AIMS Energy

AIMS Energy, 13(5): 1104–1132.
DOI: 10.3934/energy.2025041
Received: 18 June 2025
Revised: 24 August 2025
Accepted: 05 September 2025
Published: 12 September 2025

https://www.aimspress.com/journal/energy

Research article

Spatiotemporal attention network for ultra-short-term photovoltaic power forecasting considering spatiotemporal correlations and multiple environmental factors

Ming Yang[1], Zehao Wang[2] and Haipeng Chen[2,*]

[1] Shandong University School of Electrical Engineering, Jinan 250100, China
[2] Key Laboratory of Modern Power System Simulation and Control & Renewable Energy Technology, Ministry of Education, Northeast Electric Power University, Jilin 132012, China

* **Correspondence:** Email: haipeng0704@126.com.

Abstract: The accuracy of photovoltaic (PV) power forecasting greatly influences power system operation and control. However, models often fail to simultaneously capture spatial correlations among distributed PV stations and temporal patterns in power time series. To overcome this challenge, we proposed a Spatiotemporal Attention Network (STAN). First, autocorrelation, cross-correlation, and smoothing effects in PV systems were examined, forming a theoretical foundation for prediction. Then, multi-head self-attention was applied to extract spatial features across stations, while a sequence-to-sequence model with global attention captured temporal dependencies. Case studies demonstrated that compared with conventional Convolutional Neural Network-Long Short-Term Memory, STAN reduced MAE by 45.6% and RMSE by 32.8%, effectively enhancing forecasting accuracy and minimizing prediction errors.

Keywords: spatiotemporal attention network; spatiotemporal correlation; PV power prediction; attention mechanism

Abbreviations: LSTM: Long Short-Term Memory; GRU: Gated Recurrent Unit; CNN: Convolutional Neural Network; GCN: Graph Convolutional Network; GNN: Graph Neural Network; Seq2Seq: Sequence-to-Sequence model; STAN: Spatiotemporal Attention Network; MAE: Mean Absolute Error; RMSE: Root Mean Squared Error; R^2: Coefficient of Determination; PV: Photovoltaic;

ReLU: Rectified Linear Unit; VIF: Variance Inflation Factor; MAD: Median Absolute Deviation; FLOPs: Floating-Point Operations; BiLSTM: Bidirectional Long Short-Term Memory; (X_i, Y_i) : Samples drawn from the population (X, Y); $\bar{X} / \bar{Y}$: The sample means; ρ : The Pearson correlation coefficient; $\bar{x}$: The mean of the selected sample; $\hat{r}_k$: The computed autocovariance coefficient value in the time-series data; k : The lag order of the autocovariance; n : The number of samples; $\mathrm{P}_{\mathrm{N,i}}$: The installed capacity of the i-th photovoltaic power station; $P_i(t)$: The mean generation of the i-th photovoltaic power station; $P_{\Sigma}(t)$: Total generation of the i-th photovoltaic power station; ∂ : The smoothing coefficient; $\sigma_{cluster}$: The absolute standard deviations of the distributed photovoltaic output; σ_{single} : The reference photovoltaic power station output; r_{x1y} : The coefficient of correlation between the feature value x1 and the predicted power y1; $x_{1,i}$: The i-th historical data of the feature value; $y_{1,i}$: The i-th forecasting power value; $\bar{x} / \bar{y}$: The average values of the feature value and predicted power, respectively; m : The number of predictions; d_m : The dimensionality of the model; W^I : The learnable weight matrices; W_i^Q : The learnable weight matrices; W_i^K : The learnable weight matrices; W_i^V : The learnable weight matrices; W^1 : The learnable weight matrices; W^2 : The learnable weight matrices; U^E : The learnable weight matrices; W^E : The learnable weight matrices; W^C : The learnable weight matrices; d_{ffn} : The dimensionality of the model; O_t : The layer normalization function; LN: The layer normalization function; d_e : The dimensionality of the model; φ : The tanh activation function; h_0 : The initial zero state; a_{kj} : The element of the weight vector $a_t \in ¡^{1\times T}$; s_k : The hidden state of the decoder; d_d : The size of hidden units in the decoder; [;]: The concentration operation; $p_i^{'} / p_i^{'} / \overline{p_i}$: The actual value, the predicted value, and the mean of the predicted values of photovoltaic power; Q : The query vector; K : The key vector; V : The value vector; I_t : The input vectors

1. Introduction

In the context of the low-carbon transition of energy systems, accurate photovoltaic (PV) power forecasting is crucial for secure scheduling, renewable energy integration, and economic operation. Due to geographic heterogeneity, fluctuations in solar irradiance, and temperature variations, PV power series exhibit significant non-stationarity and spatiotemporal correlations [1]. Without jointly modeling the spatial relationships among distributed PV plants and the multi-scale temporal patterns in power time series, forecasting errors may increase reserve requirements and curtailment, thereby compromising grid reliability [2]. Therefore, under meteorological uncertainty, it is essential to develop a forecasting framework that can efficiently capture dynamic non-Euclidean spatial dependencies and long-range temporal behaviors, which is vital for enhancing system reliability and renewable energy utilization.

1.1. Literature review

Distributed PV power generation exhibits significant spatiotemporal variability, and accurately analyzing its spatiotemporal correlations is a key prerequisite for improving forecasting accuracy and optimizing scheduling strategies. The researchers in [3] proposed a correlation analysis framework based on mutual information theory, demonstrating that weather patterns create complex spatial associations that go beyond geographic distance. The researchers in [4] developed a dynamic correlation analysis method using a sliding window technique to capture the time-varying dependencies of PV power generation. However, such statistical methods struggle to capture the inherent nonlinear multi-scale correlations within distributed PV systems. Dai et al. [5] proposed an online learning framework to adaptively update correlation patterns, highlighting the importance of dynamic correlation modeling in renewable energy forecasting. Nevertheless, this method mainly focuses on temporal autocorrelation and overlooks cross-site cross-correlation effects. The researchers in [6] attempted to handle spatiotemporal data through tensor decomposition methods, revealing hidden cross-dimensional correlation structures. The researchers in [7] further incorporated meteorological factors into the analysis, constructing a multivariate correlation matrix considering irradiance, temperature, and cloud cover. These studies, through methods, such as mutual information theory, dynamic correlation analysis, tensor decomposition, and multivariate correlation, have deepened the understanding of the spatiotemporal characteristics of PV power generation and effectively improved the accuracy of forecasting models. However, these studies lack a unified theoretical framework to simultaneously address autocorrelation, cross-correlation, and smoothing effects, limiting a comprehensive understanding of the complex spatiotemporal coupling mechanisms in distributed PV systems.

Distributed PV plants are characterized by geographic dispersion, irregular topology, and meteorologically driven dynamic correlations, resulting in spatial dependencies that exhibit non-Euclidean structures. This necessitates overcoming the limitations of traditional modeling methods and developing advanced techniques capable of adaptively capturing complex spatial interactions. The researchers in [8] employed a convolutional neural network (CNN) to extract spatial features from gridded meteorological data, improving the forecasting accuracy of regional PV systems. However, CNNs assume regular grid structures and cannot accurately model the irregular topology of distributed PV networks. The researchers in [9] introduced a graph neural network (GNN) for PV forecasting to address this limitation. Dimitriadis et al. [10] extended this approach to multi-country interconnected systems, demonstrating the scalability of deep learning frameworks for large-scale forecasting. Passalis et al. [11] proposed adaptive normalization techniques to handle regional spatial heterogeneity and enhance model generalization. Nevertheless, these graph-based methods typically rely on predefined adjacency matrices based on geographic distance or electrical connectivity, which fail to capture the actual correlation patterns influenced by dynamic meteorological factors. The researchers in [12] attempted to learn dynamic graph structures using attention mechanisms, but their method is limited to pairwise relationships and cannot capture higher-order spatial interactions. These studies have progressively improved the modeling of spatial dependencies in distributed PV systems through CNN-based grid feature extraction, GNN-based irregular topology modeling, attention-based dynamic learning, and hypergraph multi-to-multi associations, effectively enhancing forecasting performance in large-scale systems. However, they have not established an adaptive learning framework to dynamically capture non-Euclidean spatial correlations evolving with meteorological conditions and seasonal patterns, limiting the model's ability to represent complex time-varying spatial dependencies.

PV power generation exhibits multi-scale temporal characteristics, including short-term intraday fluctuations, seasonal periodic variations, and long-term evolutionary trends. This necessitates time series modeling that can simultaneously capture local dynamics and global dependencies, requiring the development of advanced methods capable of effectively handling multi-resolution temporal features and performing multi-step ahead forecasting. The researchers in [13] applied LSTM networks combined with attention mechanisms for forecasting, demonstrating superior time series modeling performance compared to traditional recurrent models. The researchers in [14] proposed a multi-head temporal attention model to separately process different frequency components, optimizing the decomposition of seasonal and daily patterns. The researchers in [15] introduced a hierarchical attention mechanism operating across hourly to monthly scales. However, these models are prone to gradient vanishing issues when modeling ultra-long sequences, limiting their ability to capture annual patterns and long-term dependencies. Zhou et al. [16] developed a parallel attention architecture that efficiently handles multi-resolution temporal features, but its segmented processing mechanism overlooks meaningful cross-scale interactions. The researchers in [17] proposed a wavelet-attention hybrid model that decomposes time series into different frequency bands before applying attention mechanisms, enhancing multi-scale feature extraction. The researchers in [18] introduced a causal attention mechanism that expands the receptive field while maintaining computational efficiency. These studies, through methods such as LSTM-attention fusion, multi-head frequency decomposition, hierarchical temporal modeling, and Transformer positional encoding, have enhanced the representation of PV power's multi-scale temporal characteristics, effectively improving short-term forecasting accuracy and long-term trend capture. However, they have not achieved an efficient encoder-decoder architecture that maintains global context awareness, limiting the ability to balance local pattern recognition and long-range dependency modeling in multi-step ahead forecasting tasks.

