain the prediction results, and perform inverse normalization before outputting the corresponding curve graph, 4) Use Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 to evaluate the model's error and analyze the results. Validate the performance of this model for photovoltaic power prediction under extreme weather conditions.

2.6.2 Error Validation

The accuracy of the model is typically reflected by the error between the predicted results and the actual values. In this paper, the model's performance is evaluated using indicators such as MAE and RMSE. The calculation formulas are as follows,

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{y_i} - y_i\right|, \tag{29}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{y_i} - y_i\right)^2}, \tag{30}$$

where $\widehat{y_i}$ represents the predicted value, y_i represents the actual value, and n represents the number of prediction sample points.

The level of fit between the predicted results and the actual values can generally be observed visually through graphs. The coefficient of determination (R^2) quantifies the goodness of fit. The calculation method is as follows,

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{m}\left(\widehat{y_i} - y_i\right)^2}{\sum_{i=1}^{m}\left(\widehat{y_i} - \overline{y_i}\right)^2}, \tag{31}$$

where $\overline{y_i}$ represents the mean of the actual values.

3 Test and Analysis

3.1 Generated Data Analysis

The quantity and quality of the dataset are the most fundamental factors influencing the prediction results of data-driven models. A total of 32,714 samples were generated from the 1179 cold-wave weather sample data described in Section 2.2.2 of this paper. After performing quality detection and preliminary screening of the data using K-medoids clustering, the comparison between the generated data and the original data values is shown in Fig. 7. From the images below, it can be intuitively seen that the data samples generated by TimeGAN exhibit a distribution pattern similar to that of the original data samples.

