

Article

Deep Learning for Solar Power Forecasting by Integrating Historical and Meteorological Data

Cheng He ¹, Siyuan Zhao ¹, Zhenshuo Guo ^{1,2}, Xun Li ³, Chuanyu Sun ² and Mingming Ge ^{1,3,*}¹ Sustainable Energy Technology and Management, Beijing Normal-Hong Kong Baptist University, Zhuhai 519087, China² School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin 150001, China³ Environmental Science, Guangdong Provincial Key Laboratory IRADS, Beijing Normal-Hong Kong Baptist University, Zhuhai 519087, China

* Correspondence: mmge@vt.edu

Abstract

With the global shift toward green energy, solar photovoltaic (PV) power has expanded rapidly. However, the unpredictable nature of PV generation, caused by changing weather conditions like irradiance and temperature, challenges grid stability and power scheduling. Therefore, developing smart forecasting models for high-precision PV power prediction is essential for modern grid management. This paper introduces an optimized forecasting framework using Long Short-Term Memory (LSTM) networks. By integrating historical power generation data with localized meteorological factors, a multivariate predictive model was developed and validated using empirical data from a 100 MW PV plant. Based on Pearson correlation analysis, four key features—temperature, direct normal irradiance (DNI), relative humidity, and cloud cover—were chosen as the main drivers of PV output. A multivariate LSTM model was then trained and carefully tested using time series cross-validation. Results show the multi-feature architecture consistently outperforms and surpasses single-feature benchmarks. Specifically, the model achieved a peak R^2 of 0.9864 and minimum MAE of 1.4057 kW. Across four validation sets, R^2 remained stable (0.9755–0.9836), with most errors tightly bounded within ± 5 kW. These findings confirm the model's excellent accuracy and its value for improving power system dispatch and resource planning.

Keywords: solar power forecasting; multi-feature fusion; meteorological factors; deep learning; LSTM



Academic Editor: Stéphane Grieu

Received: 10 June 2026

Revised: 17 July 2026

Accepted: 19 July 2026

Published: 22 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The large-scale adoption of renewable energy has emerged as a critical global consensus for accelerating the energy transition and mitigating climate change, driving its rapid deployment worldwide. Solar energy, with its cleanliness and abundance, stands out among many energy options as one of the most ideal alternative energy sources. As reported by the International Energy Agency (IEA), global solar photovoltaic power generation is expected to triple by 2030, surpassing coal to become the world's largest source of electricity. Furthermore, China's photovoltaic manufacturing capacity is expected to make up over 80% of the world's total [1].

Figure 1 illustrates the annual new renewable energy capacity additions worldwide from 2015 to 2024 and China's share of them. The global annual installed capacity of renewable energy surged from about 162 GW in 2015 to about 660 GW in 2024, an increase

of more than four times, with the growth rate being particularly significant after 2020. Solar photovoltaic was the main driver of this expansion, with annual new installed capacity jumping from about 56 GW in 2015 to about 505 GW in 2024. Its share of new global installations soared from about 35% to about 76%. It is worth noting that since 2023, the annual newly installed capacity of photovoltaics has far exceeded the overall global increase in renewable energy installations in 2019. In 2024, China's annual PV installations are expected to account for more than two-thirds of the global total.