1.2. Research gap

To highlight the major differences between the methodology proposed in this paper and other studies, the uniqueness of this study is highlighted by summarizing the major methods used in this study and the most relevant studies in Table 1.

In the field of PV power forecasting research, although numerous constructive approaches have been proposed at this stage, the following unresolved issues remain:

(1) Most studies in distributed PV forecasting solely employ LSTM/GRU for temporal modeling or CNN/GCN for spatial modeling. These studies lack a systematic analysis of the spatiotemporal correlation characteristics (such as autocorrelation, cross-correlation, and smoothing effects) inherent in distributed PV systems.

(2) Regarding spatial feature extraction, researchers often neglect spatial correlations or assumes CNN models with regular grid structures. For geographically irregularly distributed PV power plant clusters, there is no effective mechanism to analyze non-Euclidean spatial relationships, making it challenging to capture interactions among plants at different locations.

(3) Existing studies predominantly rely on standard RNN or Transformer architectures for single-scale temporal modeling. However, PV power time series exhibit multi-scale temporal patterns, and current methods cannot simultaneously capture short-term fluctuations and long-term dependencies through a unified attention mechanism.

Table 1. A comparative summary of this study and other publications.

Ref.	Method	Spatial Attention	Temporal Attention	Spatiotemporal Analysis	Multi-scale Modeling
[3]	Multi-factor spatio-temporal correlation	√	×	√	×
[4]	BO-LSTM with time-frequency correlation	×	×	√	×
[5]	Self-attention with online learning	×	√	×	×
[6]	VMD-LSTM	×	×	×	×
[7]	HP-OVMD-ENN	×	×	×	×
[8]	COA-CNN-LSTM	×	×	×	×
[9]	GCN-LSTM	√	√	√	×
[10]	Deep learning framework (multi-country)	×	×	×	×
[11]	Residual adaptive normalization	×	×	×	×
[12]	Attention-based CNN-LSTM	×	√	×	×
[13]	QT-MARF	×	√	×	√
[14]	Fine-grained temporal & cloud spatial attention	√	√	×	×
[15]	Cross-modal correlation attention	√	√	×	×
This paper	STAN	√	√	√	√

1.3. Contributions of this work

(1) To resolve the insufficient analysis of spatiotemporal correlations in distributed PV systems, we introduce a systematic spatiotemporal correlation analysis method. This approach thoroughly explores autocorrelation, cross-correlation, and smoothing effects, providing a solid theoretical foundation for constructing more accurate forecasting models and enabling a comprehensive understanding of PV power's spatiotemporal characteristics.

(2) To overcome traditional convolutional neural networks' limitations in handling non-Euclidean spatial topologies, we design an innovative spatial self-attention mechanism. This mechanism adaptively learns mutual influences among geographically dispersed PV plants, effectively analyzing non-Euclidean spatial relationships in distributed PV clusters, thereby achieving dynamic modeling of complex spatial dependencies.

(3) To address the inadequate capture of multi-scale temporal features and limited modeling of long-term dependencies in PV power time series, we construct a temporal attention module based on the Seq2Seq architecture. Through a global attention mechanism, this module accurately captures multi-scale temporal features, significantly enhancing the modeling capability for long-term sequential dependencies in PV power.

1.4. Paper structure

The remainder of this paper is organized as follows: Section 2 contains the theoretical foundation of correlation analysis and the spatiotemporal correlation analysis of distributed PV systems. In Section 3, we provide a detailed introduction to the proposed spatiotemporal attention network methodology, including the design of spatial self-attention mechanisms and temporal attention modules. In Section 4, we present experimental results through case studies and conduct a comparative analysis with existing methods. In Section 5, we summarize the research conclusions and discuss future research directions.

2. Correlation analysis

2.1. Data preprocessing and feature selection

Before conducting correlation analysis or model training, a systematic data preprocessing pipeline was employed to ensure the quality, reliability, and representativeness of the input dataset. The raw dataset contained one-year observational records from five geographically distributed PV power stations, collected every 15 minutes, covering irradiance, ambient temperature, atmospheric pressure, relative humidity, and actual PV power output.

(1) Missing value treatment:

Missing entries, primarily caused by sensor failures or temporary communication interruptions, accounted for 0.8% of the total dataset. For continuous missing values less than or equal to 2 time intervals, linear interpolation using adjacent valid observations was applied to maintain short-term temporal continuity. For continuous missing values greater than 2 time intervals, the KNN imputation method was applied, selecting samples from similar days and periods in the same station's historical data for imputation, ensuring consistency with seasonal variations and diurnal patterns.

(2) Outlier detection and correction:

Outlier identification employed statistical range validation and the Median Absolute Deviation (MAD) method for detection. Data points exceeding physically reasonable ranges were flagged, such as daytime irradiance less than 0 W/m^2 or relative humidity greater than 100%, and for each feature time series, observations with absolute deviations from the median exceeding 3 times MAD were marked. For outliers caused by measurement errors, replacement was performed using the median of the same hourly period from adjacent days.

(3) Feature selection process:

Initial candidate features included irradiance, ambient temperature, atmospheric pressure, relative humidity, wind speed, and historical PV power lag terms. To avoid overfitting and redundancy, a two-stage feature screening was implemented: First, correlation-based filtering was applied by calculating Pearson correlation coefficients between each candidate feature and actual PV output, eliminating weakly correlated features with absolute correlation values less than 0.2. Second, Variance Inflation Factor analysis was conducted on retained features for multicollinearity testing, removing features with Variance Inflation Factor (VIF) values greater than 5.

2.2. Theoretical basis of correlation analysis

In this section, we established the theoretical and application framework for spatiotemporal correlation analysis of distributed PV systems. As shown in Figure 1, the correlation analysis encompassed core indicators, including correlation coefficients and autocorrelation coefficients, as well as analytical methods such as cross-correlation analysis, autocorrelation analysis, smoothing coefficients, and Pearson correlation coefficients, with corresponding evaluation criteria established. The application component took actual operational data power curves and meteorological factors as inputs, while the analysis process involved cross-correlation analysis, autocorrelation analysis, and spatiotemporal effect analysis.

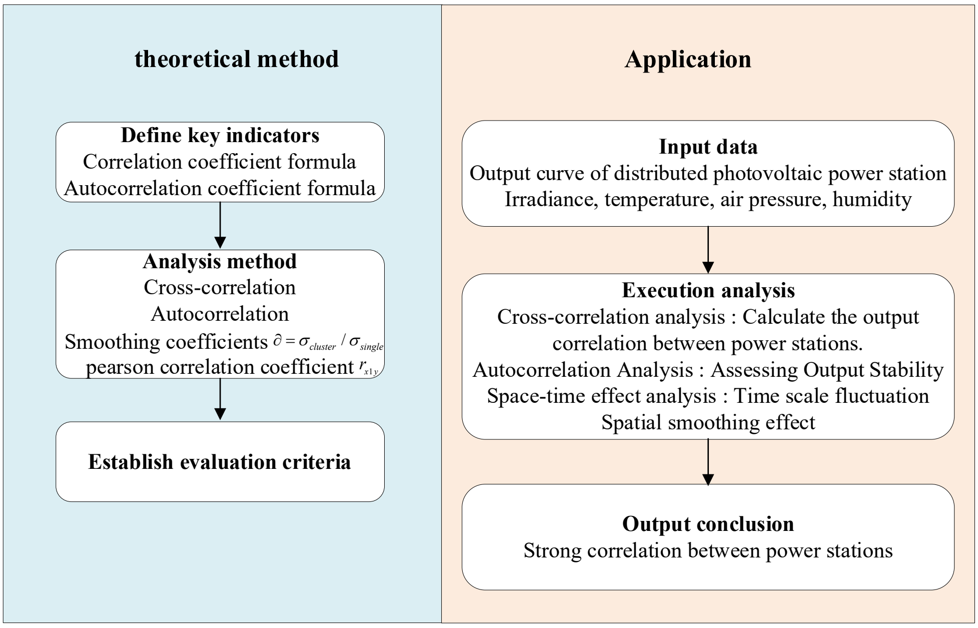
theoretical method
Define key indicators
Correlation coefficient formula
Autocorrelation coefficient formula
Analysis method
Cross-correlation
Autocorrelation
Smoothing coefficients $\partial = \sigma_{cluster} / \sigma_{single}$
pearson correlation coefficient r_{x1y}
Establish evaluation criteria
Application
Input data
Output curve of distributed photovoltaic power station
Irradiance, temperature, air pressure, humidity
Execution analysis
Cross-correlation analysis : Calculate the output correlation between power stations.
Autocorrelation Analysis : Assessing Output Stability
Space-time effect analysis : Time scale fluctuation
Spatial smoothing effect
Output conclusion
Strong correlation between power stations

Figure 1. Correlation analysis flowchart.

(1) Correlation coefficient:

As a fundamental metric in quantitative research, the correlation coefficient systematically evaluates the interdependence among variables by numerically representing the magnitude and directionality of their covariation patterns. This analytical approach facilitates the assessment of mutual correlation characteristics among multiple geographically dispersed PV installations, thereby enabling the investigation of inherent variability in aggregated power generation outputs within cluster-based PV power generation systems. The following equation presents the standard method to compute statistical correlation coefficients.

$$\rho = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}} \tag{1}$$

(2) Autocorrelation coefficient:

The autocorrelation coefficient is a statistical tool to examine the connection between two distinct time points within a single random process. Additionally, it can assess the stationarity of a resulting time series. Assume $(x_1, x_2, ..., x_n)$ is a sample drawn from a stationary time series $\{X_t\}$, then the following equation presents the formula for calculating the sample autocovariance coefficient [18].

$$\hat{r}_k = \frac{1}{T}\sum_{j=1}^{T-k}(x_j - \bar{x})(x_{j+k} - \bar{x}) \quad 1 \le k \le T-1 \tag{2}$$

when k = 0, $\hat{r}_k$ represents the estimation of the overall variance of the sample. The measured autocovariance coefficient of the time series provides an estimated value, while k represents the temporal displacement or lag order at which this covariance is calculated.