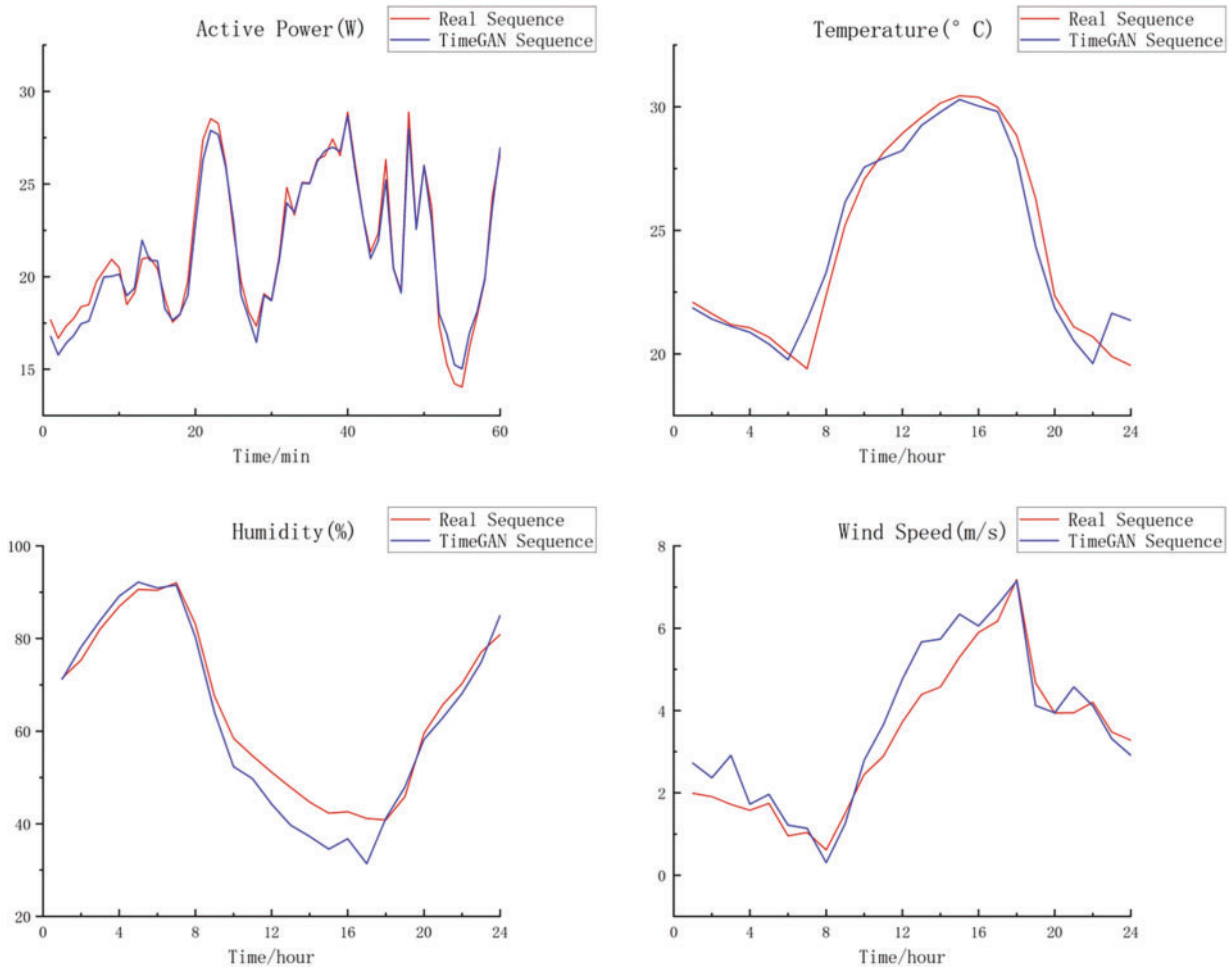

Figure 7: Comparison of generated data with original data values

To further verify the reliability of TimeGAN in data augmentation under weather sample conditions, other data augmentation methods that generate more similar data using a small amount of data and prior knowledge to expand the training dataset are used to address the small sample problem in real-world scenarios. These methods include Smote-based augmentation and GANs that learn sample features through neural networks. By comparing the augmented samples with the validation set, the LSTM neural network prediction results under different strategies are shown in Table 1.

Table 1: Comparison of prediction results for different data augmentation methods

Type	MAE	RMSE	R^2
Smote	11.35	12.35	0.669
GAN	5.32	6.72	0.768
TimeGAN	4.26	4.72	0.796

Unlike Smote, which may suffer from overfitting, TimeGAN outperforms traditional GANs when handling time series data. This is because TimeGAN is specifically designed to learn temporal features, making it more effective in time series generation and other time-dependent tasks. In conclusion, this indicates that TimeGAN can effectively learn the meteorological and output characteristics under severe weather conditions and generate sample data that aligns with the characteristics of real data.

3.2 Prediction Results Analysis

The training and validation of the model used in this study were conducted on the TensorFlow platform. Firstly, we set the parameters of the WOA optimization model. Besides the parameters of the algorithm itself, we first impose range restrictions on the relevant hyperparameters of the CNN-LSTM model. These ranges neither cause search difficulties nor limit the improvement space due to being too small. The specific parameter settings are as shown in Table 2.

Table 2: WOA model parameter settings

Parameters	Population_size	Max_iter	Learning_rate	CNN_filters	Dropout_rate
Value	30	100	[0.0001,0.01]	[16,128]	[0.1,0.5]

In this study, the WOA algorithm serves as the outer optimizer. After generating a set of hyperparameter combinations each time, the TensorFlow platform builds and trains the CNN-LSTM model based on this hyperparameter combination once, and feeds back the model performance metrics (such as RMSE) calculated on the validation set as fitness to the WOA algorithm. Subsequently, WOA dynamically adjusts the position of the whale group based on the fitness results corresponding to each hyperparameter combination, iteratively searches for a better hyperparameter configuration until the optimization termination condition is met. Throughout the entire process, hyperparameter optimization and model training are executed in a nested manner. That is, each hyperparameter update is accompanied by a complete CNN-LSTM model training and evaluation, thereby achieving continuous improvement of the model performance.