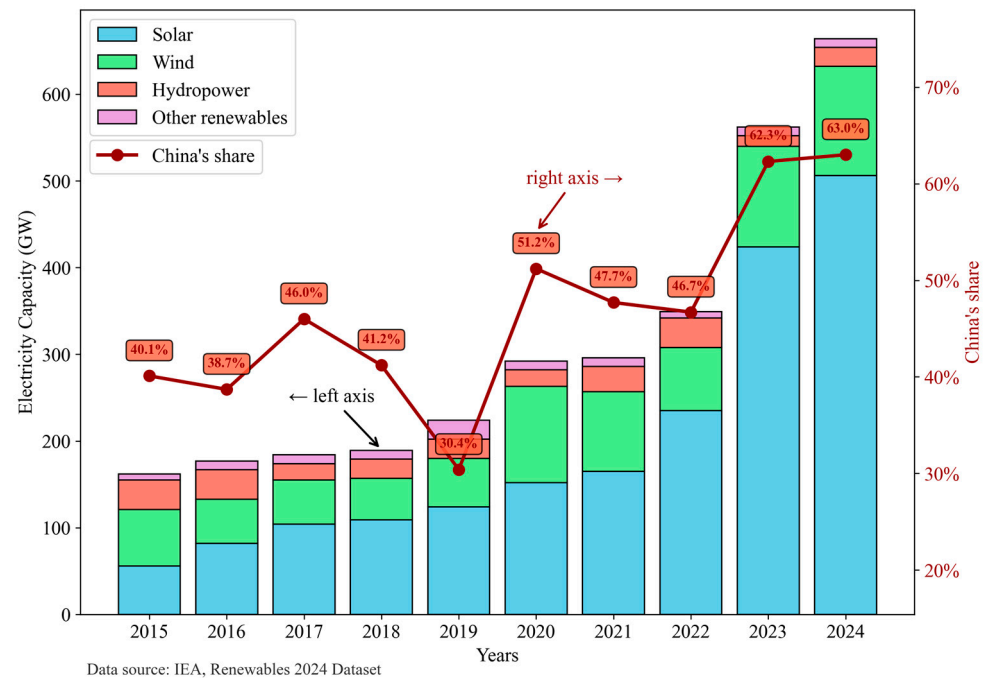


Figure 1. Annual Additions to Renewable Electricity Capacity by Technology and China's Corresponding Share, 2015–2024 (Data source: IEA, Renewables 2024 Dataset).

However, as the proportion of renewable energy sources like photovoltaics continues increasing within the energy mix, the net load curve of the power grid often shows a duck shape, a phenomenon widely known as the “duck curve” [2]. Furthermore, random fluctuations in photovoltaic power generation may threaten the secure and reliable operation of power systems along with their economic efficiency [3]. Consequently, a precise PV power forecast is important to ensure seamless integration with the grid and maintain operational stability. To meet this demand, the industry has developed a variety of tools and methods to reduce forecast errors. The primary methods of time series prediction include conventional statistical approaches, machine learning modelling, and deep learning technology, as well as ensemble and hybrid methods [4]. Among these, extensive empirical research demonstrates that the Autoregressive Integrated Moving Average (ARIMA) model exhibits significant effectiveness for forecasting high-frequency time series. By capturing potential stochastic dependencies in the data, these models have shown significant prediction accuracy in short-term predictions. In comparison to statistical methods like ARIMA, artificial neural networks perform better in capturing complex nonlinear relationships.

Artificial intelligence (AI) technology has experienced accelerated growth in recent years, with artificial neural networks (ANNs) modelled after biological neuron structures emerging as a crucial computational model. ANNs are particularly suitable for processing complex nonlinear data, excelling in tasks such as data organisation, simulation, and optimal solution. Therefore, ANNs are increasingly widely applied in the area of timeseries forecasting. With its inherent adaptability and ability to model complex nonlinear dynamics,

artificial neural networks have become a robust and widely recognised framework in the area of predictive analytics [5]. In the past few years, studies in the area of artificial intelligence have deepened continuously, giving rise to deep learning technology powered by artificial neural networks. This technology has achieved widespread application across various fields due to its remarkable ability to significantly enhance the efficiency of solving complex computational problems [6,7].

Although multilayer perceptron (MLP)-type networks exhibit strong performance in function approximation, they remain constrained by their inherent structure when handling the time series correlations within time series data. Specifically, these dependencies relate to the capability of the model to capture multiscale temporal characteristics. To predict the behaviour patterns of sequence data, specialised neural network architectures are required. In response to this challenge, recurrent neural networks (RNNs) were introduced. RNNs are network architectures designed for sequence data processing. Its core feature lies in the internal feedback connection mechanism, which enables the network to retain memories of previous inputs, thereby capturing time dependencies within sequences. However, the sensitivity of RNNs to gradient explosion and vanishing gradient problems constitutes one of the main challenges facing this method [8].

More specifically, during gradient backpropagation training of neural networks, the gradient propagates backwards through time and is iteratively multiplied by the chain rule and derivatives. When these derivatives are consistently less than 1, repeated multiplication can cause gradients to shrink exponentially toward zero; this phenomenon is referred to as the gradient vanishing problem. In contrast, if the derivative consistently exceeds 1, the gradient may grow exponentially, resulting in gradient explosion. Both of these problems can seriously hinder the effective training of recurrent neural networks [9]. LSTM networks are capable of addressing the aforementioned issues. This approach introduces a circular connection for storage units within the network. Each memory cell contains multiple storage blocks that store the network time state, and also integrates a gating unit that controls the flow of information [10].