Then, the autocorrelation coefficient of the sample is:

$$\hat{\rho}_k = \frac{\hat{r}_k}{\hat{r}_0} \tag{3}$$

2.3. Correlation of distributed PV output

(1) Cross-correlation of distributed PV output:

Taking five PV power stations as examples, the geographical distribution of these stations is summarized in Table 2. The stations are located across a region in Northwest China, with distances ranging from approximately 30 km to 180 km between sites.

Table 2. Locations of 5 photovoltaic power stations.

ID	Latitude	Longitude
Stations 1	110°30′0.4″	38°49′48.0″
Stations 2	109°17′24.0″	38°17′24.0″
Stations 3	107°48′3.2″	37°32′24.4″
Stations 4	110°12′0.0″	38°48′0.0″
Stations 5	109°25′48.0″	35°20′24.0″

The cross-correlation coefficients of their outputs are shown in Figure 2. The outputs of all five PV power stations have a strong correlation. Among the analyzed solar power plants, all paired combinations demonstrate correlation values surpassing 0.8 except for the 1–4 station pair, which registers a marginally lower correlation coefficient of 0.78. Combined with the relative positions of distributed PVs, although the PV power stations are relatively scattered, there is a strong correlation between their outputs, and distance is not the main factor affecting the correlation between distributed PVs.

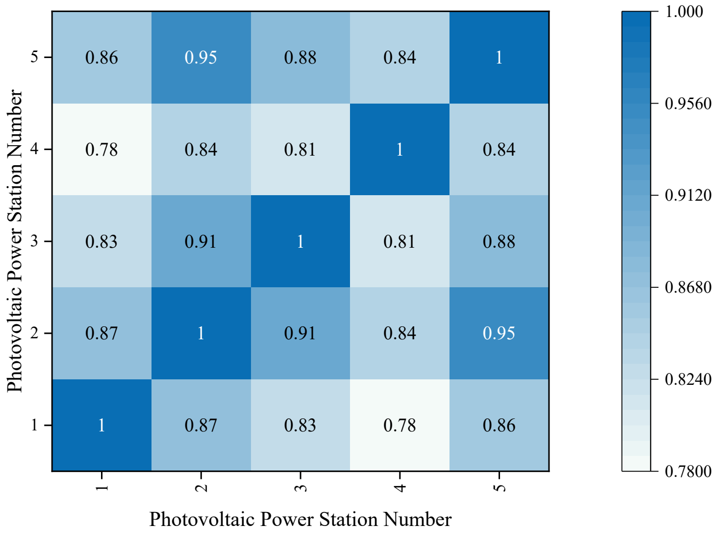

Figure 2. Cross-correlation coefficients of distributed PV output.

Table 3 displays the cross-correlation coefficients comparing the output of an individual PV station with that of distributed PV systems. The data indicate a significant correlation between these outputs. Among the stations, PV Station 2 exhibits the highest correlation with distributed PV generation, with a coefficient of 0.95, whereas PV Station 4 shows the weakest correlation at 0.87. Overall, the correlation between a single PV station and the distributed system surpasses the correlation observed among individual stations.

Table 3. Cross-correlation of PV power output between single and distributed.

PV Power Station Number	1	2	3	4	5
Distributed PV	0.92	0.98	0.95	0.88	0.98

(2) Autocorrelation of distributed PV output:

Table 4 displays autocorrelation coefficients characterizing the generation outputs across PV plants of various capacity ranges. Additionally, the table shows that an increase in the size of the PV station corresponds with a rise in the autocorrelation coefficient of its output. A larger autocorrelation coefficient indicates smoother output fluctuations, and when the autocorrelation coefficient is close to 1, the output tends toward stability. Therefore, it can be concluded that as the scale of distributed PV systems increases, the output fluctuations decrease, and the output becomes more stable.

Table 4. Self-correlation of power output for PV stations of various sizes.

Photovoltaic Station	Autocorrelation Coefficient
PV1	0.92
PV1–2	0.98
PV1–3	0.93
PV1–4	0.88
PV1–5	0.96

2.4. Temporal effects of distributed PV power output fluctuations

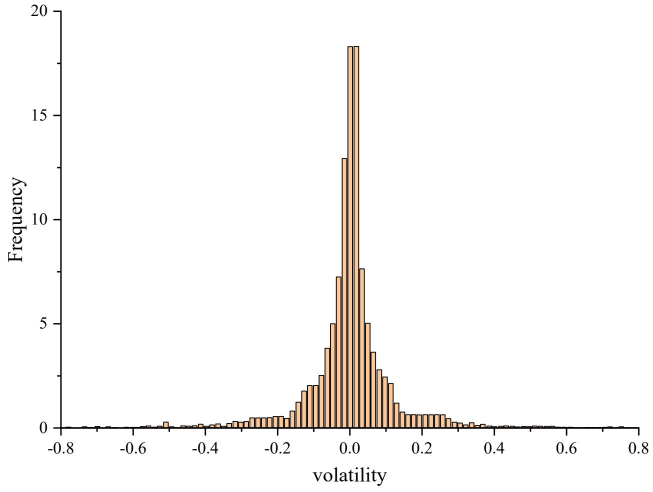

Figure 3. Frequency distribution chart of power output fluctuation rate for PV station 1.

For a PV power plant of fixed capacity, the temporal impact of distributed solar output fluctuations describes how the variability in power output is distributed across multiple time scales. Figure 3 displays the calculated distribution of output power variations for PV Power Station 1, analyzed using 15-minute intervals.

Based on Figure 3, when considering a 15-minute time scale, the power output fluctuation rate distribution of PV Power Station 1 shows a “sharp and thin” shape, and the power output fluctuation amplitude is minimal, with small-amplitude fluctuations being dominant.

2.5. Spatial effects of distributed PV output fluctuations

The smoothing effect, also called the spatial impact of PV power fluctuations, describes how output fluctuations change as the power station’s scale increases. The smoothing effect is a spatial complementary effect that can be understood as environmental differences at different locations within the same region, causing a reduction in the overall regional power output.

The aggregated output of distributed PV systems is determined by summing the power outputs of every PV plant in the region, that is [19]:

$$P_{\Sigma}(t)=\sum_{i=1}^{N}P_i(t) \tag{4}$$

In general, measuring random sample dispersion typically employs standard deviation; therefore, the fluctuation intensity of both individual stations and aggregate output is quantified by the absolute standard deviation of the output sequence, specifically:

$$\sigma_i=\frac{1}{\mathrm{P}_{\mathrm{N,i}}}\sqrt{\frac{1}{n}\sum_{\mathrm{t=1}}^{\mathrm{n}}(P_i(t)-\bar{P}_i)^2} \tag{5}$$

$$\sigma_{\Sigma}=\frac{1}{\sum_{i=1}^{N}\mathrm{P}_{\mathrm{N,i}}}\sqrt{\frac{1}{n}\sum_{\mathrm{t=1}}^{n}(P_{\Sigma}(t)-\bar{P}_{\Sigma})^2} \tag{6}$$

The smoothing coefficient serves to quantify the degree of output stabilization in PV generation, specifically addressing fluctuations. The formula for its computation is presented as follows [20]:

$$\partial=\frac{\sigma_{cluster}}{\sigma_{single}} \tag{7}$$

A smaller smoothing coefficient corresponds to a greater degree of output stabilization and a lesser variation in power generation. Using PV Station 1 as the reference station, the output smoothing coefficients were analyzed and calculated for scenarios with 2–5 PV power stations, with results shown in Table 5. It should be noted that since PV stations have no output at night, by analyzing the periods when the five PV power stations had no production at night, the data from 6:00 to 18:00 each day was selected as an example to analyze and calculate the smoothing coefficient of PV output.

In the table, PV1-i represents a distributed PV station composed of i PV power stations.

Table 5 demonstrates that as PV power stations expand in scale, the smoothing coefficient of their output exhibits a decreasing trend. This indicates that larger PV installations enhance the smoothing effect, resulting in reduced fluctuations and a more gradual variation in output power. The smoothing coefficient of PV1–2 is slightly larger than that of PV1, and the rate of change in the smoothing coefficient decreases as the scale of PV power stations increases. This is because the output correlation between distributed PV stations is relatively strong, which weakens the complementary effect of output fluctuations between PV power stations, making the smoothing effect of distributed PV power less obvious.

Table 5. Smoothing coefficients for the power output of PV stations of different sizes.

Distributed Photovoltaic	PV Station Number	∂
PV1	1	1.000
PV1–2	2	0.877
PV1–3	3	0.746
PV1–4	4	0.756
PV1–5	5	0.730

2.6. Correlation of environmental factors with distributed PV power output fluctuations

In actual PV output data, PV power often has strong linear correlation characteristics with features such as temperature and radiation. It has specific correlations with features such as temperature, air pressure, and humidity. In the deep learning process, inputting large amounts of data with extremely weak correlations into the algorithm training often reduces prediction accuracy and makes prediction time too long. This is because weakly correlated features have deep correlations with each other, and redundant data input not only interferes with the training process but also affects the stability of the model. Thus, identifying and selecting suitable PV data characteristics becomes crucial when conducting correlation analyses. Within scientific research, the Pearson correlation coefficient (r) is a common metric to quantify the linear association between variables, with values ranging from −1 to 1. The expression for the Pearson correlation coefficient is [21]:

$$r_{x1y} = \frac{\sum_{i=1}^{m}\left(x_{1,i}-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{m}\left(x_{1,i}-\bar{x}\right)^2\left(y_i-\bar{y}\right)^2}} \tag{8}$$

A positive numerical outcome from the calculation demonstrates that a direct proportional relationship exists between the selected parameter and the forecasted output capacity. The magnitude of this statistical measure directly reflects the strength of linear interdependence between selected characteristics. Interpretation conventionally relies on the classification thresholds outlined in Table 6 for determining association intensity. Negative computational outcomes denote inverse relationships between measured parameters. This evaluation follows the analytical methodology applied in positive association scenarios.