Finally, the optimal Learning_rate was selected as 0.0015 and the optimal CNN_filters as 72 as the hyperparameter combination of this prediction model.

Furthermore, in order to better reflect the accuracy of the WOA optimization algorithm we used, we compared it with two other optimization algorithms (GA and PSO). Meanwhile, the optimized CNN-LSTM model was used for the prediction of the same data set, and the relevant data indicators were compared, as shown in Table 3.

Based on the modeling described above, the expanded data samples were used for prediction to further validate the feasibility of TimeGAN for small sample expansion in photovoltaic power prediction under

severe weather conditions. At the same time, comparisons were made with the other two prediction models, and the comparison between predicted results and actual values is shown in Fig. 8.

Table 3: Verification of model accuracy under different optimization algorithms

Type	MAE	RMSE	R^2
GA+CNN-LSTM	4.28	4.79	0.649
PSO+CNN-LSTM	3.56	4.32	0.792
WOA+CNN-LSTM	3.17	4.29	0.875

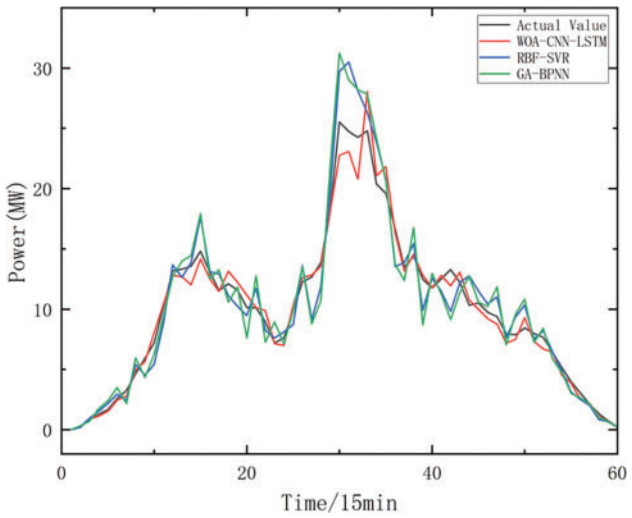

Figure 8: Comparison of photovoltaic output predictions based on different models

By observing the curve graph, we can intuitively find that the power peak is about 25 MW. When the power is low, the curve fitting of the three models is high and the prediction error is low. The curve fitting at the peak is low and the prediction error is the largest, among which the BP model has the largest curve interpolation.

In addition to the intuitive comparison of the predicted result curves, error metrics can be used to quantify the accuracy of the models. Since the photovoltaic power output can have values of 0 during the nighttime, making it impossible to calculate the MAPE value, this metric is not considered. The SVM decomposition method can sometimes dilute data features, leading to poorer results, so this method is not considered in this model. The R^2 values and prediction errors for the three models from the photovoltaic power station data are shown in Table 4.

From the model perspective, the GA-BPNN model has the lowest fit, followed by the RBF-SVR model, and the WOA-CNN-LSTM model has the highest fit. According to the MAE and RMSE metrics, the GA-BPNN model has the largest error. The MAE of the WOA-CNN-LSTM model is reduced by 0.22 compared to the CNN-LSTM model, with a reduction ratio of 6.49%; the RMSE is reduced by 0.49, with a reduction ratio of 10.25%. The WOA-CNN-LSTM model can reduce prediction errors by approximately 30% compared to traditional neural network models.

Table 4: Predictions error metrics for three models of photovoltaic power stations

Type	MAE	RMSE	R^2
TimeGAN+GA-BPNN	6.45	6.89	0.598
TimeGAN+RBF-SVR	3.39	4.78	0.809
TimeGAN+WOA-CNN-LSTM	3.17	4.29	0.875

4 Conclusion

This paper primarily studies the photovoltaic output prediction model based on WOA-CNN-LSTM under cold-wave weather conditions with small sample sizes. First, TimeGAN was used to generate and expand the small sample data, and K-medoids was applied to analyze the quality of the generated data. A horizontal comparison was made between different data augmentation methods, revealing that the TimeGAN algorithm performs better for small sample expansion of photovoltaic output under cold-wave weather. Then, a CNN-LSTM prediction model was built, and the WOA algorithm was used to optimize the hyperparameters of the model. Finally, the dataset and validation set were divided to verify the accuracy of the model. The results showed that the WOA-CNN-LSTM model performs better than other neural network prediction methods for few-short photovoltaic output prediction under cold-wave weather, providing a practical and reliable prediction method.