As shown in Figure 2, neural networks are employed to predict and schedule intermittent renewable energy sources, a process that facilitates the conversion of fluctuating inputs into a stabilised output for the grid. (Note: During the manuscript preparation, Gemini 1.5 was utilized to enhance the visual clarity of Figure 2). As shown in the figure, the left side displays the power produced by renewable energy sources like solar power and wind. Their output power is directly affected by meteorological factors like sunlight, rainfall, and wind speed, leading to fluctuations in power flow. The system uses neural network technology to predict these fluctuating power flows and schedule grid operations, transforming unstable inputs into steady energy streams. The electricity is finally transmitted to the urban power grid through substations and transmission towers, providing a reliable guarantee for various urban electricity needs. Accurate power generation forecasting holds critical importance to enhancing photovoltaic energy integration within urban smart grids. Within energy markets, when renewable energy generation deviates from the forecasted dispatch, the grid will face challenges such as peak shaving, voltage regulation, and scheduling of traditional coal-fired power plants. Due to the unpredictable and intermittent properties of photovoltaic power, its grid connection increases the complexity of grid management and affects the dynamic balance of supply and demand. Accordingly, a series of methods and tools have been designed to forecast photovoltaic output over different time periods.

PV forecasting has a crucial impact on the smart grid system's planning, management, and monitoring. However, as photovoltaic power generation exhibits random and intermittent characteristics, traditional methods struggle to achieve satisfactory results. ANNs have shown significant advantages in mathematical modelling of renewable energy-based

systems [11,12]. Published research in this area has focused on solar radiation prediction. For example, Gao et al. developed a prediction approach using a recurrent neural network framework. Through the design and optimisation of the model architecture, its explainability was enhanced. The neural network successfully captured the variables in the dataset that were most correlated with solar radiation. The results indicate that solar radiation is strongly linked to multiple factors like temperature, precipitation, vapour pressure and radiation duration [13].

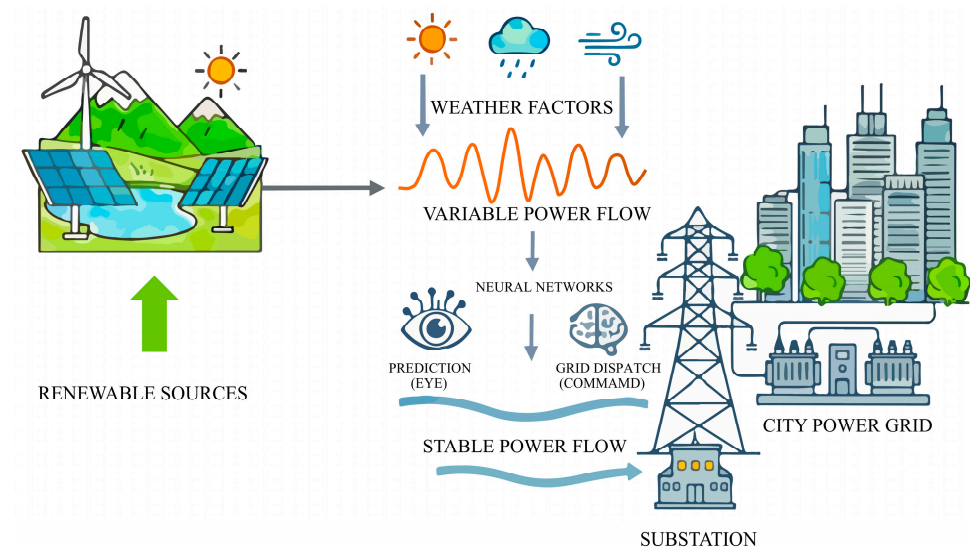


Figure 2. Schematic diagram of predicting photovoltaic power generation using neural networks. (Note: Graphic elements were partially assisted by AI tools, while the overall layout was designed by the authors).

However, the actual power output of a PV module is not solely determined by solar irradiance; it is also affected by the combined action of multiple environmental and operational factors. Including solar cell type, ambient temperature, angle of incidence of sunlight, weather conditions and other parameters, all will affect the final power generation [14]. Wang et al. developed a power forecasting model for photovoltaic systems applying the Elman neural network. The method uses four physical parameters of wind speed, wind direction, atmospheric pressure and humidity as model inputs. Compared with traditional backpropagation networks, it has fewer iterations and is less computationally intensive. The results show that the introduction of additional meteorological factors in low-sunlight areas can significantly enhance the model's predictive precision [15].

In comparison to conventional statistical models like ARIMA, deep learning algorithms represented by LSTM networks show significant advantages in dealing with regression and classification problems in time series. This is due to its capability to self-adjust internal hyperparameters via supervised learning, typically achieving lower prediction errors [16]. Consequently, this technology has been extensively applied in renewable energy forecasting and optimisation [17]. In recent years, many studies have applied LSTM networks to wind speed prediction, and it has become the most mainstream artificial neural network model in the field of short-term wind speed prediction. To solve the common gradient disappearance or explosion problem in RNN training, Elsaraiti et al. developed a prediction model using LSTM and deep neural network. This model can realise multi-step wind speed prediction. Experimental results indicate that the model notably enhances the precision of short-term wind velocity forecasting [18]. Liu et al. introduced a newly developed hybrid deep learning modelling approach. An LSTM neural network based on deep learning methods is employed to forecast the low-frequency wind velocity sublayer;

meanwhile, an Elman neural network is used to predict the high-frequency sublayer. Experimental findings show that the model exhibits outstanding prediction precision and stability in high-precision wind speed forecasting tasks [19].