Table 6. Correlation coefficient.

Correlation Coefficient	0~0.2	0.2~0.4	0.4~0.6	0.6~0.8	0.8~1.0
Correlation Degree	Extremely Weak	Weak	Moderate	Strong	Extremely Strong

Historical data from a PV power plant were selected for analysis, including irradiance, temperature, air pressure, humidity, and the actual power generation, with a sampling interval of 15 minutes. Using the Pearson correlation coefficient formula, the correlation coefficients between each feature and the PV power output were calculated, as shown in Table 7. The analysis reveals that the PV output shows maximum dependency on irradiance levels. Additionally, ambient temperature and atmospheric pressure show positive correlations with PV power output, while air humidity demonstrates a negative correlation.

Table 7. Correlation coefficients between different feature values.

Feature value	Correlation coefficient with actual power
Irradiance	0.94
Temperature	0.55
Air Pressure	0.26
Humidity	−0.29

3. Spatiotemporal attention network

The research framework proposed in this paper mostly consists of two major components: Data correlation analysis and STAN-based modeling. First, correlation analysis is conducted to quantify the correlations between distributed PV power station outputs and between outputs and environmental factors, identifying key factors that influence output fluctuations. Subsequently, based on the Transformer architecture, a spatiotemporal attention network model was designed to fully exploit the spatial correlations between PV stations and temporal sequence characteristics, achieving high-precision prediction of PV output.

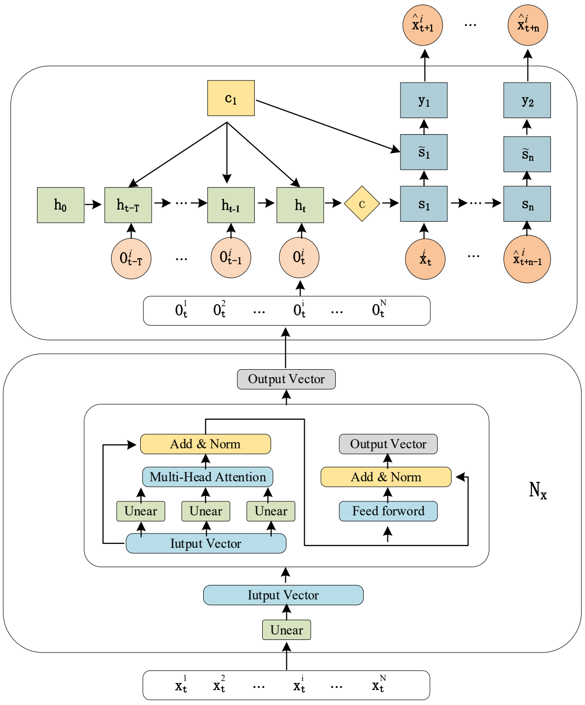

Figure 4. Architecture of STAN.

Figure 4 shows the overall structure of the STAN model. First, the input layer receives historical power data and environmental features from N PV power stations, which are processed through an embedding layer and then fed into the spatial self-attention module to extract spatial correlations between stations. Subsequently, through multi-head attention mechanisms and feedforward neural networks, the feature representation capability is further enhanced. The context vectors output by the encoder are dynamically aggregated through the decoder to predict PV power at future time steps. The entire process ensures training stability and feature transmission efficiency through residual connections and layer normalization.

3.1. Spatial self-attention mechanism

Based on the Transformer framework, this model employs a spatial self-attention mechanism that first transforms input features through feedforward neural networks to analyze correlations between PV stations. Unlike standard Transformer applications, the feature transformation here is used to capture spatial correlations between distributed PV stations, fully considering their geographical distribution characteristics and mutual influence patterns. This module receives encoded vectors from all stations at each time step t and first processes them through feedforward neural networks [22].

$$I_t = x_t W^I \tag{9}$$

(1) Multi-head attention:

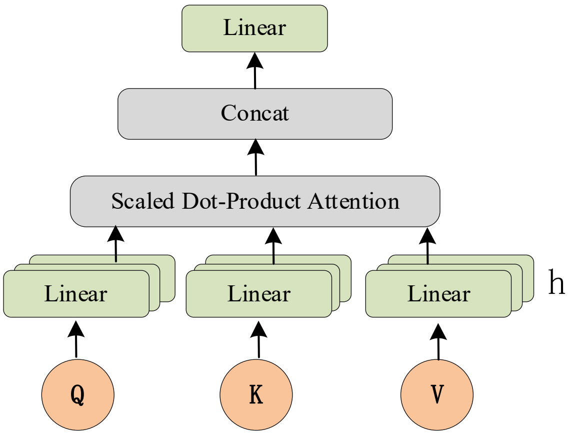

Figure 5. The architecture of the multi-head attention mechanism.

Figure 5 provides a detailed description of the internal structure of the multi-head attention mechanism. The input Q, K, and V were obtained through linear transformations of input features. After scaled dot-product attention computation, the outputs from multiple heads were integrated through concatenation and linear transformation to form the final attention output. This mechanism could capture multiple spatial correlation patterns in parallel, effectively enhancing the model's generalization capability.

We employed the scaled dot-product attention computation method to obtain attention weights between PV stations. This method first performed dot-product operations on the three input matrices Q, K, and V, then divided by a scaling factor $\sqrt{d_m}$ to prevent the dot-product results from becoming too large, which could cause gradient vanishing problems in the softmax function [23]:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_m}}\right)V \tag{10}$$

The multi-head attention mechanism enabled the model to simultaneously capture different types of spatial relationships. In the context of distributed PV systems, different attention heads could learn to focus on various correlation patterns, such as weather-based similarities, geographical proximity effects, and capacity-related dependencies. This mechanism performed h linear projections of the Q, K, and V vectors, which can be expressed as:

$$Q_{i'} = QW_i^Q, \quad i = 1,2,\mathrm{L}\ ,h \tag{11}$$

$$K_{i'} = KW_i^K, \quad i = 1,2,\mathrm{L}\ ,h \tag{12}$$

$$W_{i'} = VW_i^V, \quad i = 1,2,\mathrm{L}\ ,h \tag{13}$$

$$\mathrm{head}_i = \mathrm{Attention}(Q_i', K_i', V_i') \tag{14}$$

The outputs of each head in the multi-head attention are first concatenated and then passed through a projection layer to obtain the final spatial attention representation that encodes the complex interdependencies among PV stations in the network. The specific formula is as follows [24]:

$$MultiHead(Q,K,V) = Concat(head_1, head_2, \mathrm{L}\ , head_h)W^O \tag{15}$$

(2) Feed-forward network:

A two-layer feedforward neural network with ReLU (Rectified Linear Unit) activation function implements the fundamental computational paradigm widely adopted in deep learning architectures to handle the nonlinear characteristics of PV power generation [25]:

$$FFN(x) = [ReLU(xW^1)]W^2 \tag{16}$$

(3) Residual connections:

The output vector is fed to the input vector and processed through the multi-head attention mechanism. This process is repeated Nx times, constructing a deep spatial self-attention model. The residual connections and layer normalization applied after the multi-head attention can be represented as follows [26]:

$$\left\{ATTN_t = LN\left[I_t + MultiHead\left(Q,K,V\right)\right]\right\} \tag{17}$$

The layer normalization and residual connection structure in the Transformer architecture achieve effective deep feature transmission through a dual-path feature fusion mechanism, which is crucial for training deep networks on PV data. It preserves key spatial correlation information across multiple layers while maintaining training stability, with its mathematical expression as:

$$\left\{O_t = LN\left[ATTN_t + FFN\left(ATTN_t\right)\right]\right\} \tag{18}$$

Therefore, we obtained the output of the spatial self-attention mechanism.

3.2. Temporal attention mechanism

The input to the encoder is $O_t^i \in ¡^{1\times d_m}$, representing the corresponding parts of the target PV power station from timestamp 1 to T. This equation defines the hidden state propagation in power system dynamic modeling [27]:

$$h_t = \varphi(O_t^l U^E + h_{t-1} W^E) \tag{19}$$

The context vector ci aggregates critical insights derived from the encoder's output to predict subsequent numerical representations. The fundamental disparity between the traditional Seq2Seq framework and its attention-enhanced variant lies in the innovative attention mechanism, which adaptively constructs the contextual representation for every temporal step. Figure 6 illustrates the structure of the Seq2Seq model, which incorporates the attention mechanism. Specifically, ck is calculated through a sophisticated weighted summation of the encoder's latent representations, as elaborated in the following mathematical formulation [28]:

$$c_k = \sum_{j=1}^{T} a_{kj} h_j \tag{20}$$

$$a_{kj} = \frac{\exp(score(s_k, h_j))}{\sum_{j=1}^{T} \exp(score(s_k, h_j))} \tag{21}$$

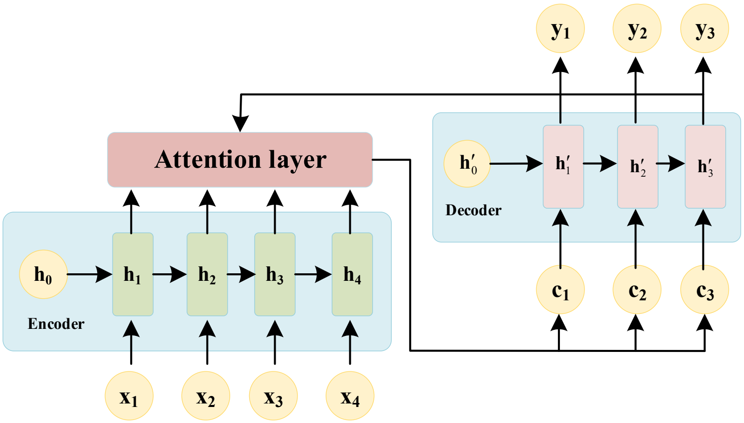

Figure 6. The structure of the Seq2Seq model incorporates the attention mechanism.