The experimental results demonstrate that our algorithm effectively addresses the issue of small sample prediction under extreme weather conditions, and by using the optimized neural network algorithm for prediction, the accuracy of the model is improved. Looking ahead, the prediction of new energy photovoltaic output under extreme weather conditions holds immense potential. With the development of technology, we expect this field to continue advancing, adapting to new challenges and opportunities, and revolutionizing various industries while enhancing our daily lives.

Acknowledgement: Not applicable.

Funding Statement: This study is supported by Science and Technology Projects of Jiangsu Province (No. BE2022003) and Science and Technology Projects of Jiangsu Province (No. BE2022003-5).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Ruiheng Pan and Gang Ma; data collection: Ruiheng Pan and Shuyan Wang; analysis and interpretation of the results: Ruiheng Pan and Yihan Huang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data supporting this study are included within the article.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Liao PY, Lin T, Ali Zargar O, Hsu CJ, Chou CH, Shih YC, et al. Energy consumption and carbon emission reduction in HVAC system of a dynamic random access memory (DRAM) semiconductor fabrication plant (fab). IEEE Trans Semicond Manufact. 2024;37(2):174–84. doi:10.1109/tsm.2024.3379949.
2. Chen Y, Lin C, Zhang Y, Liu J, Yu D. The balance issue of the proportion between new energy and traditional thermal power: an important issue under today's low-carbon goal in developing countries. Renew Energy. 2024;231(1):121018. doi:10.1016/j.renene.2024.121018.