In the process of modelling and simulation of solar energy systems, AI tools are also widely used. Iheanetu offers a systematic and comprehensive review of the latest advancements in photovoltaic power generation prediction technology, along with an analysis of the strengths and limitations across various solar PV forecasting methods [20]. Dhaked et al. employed LSTM and BPNN for predicting electricity generation at solar power plants, and the results showed that in MAE, MAPE, RMSPE and R^2 metrics, the prediction values of this model show a close match to the actual power generation. The study also conducted a comparative evaluation on the performance characteristics across various meteorological seasons for the multi-layer LSTM model. The error metrics of both LSTM and BPNN models show that LSTM prediction accuracy is higher [21]. Elsaraiti et al. explored the application of deep learning techniques for predicting short-term power generation at photovoltaic power plants. They employed the LSTM algorithm to assess its capability for forecasting solar power generation data; evaluation metrics like RMSE and R^2 were used to compare the performance of LSTM networks versus MLP networks. The findings indicate that among all types of daily forecasts, the LSTM network performs the best [22].

Sharadga et al. conducted a comparative assessment of statistical methods (ARMA, ARIMA, SARIMA) and AI approaches to predict time series of power generation for large-scale photovoltaic power plants, based on 3640 h of operational dataset obtained at a 20 MW connected PV power station in China. The comparative analysis indicates that neural networks demonstrate higher accuracy in photovoltaic power output forecast compared to traditional statistical models, while also requiring shorter computation times. In the absence of solar radiation measurements or meteorological parameters, both neural networks and statistical models can efficiently predict the hourly power generation of photovoltaic power plants. However, time series predictions for photovoltaic power stations are only reliable when the forecast range is within one hour [23].

Zhang et al. proposed a model using the dynamic Bayesian network (DBN) method, which comprehensively integrates multi-source information such as sensor data, operating parameters, meteorological data, historical power generation and model errors. It aimed to increase the prediction reliability of short-term and medium-term electricity generation at photovoltaic power plants. The outcomes indicate that the DBN model generally achieves high predictive accuracy within the prediction period of 1 h to 1 week, with a prediction accuracy of 92% to 95% [24].

Despite these methodological advancements, a critical research gap remains in current literature. Most existing studies focus either on predicting pure solar irradiance or on using standalone historical power sequences, without fully addressing how different weather variations affect large-scale commercial power plants. As summarized in Table 1, existing methods fail to balance computational efficiency with stability under abrupt weather changes, limiting their real-time edge deployment. While LSTMs offer an optimal balance, current applications often ignore the operational reality of utility-scale plants under volatile weather, leaving a critical gap for practical grid scheduling.

To bridge this gap, this study proposes a robust multi-feature LSTM framework specifically optimized for short-term, day-ahead PV power forecasting. The model is validated using four months of real-world operational data (1 June to 30 September 2025) collected from a 100-megawatt utility-scale PV power plant in Hebei Province, China. By integrating historical power output with multi-source meteorological variables at a 30 min resolution, our model seeks to improve forecasting precision under diverse weather

conditions. The remainder of this paper is organized as follows: Section 2 details the data preprocessing procedures and meteorological factor correlation analysis; Section 3 introduces the LSTM-based deep learning framework and its implementation; Section 4 presents the experimental results and a detailed discussion; finally, Section 5 concludes the paper with primary contributions and outlines potential directions for future study.

Table 1. Comparative analysis of existing methods for forecasting photovoltaic power generation.

Methodology Category	Representative Studies	Main Advantages	Key Limitations
Statistical Methods	Sharadga et al. [23], Elsaraiti et al. [16]	Computationally efficient; Easy to implement; Highly accurate for ultra-short-term horizons (≤ 1 h).	Struggle to capture non-linear weather fluctuations; Error rates increase sharply beyond a 1 h horizon.
Traditional Machine Learning	Wang et al. [15], Zhang et al. [24]	Fewer iterations than deep networks; Robust on small datasets with low computational demand.	Limited feature extraction capabilities; Prone to local optima; Cannot effectively model long-term temporal dependencies.
Standard Deep Learning	Gao et al. [13], Dhaked et al. [21], Elsaraiti et al. [22]	Excellent at capturing non-linear patterns and long-term dependencies; Lower prediction errors on large datasets.	Exhibiting a noticeable time lag during sudden meteorological transitions; Vulnerable to data scarcity under extreme weather.