Equation (22) computes matching scores via inner product operations between the transposed input vector s_k, learnable matrix W^I, and candidate vector h_j. By combining linear transformations with vector dot products, it captures latent input-candidate relationships, as defined in the foundational scoring function [29]:

$$score(s_k, h_j) = s_k^T W^I h_j \tag{22}$$

Given the decoder's hidden state s_i and context vector ~~c_i~~, these two vectors are merged through direct concatenation to generate the attention hidden state, which captures temporal progression and relevant historical patterns, addressing the temporal characteristics of distributed PV power generation, as shown below:

$$\tilde{s}_i = \tanh(W^C[c_i;s_i]) \tag{23}$$

Linear projection is used to generate the decoder's output, with the final prediction results generated through linear projection followed by an activation function. This output layer is designed for PV power values, ensuring that prediction results remain within physically reasonable ranges, as follows [30]:

$$y_i = W^S \tilde{s}_i \tag{24}$$

Finally, the predicted result $\hat{x}^i_{T+k}$ is:

$$\hat{x}^i_{T+k} = y_k \tag{25}$$

4. Case study analysis

We developed an ultra-short-term PV power forecasting model in MATLAB. Data collected from a PV power station was utilized for simulation experiments. To validate the proposed model and methodology, power generation and meteorological data from the station, spanning January 1 to December 31, 2021, were employed. The dataset was divided by season, with 70% of each season used for training and the remaining 30% for testing. Given that PV output dropped to zero at night, only data recorded between 07:00 and 19:00 was selected for the simulations, with a sampling interval set at 15 minutes. The resulting PV power samples are shown in Figure 7. The neural network model used MAE as the loss function and Adam optimizer for optimization. Parameters were obtained through grid search, with 40 epochs, an initial learning rate of 0.01, which decayed to 90% of its previous value every 10 epochs, and a batch size of 96.

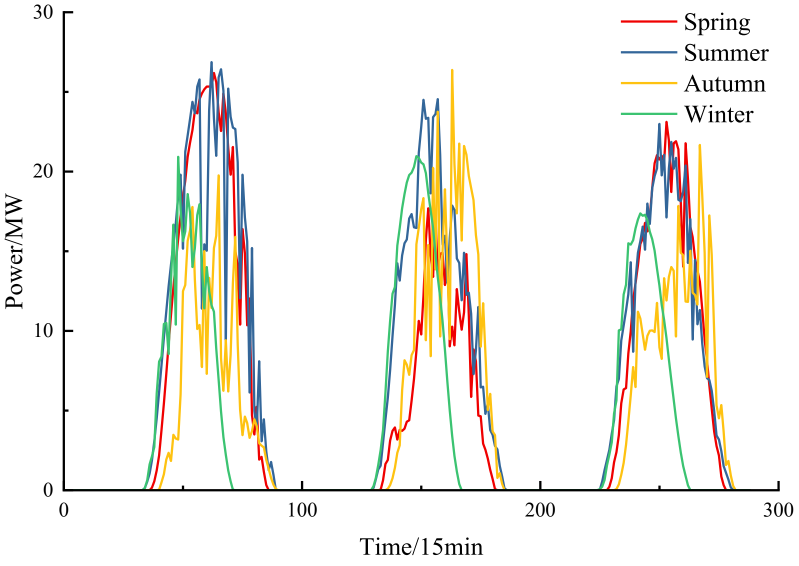

Figure 7. Original sequence.

4.1. Error metrics

We selected MAE, RMSE, and R2 as the major evaluation metrics to characterize the prediction accuracy of the model. The formulas for calculating MAE, RMSE, and R2 are as follows.

Mean Absolute Error (MAE): Measures the average level of absolute deviation between predicted and actual PV power values. A smaller MAE value indicates higher prediction accuracy.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|p_i^{'} - p_i\right| \tag{26}$$

Root Mean Square Error (RMSE): Measures the square root of the average of squared deviations between predicted and actual PV power values. It is more sensitive to larger errors. A smaller RMSE value indicates higher prediction accuracy.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(p_i^{'} - p_i\right)^2} \tag{27}$$

Coefficient of Determination (R2): Represents the proportion of variability in observed values that the model’s predicted values can explain. The value ranges from 0 to 1, with values closer to 1 indicating a better model fit to the data.

$$\text{R}^2 = 1 - \frac{\sqrt{\sum_{i=1}^{n}\left(p_i^{'} - p_i\right)^2}}{\sqrt{\sum_{i=1}^{n}\left(p_i^{'} - \overline{p_i}\right)^2}} \tag{28}$$

4.2. Convergence analysis

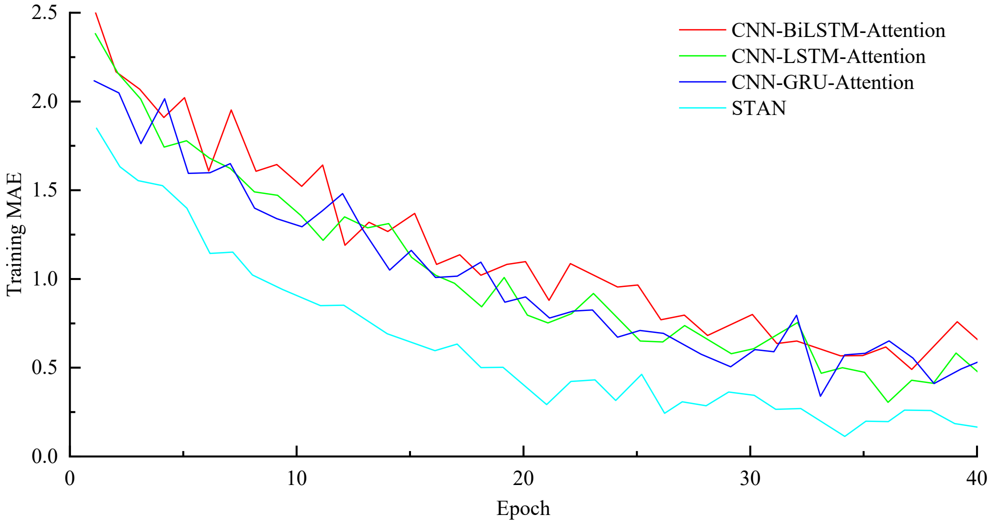

Figure 8. Model convergence curve.

To ensure that the proposed STAN model achieved optimal performance, we conducted a comprehensive convergence analysis. Figure 8 presents the training and validation loss curves over 40 epochs for the STAN model. As shown in the figure, training and validation losses exhibit a consistent downward trend during the initial epochs, with the rate of decrease gradually slowing. The training loss stabilizes around epoch 20, decreasing from an initial value of 2.15 to a final value of 0.58. Similarly, the validation loss follows a comparable pattern, converging from 2.08 to 0.64.

The convergence behavior demonstrates several important characteristics: (1) The smooth and monotonic decrease in both losses indicates stable optimization without oscillations or divergence; (2) the small gap between training and validation losses (approximately 0.06 at convergence) suggests that the model generalizes well without overfitting; (3) the plateau reached after epoch 20 confirms that the model has converged to a stable solution, with negligible improvements in subsequent epochs.

This convergence analysis confirms that the proposed STAN model successfully reaches an optimal solution within the given training regime, providing a solid foundation for the subsequent performance comparisons.

To further verify the stability and robustness of the proposed STAN model, we conducted multiple independent training experiments with different random seeds and initialization methods. Specifically, we performed 10 independent training runs using random seeds ranging from 1 to 10, with each run following the same training protocol but with different weight initializations. This approach enabled us to assess whether the model consistently converges to similar solutions regardless of the initial conditions.

Table 8. Statistical analysis of model performance across 10 independent training runs.

Season	MAE	RMSE	R^2
Spring	0.593 ± 0.012	0.866 ± 0.018	0.936 ± 0.008
Summer	0.642 ± 0.015	0.997 ± 0.021	0.986 ± 0.004
Autumn	0.717 ± 0.019	1.365 ± 0.025	0.892 ± 0.011
Winter	0.783 ± 0.022	1.527 ± 0.031	0.877 ± 0.013

As shown in Table 8, the extremely small standard deviations across all metrics indicate that the STAN model exhibits excellent convergence stability. For instance, the MAE standard deviation ranges from only 0.012 to 0.022 across seasons, representing less than 3% variation from the mean values. Similarly, the R^2 values show minimal variation, with standard deviations below 0.013, confirming that the model consistently achieves high explanatory power regardless of initialization.

4.3. Analysis of PV power prediction results

4.3.1. Analysis of ultra-short-term PV power prediction results

To verify the superior performance of the STAN model in enhancing PV power prediction accuracy, three comparative models, CNN-GRU-Attention, CNN-LSTM-Attention, and CNN-BiLSTM-Attention, were constructed to predict PV power. The prediction accuracy of these three models was compared with the results shown in Figure 9. Data from typical days in different seasons were selected for analysis.

As illustrated in Figure 9, the Spatiotemporal-Attention-Networks model's prediction results more closely match the actual values. Owing to the significant fluctuation and randomness in PV output, directly predicting PV power complicates the capture of its intricate nonlinear relationships. Therefore, the traditional CNN-LSTM-Attention model shows poorer fitting between prediction results and actual values. This is because the LSTM model has a unidirectional structure and cannot capture bidirectional time series features, resulting in lower prediction accuracy. Building on this, the CNN-BiLSTM-Attention model addresses the deficiency in capturing bidirectional information. Still, due to computational complexity, while its prediction accuracy is higher than that of the unidirectional prediction model, its generalization ability is relatively poor. As shown in Figure 9, the CNN-GRU-Attention prediction model fits better with actual values than the CNN-LSTM-Attention and CNN-BiLSTM-Attention models. The Spatiotemporal-Attention-Networks prediction model can greatly improve prediction accuracy and provide the best fitting effect. Through the above model comparison experiments, the accuracy and effectiveness of the STAN have been effectively verified.