3. Hu H, Wen L, Zheng K. Short-term optimization of multi-source system with high proportion of wind power and photovoltaic. In: 2021 IEEE 4th International Conference on Renewable Energy and Power Engineering (REPE); 2021 Oct 9–11; Beijing, China. p. 371–5. doi:10.1109/repe52765.2021.9617052.
4. Pereira S, Canhoto P, Oozeki T, Salgado R. Comprehensive approach to photovoltaic power forecasting using numerical weather prediction data and physics-based models and data-driven techniques. Renew Energy. 2025;251:123495. doi:10.1016/j.renene.2025.123495.
5. Ding M, Wang WS, Wang XL, Song YT, Chen DZ, Sun M. A review on the effect of large-scale PV generation on power systems. Proc CSEE. 2014;34(1):1–14. (In Chinese). doi:10.13334/j.0258-8013.pcsee.2014.01.001.
6. Mellit A, Pavan AM, Lughi V. Deep learning neural networks for short-term photovoltaic power forecasting. Renew Energy. 2021;172(2):276–88. doi:10.1016/j.renene.2021.02.166.
7. Wang H, Yan J, Zhang J, Liu S, Liu Y, Han S, et al. Short-term integrated forecasting method for wind power, solar power, and system load based on variable attention mechanism and multi-task learning. Energy. 2024;304(1):132188. doi:10.1016/j.energy.2024.132188.
8. Li B, Zhao H, Wang X, Zhao Y, Zhang Y, Lu H, et al. Distributionally robust offering strategy of the aggregator integrating renewable energy generator and energy storage considering uncertainty and connections between the mid-to-long-term and spot electricity markets. Renew Energy. 2022;201(2):400–17. doi:10.1016/j.renene.2022.10.117.
9. Masuta T, Oozeki T, da Silva Fonseca JG, Murata A. Impact of forecast error of photovoltaic power output on demand and supply operation in power systems. In: 2014 Power Systems Computation Conference; 2014 Aug 18–22; Wroclaw, Poland. p. 1–7.
10. Chindarkkar A, Priyadarshi S, Shiradkar NS, Kottantharayil A, Velmurugan R. Deep learning based detection of cracks in electroluminescence images of fielded PV modules. In: 2020 47th IEEE Photovoltaic Specialists Conference (PVSC); 2020 Jun 15–21; Calgary, AB, Canada. p. 1612–6. doi:10.1109/pvsc45281.2020.9300615.
11. Abdel-Fattah MF, Sigurðsson GÁ. Extreme weather effects on transmission grids and the potential of the utilization of real time monitoring for smart grids applications. In: 2023 23rd International Scientific Conference on Electric Power Engineering (EPE); 2023 May 24–26; Brno, Czech Republic. p. 1–6. doi:10.1109/EPE58302.2023.10149280.
12. Gao F, Hou H, Li S, Wei R, Wang L, He H. Prediction of power outage areas in distribution networks during typhoon disasters. In: 2023 2nd Asia Power and Electrical Technology Conference (APET); 2023 Dec 28–30; Shanghai, China. p. 506–10. doi:10.1109/APET59977.2023.10489708.
13. Zhang X, Zhao H, Yao J, Wang Z, Zheng Y, Peng T, et al. A multi-scale component feature learning framework based on CNN-BiGRU and online sequential regularized extreme learning machine for wind speed prediction. Renew Energy. 2025;242(P1):122427. doi:10.1016/j.renene.2025.122427.
14. Li S, Xin P, Zou G. Short-term forecasting of PV power based on the fuzzy clustering algorithm and support vector machine in smart distribution planning. In: 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC); 2018 Oct 12–14; Chongqing, China. p. 643–7. doi:10.1109/IAEAC.2018.8577853.
15. Jin SL, Chi YN, Wang B, Liu XL. Application research of meteorological sensing and monitoring technology in power grid. Electr Power Inf Commun Technol. 2020;18(4):84–90. (In Chinese). doi:10.16543/j.2095-641x.electric.power.ict.2020.04.010.
16. Dai Q, Duan SX, Cai T, Chen CS, Chen ZH, Qiu C. Short-term PV generation system forecasting model without irradiation based on weather type clustering. Proc CSEE. 2011;31(34):28–35. (In Chinese). doi:10.13334/j.0258-8013.pcsee.2011.34.026.
17. Yuan XL, Shi JH, Xu JY. Short-term power forecasting for photovoltaic generation considering weather type index. Proc CSEE. 2013;33(34):57–64. (In Chinese). doi:10.13334/j.0258-8013.pcsee.2013.34.011.
18. Wang P, Guo J, Cheng F, Gu Y, Yuan F, Zhang F. A MPC-based load frequency control considering wind power intelligent forecasting. Renew Energy. 2025;244:122636. doi:10.1016/j.renene.2025.122636.
19. Deng W, Dai Z, Chen R, Wang H, Lu S, Li C, et al. Wind power interval prediction based on CGAN and KELM under extreme weather scenarios. In: 2023 IEEE/IAS Industrial and Commercial Power System Asia (I&CPS Asia); 2023 Jul 7–9; Chongqing, China. p. 2341–6. doi:10.1109/ICPSAsia58343.2023.10294890.

20. Wei P, Wang X, Xiong Y, Peng X, Yang Z, Guo S, et al. Short-term wind power probabilistic prediction for newly built wind farms scenarios based on W-GAN and CNN-LSTM. In: 2022 IEEE 6th Conference on Energy Internet and Energy System Integration (EI2); 2022 Nov 11–13; Chengdu, China. p. 3111–6. doi:10.1109/EI256261.2022.10116968.
21. Xian J, Meng A, Fu J. Short-term load probability prediction based on conditional generative adversarial network curve generation. IEEE Access. 2024;12:64165–76. doi:10.1109/access.2024.3395659.