2. Methodology

This section describes an LSTM-based predictive modelling method for solar power generation, which integrates historical data with a variety of meteorological features. Meteorological data offers a scientific foundation for the potential electricity generation theory, and the power generation history data reflects the objective laws governing the actual power generation of the power station. This integration enables us to develop high-precision predictive models that align with physical principles and closely match the practical operational states of power plants. Considering the application of this model to practical products, this study employs iterative training and testing using fixed-size sliding windows to validate the model’s generalisation capability and stability. The process primarily consists of 6 stages: (1) Collection of data, (2) Processing of data, (3) Modelling and implementation, (4) Model evaluation, (5) Forecasting PV power output, (6) result visualisation and display. As shown in Figure 3, a flowchart of the proposed approach is provided. Through comparing the evaluation metrics against base models, the superiority of the LSTM model integrating historical data and multiple meteorological features in photovoltaic power generation forecasting will be demonstrated.

2.1. Dataset and Preprocessing

The solar power generation data in this paper is sourced from a grid-connected photovoltaic power station located in Hebei Province, China, with coordinates of longitude 113°67' and latitude 37°73'. Each of the plant’s photovoltaic panels consists of 32 monocrystalline silicon cells, forms an array of solar panels, and the total installed capacity is 100 MW. The photovoltaic modules are installed with fixed brackets, with an optimal tilt angle of 33 degrees. Since the PV power station was newly constructed in 2025 and lacked historical records, the collected power data only covers a total of 112 days from 1 June 2025 to 30 September 2025. Analysis shows that power values between 6:00 p.m.

and 9:00 p.m. are extremely low and often zero, while the output between 9:00 p.m. and 7:00 a.m. the next morning always remains at zero. Consequently, we restrict the forecast window to the 7:00 a.m.–6:00 p.m. frame. Including these long periods of near-zero values introduces a large volume of zeros into the time series, which can cause the standard LSTM model to overfit to zero and decrease its sensitivity during peak hours. To avoid this operational bias and maintain focus on the high-impact dispatch period, these near-zero hours were excluded.

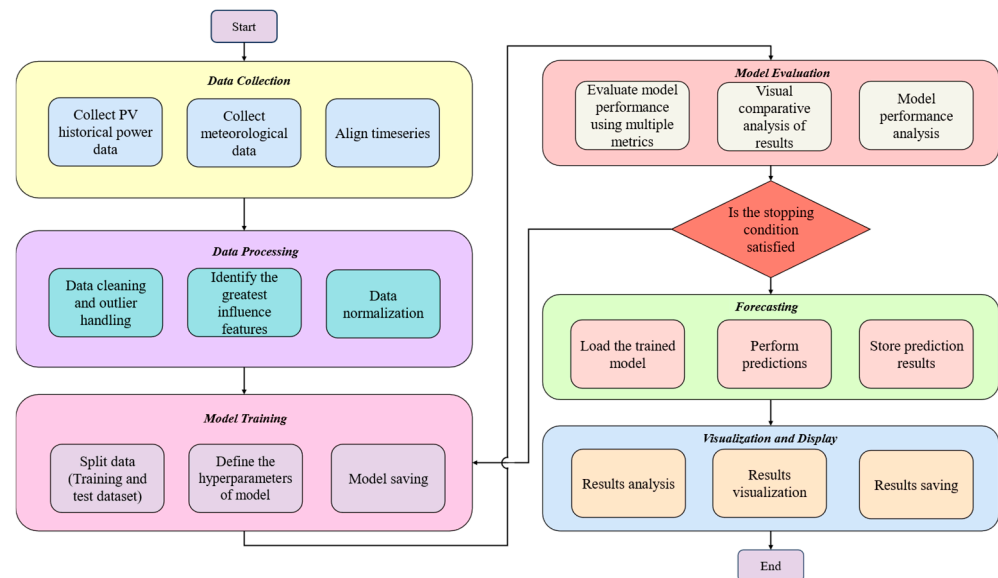


Figure 3. Flowchart of multi-input feature LSTM photovoltaic power prediction.