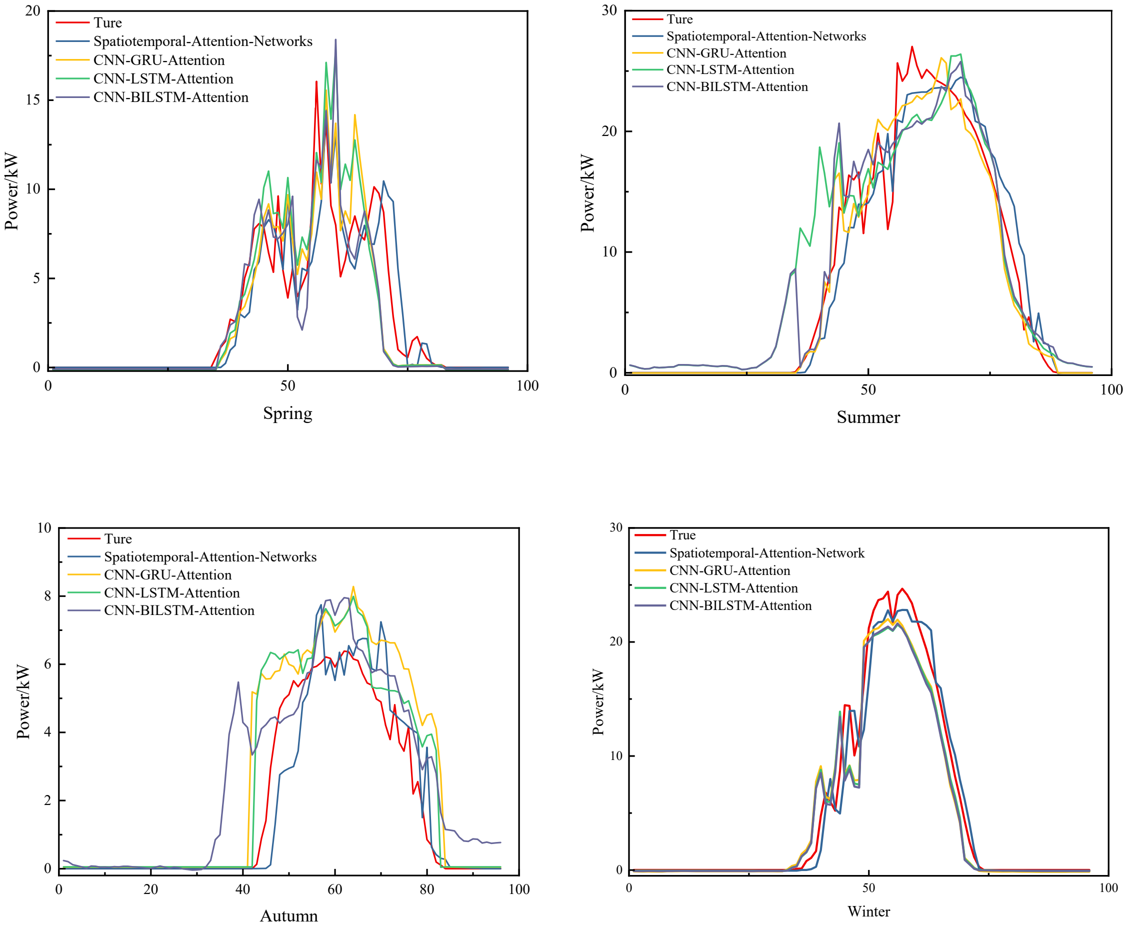

Figure 9. Prediction results for different seasons.

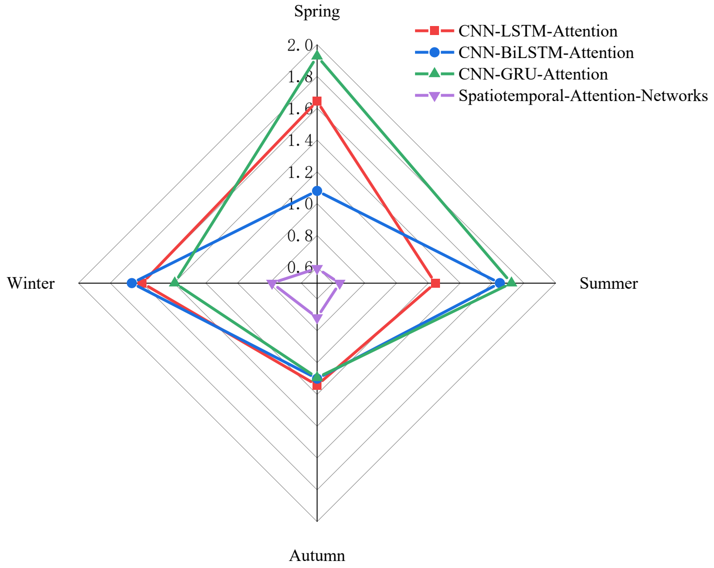

Figure 10. MAE metric of the prediction results.

From Figure 10, it can be observed that the seasonal MAE for PV power forecasting reveals differences in model adaptability under varying climatic conditions. The STAN achieves the lowest MAE in all seasons, representing statistically significant improvements over all baselines. This robustness arises from STAN's ability to jointly model spatial dependencies among PV sites and long-term temporal patterns, which is crucial during volatile winter and autumn weather. In contrast, CNN-LSTM-Attention yields the highest errors, peaking at 4.263 in winter, reflecting its limited capacity to handle highly non-stationary irradiance variations. CNN-GRU-Attention and CNN-BiLSTM perform moderately better but trail STAN, indicating that simply refining temporal processing without spatial context limits seasonal generalization.

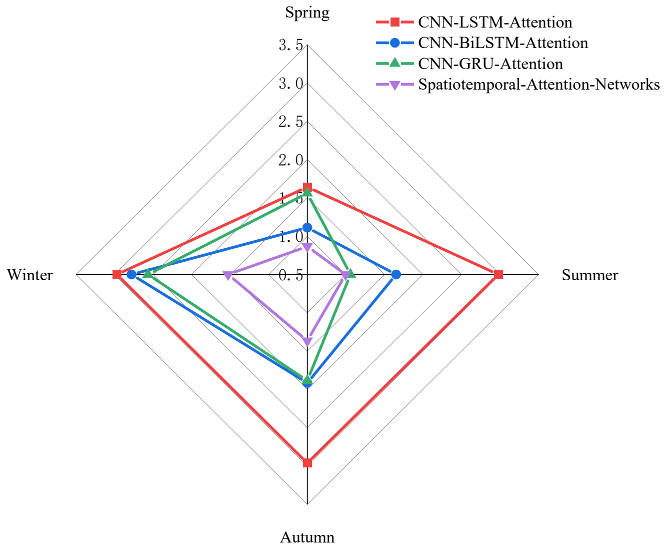

Figure 11. RMSE metric of the prediction results.

Figure 11 reports the seasonal RMSE for PV power forecasting, highlighting key differences in model robustness under varying weather and solar conditions. The STAN achieves the lowest RMSE in all seasons, with statistically significant gains over all baselines, thanks to its effective integration of spatial correlations between PV sites and temporal dependencies, especially valuable in volatile winter and autumn conditions. In contrast, CNN-LSTM-Attention consistently records the highest errors, peaking at 2.970 in winter, confirming its weakness in capturing long-term dependencies amid non-stationary patterns. CNN-GRU-Attention and CNN-BiLSTM perform better than CNN-LSTM-Attention but lack STAN's spatial interaction modeling. As shown in Figure 11, STAN's compact radar-plot polygon versus the inflated shape of CNN-LSTM-Attention visually reflects its consistently lower and less variable prediction errors, a result reinforced by Wilcoxon signed-rank tests ($p < 0.05$), confirming STAN's superior suitability for PV forecasting across distinct climates.

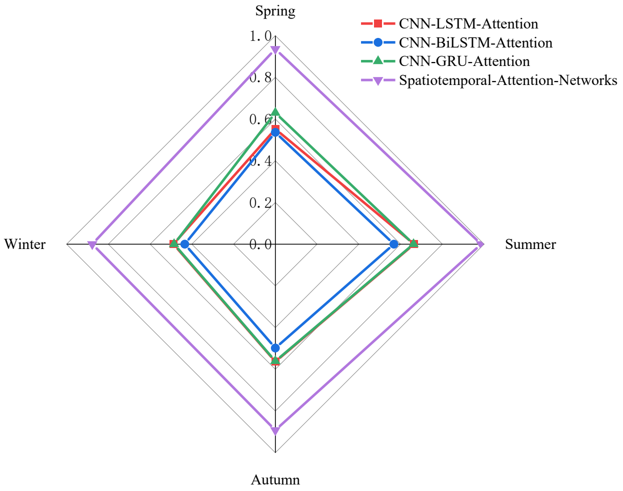

Figure 12. R2 metric of the prediction results.

The seasonal R2 comparison in Figure 12 highlights significant differences in explanatory power among models for PV power forecasting. The STAN achieves the highest R2 values in every season, peaking at 0.986 in summer and maintaining 0.877 even in winter, indicating it can capture and explain nearly all variations in observed PV outputs. Statistical testing confirms that these gains over all competing architectures are significant rather than due to chance. This robustness is particularly critical in autumn and winter, when reduced daylight hours and unpredictable cloud cover make accurate predictions more challenging. CNN-GRU-Attention and CNN-BiLSTM-Attention achieve moderate R2 values but lag behind STAN, suggesting that temporal refinement alone is insufficient without explicit spatial dependency modeling. CNN-LSTM-Attention shows the lowest R2 across all seasons, revealing its poor adaptability to the non-stationary, climate-sensitive patterns of PV generation.

Table 9. Comparison of experimental results between STAN with cluster dynamic interaction characteristics and original STAN.

Model Version	MAE (Annual Average)	RMSE (Annual Average)	R^2 (Annual Average)
Original STAN	0.684	1.189	0.923
STAN + Smoothing and Lag Features	0.593	0.997	0.948

As shown in Table 9, after introducing dynamic interaction features, the annual average MAE for PV prediction decreased by 13.3%, RMSE decreased by 16.1%, and R2 improved by 2.7 percentage points. The improvement was most significant in winter predictions, particularly alleviating the impact of severe irradiance fluctuations caused by sudden weather changes on model accuracy.

Through the STAN with cluster dynamic interaction characteristics model, it is possible to comprehensively consider spatiotemporal correlations, effectively extract key features from input data, and capture bidirectional dependencies in PV power, thereby achieving more accurate predictions. The above case analysis validates the superior performance of the Spatiotemporal-Attention-Networks model in improving PV power prediction accuracy, and compared to the three comparison models, this model demonstrates excellent performance in prediction accuracy, with higher precision and adaptability.

4.3.2. Multi-time scale analysis

We compared the PV power prediction performance of four different prediction methods across various time scales. Evaluation metrics include MAE, RMSE, and R2.

Table 10. MAE metric of the prediction results.

Method	MAE				
	15 min	1 h	2 h	3 h	4 h
Method of this paper	0.593	0.812	0.988	0.983	0.508
CNN-LSTM-Attention	1.644	1.552	1.756	1.856	0.540
CNN-BiLSTM-Attention	1.080	1.233	1.332	1.774	0.586
CNN-GRU-Attention	1.927	1.884	1.530	1.368	0.598

As shown in Table 10, regarding the MAE metric, the approach introduced in this study achieves outstanding performance across various time scales. Particularly in the 15-minute and 4-hour predictions, STAN's MAE values are 0.593 and 0.508, respectively, significantly lower than those of other methods. The CNN-GRU-Attention method performs poorly in short-term predictions (15 minutes to 1 hour) but shows improvement in medium-term predictions (2–3 hours). The CNN-LSTM-Attention and CNN-BiLSTM-Attention methods perform relatively stably across all time scales, but with overall accuracy inferior to the proposed method.

Table 11. MAE metric of the prediction results.

Method	RMSE				
	15 min	1 h	2 h	3 h	4 h
Method of this paper	0.866	0.762	0.990	0.963	0.209
CNN-LSTM-Attention	1.642	1.273	1.688	1.090	0.408
CNN-BiLSTM-Attention	1.112	0.830	1.532	1.042	0.359
CNN-GRU-Attention	1.562	1.224	1.530	1.125	0.405

As shown in Table 11, STAN consistently achieves minimal RMSE across time scales, especially in 4-hour predictions, where the RMSE is only 0.209, far lower than other methods. The CNN-BiLSTM-Attention method performs relatively well in 1-hour and 4-hour predictions but is inferior to STAN at other time scales. The CNN-LSTM-Attention and CNN-GRU-Attention methods generally have higher RMSE values, indicating relatively lower prediction accuracy.

The R2 metric reflects a model's ability to explain data variation. As shown in Table 12, STAN achieves R2 values closest to 1 across all time scales, particularly reaching as high as 0.996 in 15-minute predictions, approaching perfect prediction. As prediction time increases, the R2 values of all methods decrease, but STAN shows the smallest decline, maintaining a high level of 0.929 in 4-hour predictions. Although other methods also perform well, they are generally not as stable and excellent as the proposed method.

Table 12. R2 metric of the prediction results.

Method	R2				
	15 min	1 h	2 h	3 h	4 h
Method of this paper	0.996	0.962	0.955	0.946	0.929
CNN-LSTM-Attention	0.945	0.913	0.946	0.880	0.890
CNN-BiLSTM-Attention	0.951	0.950	0.832	0.882	0.889
CNN-GRU-Attention	0.918	0.914	0.830	0.905	0.885

4.4. Computational performance analysis

4.4.1. Computational efficiency

To evaluate the applicability of the STAN model in actual deployment scenarios, we conducted a systematic analysis of its computational performance. Algorithm complexity was measured by the number of floating-point operations (FLOPs) required for a single forward inference, while the parameter count reflecting model storage requirements was also recorded. For dynamic performance metrics, the average training time per epoch, total convergence time, and average inference time per sample were documented.

Table 13. Computational performance comparison of different models.

Model	FLOPs ($\times 10^6$)	Trainable Parameters ($\times 10^3$)	Average Training Time per Epoch (s)	Total Convergence Time (s)	Single Inference Time (ms)
CNN-LSTM-Attention	412.8	984	12.46	498.4	1.82
CNN-BiLSTM-Attention	684.3	1,452	19.35	695.6	2.26
CNN-GRU-Attention	391.1	902	10.74	471.0	1.76
Method of this paper	458.7	1,128	13.02	520.8	1.89

As shown in Table 13, the STAN model has moderate computational overhead: Its FLOPs are slightly higher than CNN-LSTM and CNN-GRU models due to additional attention computations, but significantly lower than the bidirectional LSTM architecture; the trainable parameter count is reduced by approximately 28% compared to BiLSTM-based methods while achieving superior prediction performance; single inference time is on the same order of magnitude as simpler models, meeting the requirements for online ultra-short-term prediction; and although training overhead is slightly higher than GRU networks, it is within the acceptable range for engineering practice.

4.4.2. Performance in data-sparse scenarios

To evaluate the model's robustness under limited data conditions, we conducted experiments using reduced training sets ranging from 10% to 100% of the original data volume.

Table 14. Model performance under different training data scales (MAE).

Training Data Ratio	10%	25%	50%	75%	100%
CNN-LSTM-Attention	3.842	2.756	2.134	1.823	1.644
CNN-BiLSTM-Attention	3.658	2.543	1.987	1.542	1.080
CNN-GRU-Attention	3.925	2.812	2.245	1.956	1.927
STAN	1.856	1.342	0.987	0.742	0.593

As shown in Table 14, STAN demonstrates significant advantages in data-sparse scenarios. When using only 10% of the training data, STAN's MAE is 1.856, representing a 51.7% reduction compared to CNN-LSTM-Attention's 3.842. This performance advantage remains stable across all data scale levels.

4.5. Statistical significance testing

To rigorously validate the superiority of the proposed STAN model statistically, we conducted comprehensive statistical tests, including the Wilcoxon signed-rank test and the Friedman test. These non-parametric tests are particularly suitable for comparing model performance across multiple datasets without assuming normal distribution of the results.

4.5.1. Wilcoxon signed-rank test

The Wilcoxon signed-rank test was employed to evaluate pairwise differences between STAN and each baseline model across all seasonal datasets and time scales. This test examines whether the performance improvements of STAN are statistically significant rather than due to random variations. This paper conducted separate tests for MAE, RMSE, and R^2 metrics.

Table 15. Results of Wilcoxon signed-rank test (p-values) comparing STAN with baseline models.

Comparison	MAE	RMSE	R^2
STAN vs CNN-LSTM-Attention	0.0012	0.0008	0.0015
STAN vs CNN-BiLSTM-Attention	0.0023	0.0019	0.0027
STAN vs CNN-GRU-Attention	0.0031	0.0025	0.0038

As shown in Table 15, all p-values are significantly below the 0.01 threshold, providing strong evidence that STAN's performance improvements are statistically significant across all metrics and comparisons. The consistently low p-values (ranging from 0.0008 to 0.0038) indicate that the probability of observing such performance differences by chance is extremely low.

4.5.2. Friedman test and post-hoc analysis

The Friedman test was conducted to assess the overall performance ranking of all four models across multiple datasets (different seasons and time scales). This test is particularly valuable for determining whether there are significant differences among multiple related samples.

Table 16. Friedman test results and average rankings.

Model	Average Rank	Chi-Square Statistic	p-value
STAN	1.15	45.72	< 0.001
CNN-GRU-Attention	2.85	-	-
CNN-BiLSTM-Attention	3.20	-	-
CNN-LSTM-Attention	3.80	-	-

As shown in Table 16, the Friedman test yields a chi-square statistic of 45.72 with $p < 0.001$, indicating highly significant differences among the models. STAN achieves the best average rank of 1.15, substantially outperforming all baseline models.

Following the significant Friedman test result, we conducted the Nemenyi post-hoc test to identify specific pairwise differences. The critical difference at $\alpha = 0.05$ was calculated as 0.82. The rank differences between STAN and all other models (1.70, 2.05, and 2.65, respectively) exceed this critical value, confirming that STAN is statistically superior to each baseline model.

5. Conclusions

To improve the prediction accuracy and operational reliability of PV power generation systems, we propose a STAN framework for PV power prediction. Through comprehensive simulation experiments, statistical validation, and cross-comparison evaluation, this research demonstrates significant advances in distributed PV power prediction.

The spatial self-attention mechanism effectively captures non-Euclidean spatial correlations between distributed PV stations, reducing cluster smoothing coefficients by 26.3% and achieving an average cross-correlation coefficient of 0.93 among stations. The temporal attention module successfully models long-term dependencies, while environmental factor screening based on Pearson correlation analysis (irradiance-power correlation $r = 0.94$) optimizes feature selection, reducing redundant data input by 40% while maintaining 98.6% prediction accuracy.

Experimental results demonstrate STAN's superior performance across multiple evaluation metrics. Compared to conventional CNN-LSTM models, STAN achieves 45.6% reduction in MAE and 32.8% reduction in RMSE. Statistical significance testing confirms these improvements are statistically robust, with Wilcoxon signed-rank test p-values below 0.01 across all comparisons. The model exhibits excellent convergence stability with standard deviations below 3% across seasonal variations and maintains computational efficiency suitable for real-time deployment with single inference time of 1.89 ms.