The raw power data was collected at 15 min intervals; a total of 11,712 data points were obtained ($122 \times 96 = 11,712$). To handle data anomalies, missing values were imputed using the moving average of the preceding 60 min. This localized imputation approach was selected because solar generation profiles exhibit strong temporal locality, meaning that meteorological conditions generally remain stable within short intervals. To meet the modelling demands for the next day's 30 min power values in subsequent forecasts, the data were ultimately resampled at 30 min intervals. As a result, 22 data points are generated every day, with a total of 2684 data points for the four-month study period ($22 \times 122 = 2684$). Figure 4 shows the weekly statistics of photovoltaic power generation from 7:00 a.m. to 6:00 p.m. every day from 1 June to 30 September 2025. The figure also shows that power generation exhibits periodic fluctuations, with an overall average power of 33.95 kW. Specifically, the output peaks between W28 and W30 with a maximum median of 52.8 kW, followed by a sharp decline to a minimum of 18.8 kW in W35. Additionally, box plots in Figure 4a capture wide data distributions and a notable maximum anomaly outlier of 88.3 kW in W34. These variations are closely linked to seasonal weather shifts and localized weather fluctuations.

The meteorological dataset employed in this study was collected from weather stations situated near photovoltaic power stations. The dataset contains meteorological factors like temperature, wind speed, precipitation, humidity, air pressure, solar radiation, cloud cover, and $\text{PM}_{2.5}$ concentration. This dataset covers the same observation time period as the power data. The original sampling interval was 1 h, which was converted to a 30 min interval for use in predicting photovoltaic power generation based on multiple features input from measurable meteorological conditions. Table 2 provides an overview of all meteorological data.

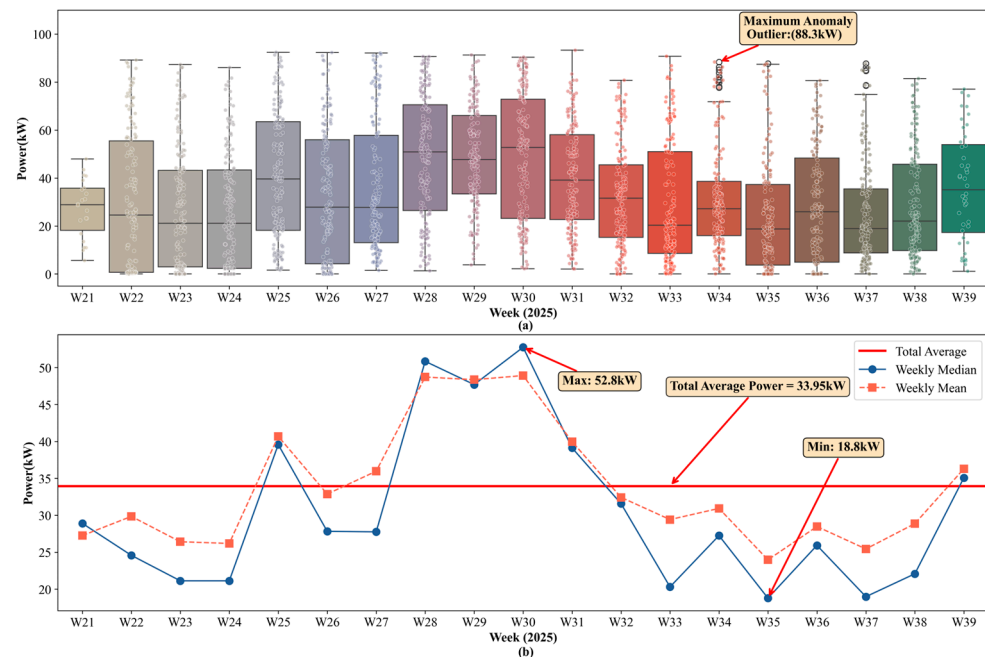


Figure 4. Photovoltaic power generation from 1 June to 30 September 2025, (a) box plot of weekly power generation distribution, (b) statistics trend chart of weekly power generation.

Table 2. Overview of the meteorological data.

Parameters	Range	Mean
Temperature (°C)	7.8~36.5	25.5
Wind Speed (km/h)	0~36	7.9
Precipitation (mm)	0~21.4	0.16
Humidity (%)	9~99	60.4
Air Pressure (hPa)	910~937	923.6
Direct Normal Irradiance(W/m ²)	0~808.7	312.9
Cloud Cover (%)	0~100	32.8
PM _{2.5}	1~65	23.3

For the purpose of minimising the impact of differing data dimensions on the model and enhancing computational efficiency, all data were normalised to the interval (0–1). Different data types were normalised using different algorithms; for instance, power data employed the MinMaxScaler method, and temperature data used the StandardScaler method. In this study, all predictive models were developed and run in a Python 3.9.7 environment. The neural network components were based on the PyTorch 2.3.0 framework, with LSTM models constructed for training.

2.2. Meteorological Features Extraction and Identification

Photovoltaic power generation is significantly dependent on meteorological conditions, which influence its output to various degrees. Consequently, the extraction and identification of meteorological features represent one of the most important elements within the field for PV power forecasting. The raw meteorological data employed contains a wide variety of features. It is essential to determine from the data which features contain the most critical information. This information helps provide an accurate prediction. Additionally, it reduces the negative impact of redundant data and optimises the computational efficiency of the model [25]. Here, Python's plotting libraries facilitate data visualization and the analysis of relationships between electricity generation and other parameters before the construction of the prediction model. Consequently, Pearson's correlation coefficient

was employed to identify the parameters exhibiting the greatest interdependence. The first step is to generate a feature-related matrix to filter out unrelated data and extract the most important features. Pearson's correlation coefficient serves as a tool of statistics for measuring the degree of linear correlation between two independent variables. As the first formally defined correlation measurement method in history, it remains one of the most widely used relevant metrics today [26]. This coefficient assesses the degree of linear relationship between two normally distributed continuous random values, and its mathematical definition may be expressed as follows [27]:

$$R_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \quad (1)$$

where $\bar{x}$, $\bar{y}$ indicate the mean of the variables x and y , respectively. R_{xy} is the correlation coefficient. The value of R_{xy} ranges from -1 to 1 and is not affected by the linear transformation of these two variables. When $R_{xy} = 0$, there is no significant linear correlation between x and y ; when $R_{xy} = 1$, it indicates a complete positive correlation between x and y ; when $R_{xy} = -1$, it represents that x and y are completely negatively correlated.

Following the data analysis, positive correlations were identified between photovoltaic power generation and several parameters within the dataset, including temperature, wind speed, air quality, visibility, direct normal irradiance, as well as PM_{2.5} and O₃ concentrations. Among them, direct normal irradiance showed the strongest correlation, with a correlation coefficient of 0.46. Meanwhile, the increase in temperature is positively correlated with the amount of electricity generated (Pearson correlation coefficient $R_{xy} = 0.43$). This is because rising temperatures mean a corresponding increase in solar radiation energy. In contrast, precipitation, wind direction, air pressure, humidity and cloud cover are negatively correlated with power generation. Among them, the negative correlation between humidity and photovoltaic power generation was the most significant, and the Pearson correlation coefficient reached -0.28 . This is because as humidity increases, water vapour in the atmosphere will scatter and reflect solar radiation, thereby decreasing the effective light intensity that reaches the solar cells and resulting in a drop in output power. Moreover, the Pearson correlation coefficient for cloud cover reached -0.26 , which is easily understood: higher cloud cover indicates less sunlight reaching the photovoltaic panels, thereby reducing power generation output. The corresponding values are shown in the Pearson heatmap in Figure 5. The different colors in the heatmap represent the direction and strength of the correlation between the variables. Specifically, pink cells indicate positive correlations, while blue cells represent negative correlations, with deeper shades reflecting stronger relationships.

2.3. Impact of Meteorological Parameters on Photovoltaic Power Output

Photovoltaic technology is the process of directly converting solar radiation into electrical energy. Its fundamental principle involves forming a P-N junction semiconductor within the photovoltaic cell using different materials. When solar radiation is incident, electrons transition from the valence band to the conduction band, thereby forming electron-hole pairs. Under the effect of a built-in electric field, a current flow is produced. However, the efficiency of solar cells is affected by multiple factors. Consequently, conventional single-junction solar cells face theoretical constraints. Under standard AM1.5 solar illumination conditions, the Shockley–Queisser limit restricts their maximum efficiency to approximately 34% for a bandgap of around 1.3 eV [28].