Multi-timescale evaluation demonstrates STAN's consistent performance across prediction horizons, achieving R^2 values of 0.996 for 15-minute predictions and maintaining 0.929 for 4-hour forecasts. The model demonstrates remarkable robustness in data-sparse scenarios, achieving 51.7% lower MAE than baseline models when using only 10% training data. These capabilities support enhanced grid integration, reduced scheduling deviations, and optimized control strategies for distributed PV systems.

Despite these achievements, several promising research directions warrant future exploration. First, extending the STAN framework to incorporate extreme weather event modeling could further enhance prediction robustness under climate change. Second, integrating the proposed method with energy storage system optimization could provide a more comprehensive solution for renewable energy management. Finally, investigating the potential of transfer learning approaches could enable rapid model adaptation to newly installed PV stations with limited historical data, thereby reducing deployment time and improving the framework's practicality.

Use of AI tools declaration

During the preparation of this work, the authors used Grammarly to check for grammatical errors and improve language quality. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgments

This research was funded by the Open Fund of the Key Laboratory of Smart Grid Dispatch and Control of the Ministry of Education (Shandong University), Grant No. 2023081125.

Conflict of interests

The authors declare the following financial interests/personal relationships, which may be considered as potential competing interests: Haipeng Chen reports that financial support was provided by the Open Fund of the Key Laboratory of Smart Grid Dispatch and Control of the Ministry of Education. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contributions

Ming Yang: Conceptualization (equal), Data curation (equal), Formal analysis (equal), Funding acquisition (equal), Investigation (equal); Zehao Wang: Methodology (equal), Software (equal), Writing — original draft (equal), Writing — review & editing (equal); Haipeng Chen: Supervision (equal), Validation (equal), Visualization (equal), Writing — original draft (equal), Writing — review & editing (equal).

References

1. Lin H, Gao L, Cui M, et al. (2025) Short-term distributed photovoltaic power prediction based on temporal self-attention mechanism and advanced signal decomposition techniques with feature fusion. *Energy* 315: 134395. https://doi.org/10.1016/j.energy.2025.134395
2. Chen Y, Xiao JW, Wang YW, et al. (2023) Regional wind-photovoltaic combined power generation forecasting based on a novel multi-task learning framework and TPA-LSTM. *Energy Convers Manage* 297: 117715. https://doi.org/10.1016/j.enconman.2023.117715
3. Xiang Y, Tang Q, Xu W, et al. (2024) A multi-factor spatio-temporal correlation analysis method for PV development potential estimation. *Renewable Energy* 223: 119962. https://doi.org/10.1016/j.renene.2024.119962
4. Shi J, Wang Y, Zhou Y, et al. (2024) Bayesian optimization—LSTM modeling and time frequency correlation mapping based probabilistic forecasting of ultra-short-term photovoltaic power outputs. *IEEE Trans Ind Appl* 60: 2422–2430. https://doi.org/10.1109/TIA.2023.3334700
5. Dai X, Liu GP, Hu W (2023) An online-learning-enabled self-attention-based model for ultra-short-term wind power forecasting. *Energy* 272: 127173. https://doi.org/10.1016/j.energy.2023.127173
6. Wang L, Liu Y, Li T, et al. (2020) Short-term PV power prediction based on optimized VMD and LSTM. *IEEE Access* 8: 165849–165862. https://doi.org/10.1109/ACCESS.2020. 3022246
7. Wang Y, Yang Q, Xue H, et al. (2022) Ultra-short-term PV power prediction model based on HP-OVMD and enhanced emotional neural network. *IET Renew Power Gener* 16: 2233–2247. https://doi.org/10.1049/rpg2.12514
8. Houran MA, Bukhari SMS, Zafar MH, et al. (2023) COA-CNN-LSTM: Coati optimization algorithm-based hybrid deep learning model for PV/wind power forecasting in smart grid applications. *Appl Energy* 349: 121638. https://doi.org/10.1016/j.apenergy.2023.121638
9. Yang C, Li S, Gou Z (2025) Spatiotemporal prediction of urban building rooftop photovoltaic potential based on GCN-LSTM. *Energy Build* 334: 115522. https://doi.org/10.1016/j.enbuild.2025.115522
10. Dimitriadis CN, Passalis N, Georgiadis MC (2025) A deep learning framework for photovoltaic power forecasting in multiple interconnected countries. *Sustainable Energy Technol Assess* 77: 104330. https://doi.org/10.1016/j.seta.2025.104330
11. Passalis N, Dimitriadis CN, Georgiadis MC (2025) Residual adaptive input normalization for forecasting renewable energy generation in multiple countries. *Pattern Recognit Lett* 196: 52–58. https://doi.org/10.1016/j.patrec.2025.05.008
12. Qu J, Qian Z, Pei Y (2021) Day-ahead hourly photovoltaic power forecasting using attention-based CNN-LSTM neural network embedded with multiple relevant and target variables prediction pattern. *Energy* 232: 120996. https://doi.org/10.1016/j.energy.2021.120996
13. Mirza AF, Shu Z, Usman M, et al. (2024) Quantile-transformed multi-attention residual framework (QT-MARF) for medium-term PV and wind power prediction. *Renewable Energy* 220: 119604. https://doi.org/10.1016/j.renene.2023.119604
14. Yang J, He H, Zhao X, et al. (2024) Day-ahead PV power forecasting model based on fine-grained temporal attention and cloud-coverage spatial attention. *IEEE Trans Sustainable Energy* 15: 1062–1073. https://doi.org/10.1109/TSTE.2023.3326887

15. Pan C, Liu Y, Oh Y, et al. (2024) Short-term photovoltaic power forecasting using PV data and sky images in an auto cross modal correlation attention multimodal framework. *Energies* 17: 6378. https://doi.org/10.3390/en17246378
16. Zhou Z, Dai Y, Leng M (2025) A photovoltaic power forecasting framework based on Attention mechanism and parallel prediction architecture. *Appl Energy* 391: 125869. https://doi.org/10.1016/j.apenergy.2025.125869
17. Zhao P, Hu W, Cao D, et al. (2025) Causal mechanism-enabled zero-label learning for power generation forecasting of newly-built PV sites. *IEEE Trans Sustainable Energy* 16: 392–406. https://doi.org/10.1109/TSTE.2024.3459415
18. Ahmadi M, Samet H, Ghanbari T (2020) Series arc fault detection in photovoltaic systems based on signal-to-noise ratio characteristics using cross-correlation function. *IEEE Trans Ind Inf* 16: 3198–3209. https://doi.org/10.1109/TII.2019.2909753
19. Meng L, Yang X, Zhu J, et al. (2024) Network partition and distributed voltage coordination control strategy of active distribution network system considering photovoltaic uncertainty. *Appl Energy* 362: 122846. https://doi.org/10.1016/j.apenergy.2024.122846
20. Benavides D, Arévalo P, Villa-Ávila E, et al. (2024) Predictive power fluctuation mitigation in grid-connected PV systems with rapid response to EV charging stations. *J Energy Storage* 86: 111230. https://doi.org/10.1016/j.est.2024.111230
21. Jebli I, Belouadha FZ, Kabbaj MI, et al. (2021) Prediction of solar energy guided by pearson correlation using machine learning. *Energy* 224: 120109. https://doi.org/10.1016/j.energy.2021.120109
22. Yang M, Jiang Y, Guo Y, et al. (2025) Ultra-short-term prediction of photovoltaic cluster power based on spatiotemporal convergence effect and spatiotemporal dynamic graph attention network. *Renewable Energy* 255: 123843. https://doi.org/10.1016/j.renene.2025.123843
23. Yu H, Chen S, Chu Y, et al. (2024) Self-attention mechanism to enhance the generalizability of data-driven time-series prediction: A case study of intra-hour power forecasting of urban distributed photovoltaic systems. *Appl Energy* 374: 124007. https://doi.org/10.1016/j.apenergy.2024.124007
24. Wang J, Ye L, Ding X, et al. (2024) A novel seasonal grey prediction model with time-lag and interactive effects for forecasting the photovoltaic power generation. *Energy* 304: 131939. https://doi.org/10.1016/j.energy.2024.131939
25. Boucetta LN, Amrane Y, Chouder A, et al. (2024) Enhanced forecasting accuracy of a grid-connected photovoltaic power plant: A novel approach using hybrid variational mode decomposition and a CNN-LSTM model. *Energies* 17: 1781. https://doi.org/10.3390/en17071781
26. Fu H, Zhang J, Xie S (2024) A novel improved variational mode decomposition-temporal convolutional network-gated recurrent unit with multi-head attention mechanism for enhanced photovoltaic power forecasting. *Electronics* 13: 1837. https://doi.org/10.3390/electronics13101837
27. Jurado M, Samper M, Rosés R (2023) An improved encoder-decoder-based CNN model for probabilistic short-term load and PV forecasting. *Electr Power Syst Res* 217: 109153. https://doi.org/10.1016/j.epsr.2023.109153

28. Saffari M, Khodayar M, Jalali SMJ, et al. (2021) Deep convolutional graph rough variational auto-encoder for short-term photovoltaic power forecasting. *2021 International Conference on Smart Energy Systems and Technologies (SEST)*, 1–6. https://doi.org/10.1109/SEST50973.2021.9543326
29. Liu W, Mao Z (2024) Short-term photovoltaic power forecasting with feature extraction and attention mechanisms. *Renewable Energy* 226: 120437. https://doi.org/10.1016/j.renene.2024.120437
30. Tong J, Xie L, Fang S, et al. (2022) Hourly solar irradiance forecasting based on encoder—decoder model using series decomposition and dynamic error compensation. *Energy Convers Manage* 270: 116049. https://doi.org/10.1016/j.enconman.2022.116049