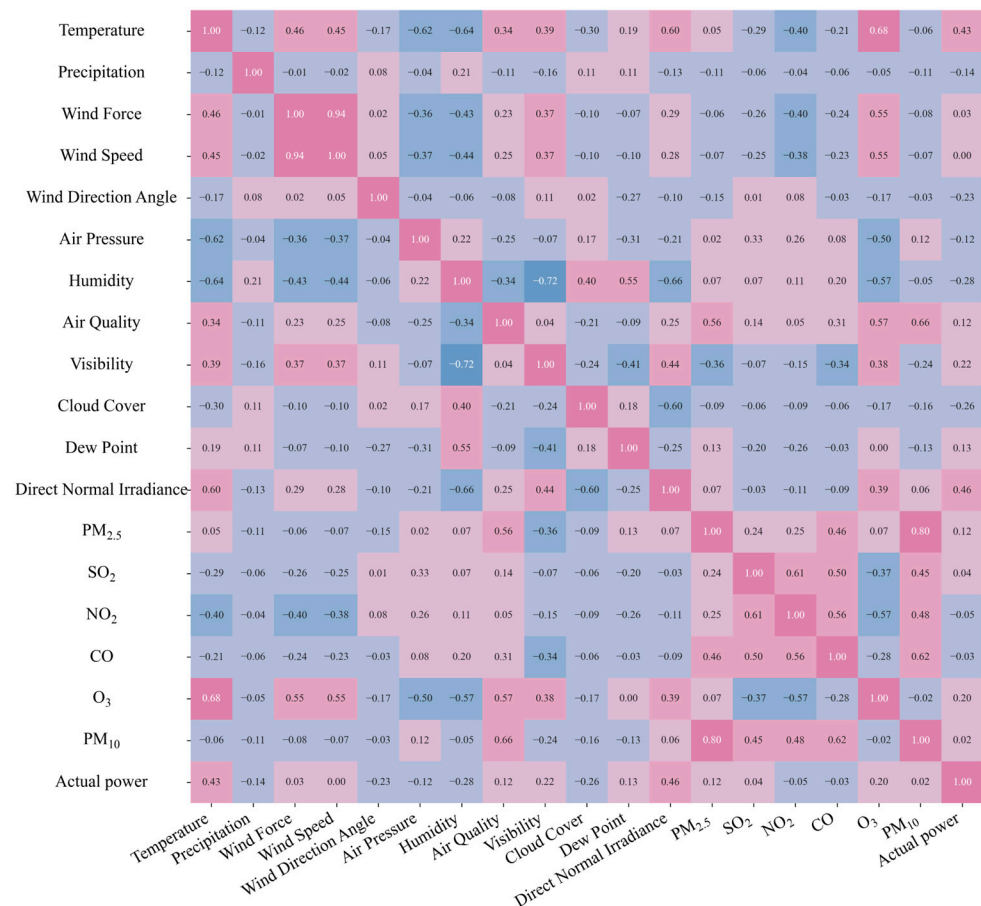


Figure 5. Pearson correlation heatmap of multiple meteorological factors.

The efficiency equation for solar cells can be expressed as:

$$\eta = \frac{V_{OC} \cdot I_{SC} \cdot FF}{P_{in}} \quad (2)$$

where I_{SC} corresponds to the maximum electric current of the cell, V_{OC} indicate the maximum voltage of the cell, and FF represents the fill factor coefficient. P_{in} is the total incoming sunlight power, the equation of which is shown below:

$$P_{in} = I \cdot A \cdot \cos(\theta) \quad (3)$$

where I is the intensity of irradiation, A is the area of the solar panel, and θ is the angle of incidence of sunlight.

From the above formula, it can be concluded that at a given location, the area of the photovoltaic panels and the intensity of solar radiation together determine the amount of electricity the photovoltaic cell can produce. Sunlight intensity fluctuates seasonally and periodically with weather changes. For example, in the northern hemisphere, in summer, the zenith of the sun is close to the Tropic of Cancer, which leads to an increase in the intensity of solar radiation and a decrease in winter. Concurrently, local meteorological conditions such as rainfall or cloudy days can also affect the incoming radiation intensity to the surface of photovoltaic panels. Under the condition that the physical conditions remain unchanged, the installation area of the photovoltaic panels remains fixed. However, precipitation, dust accumulation, and cloud cover can all affect photovoltaic power generation efficiency.

When solar radiation travels to the Earth, its distribution is driven by two primary factors: the planet's rotation and its orbital revolution around the Sun. The combined effect of these two movements leads to regular changes in solar radiation received in different geographical locations. Outside the Earth's upper atmosphere, total solar radiation is about 1361 W/m^2 [29]. However, the solar radiation flux observed at the surface is much lower. This reduction is mainly attributed to atmospheric attenuation, where radiation weakens as it passes through the atmosphere. These physical processes include: the absorption of light by various gaseous species, molecular scattering, and the absorption and scattering by dust particles. Additionally, the orientation and tilt angle of photovoltaic panels, the latitude of the solar power station's location, the time of year, the specific time of day, and the degree of cloud cover affect the level of solar radiation reaching the surface of solar panels.

Meteorological parameters like radiation from the sun, relative humidity, temperature, wind speed, cloud cover and sunshine duration both affect scattered or diffused solar radiation from all directions in the sky. Therefore, establishing a predictive and evaluative model for the impact of meteorological data on photovoltaic power generation is crucial, as this model performs a significant role in converting sunlight into electrical energy.

3. Model Description and Implementation

3.1. Artificial Neural Networks and LSTM

Artificial neural networks (ANNs) are a fundamental component of artificial intelligence and deep learning. They simulate the interconnected structure of neurons in the human brain through computer programs, thereby learning complex dependencies among inputs and outputs. This technology has been extensively applied in the modelling of non-linear and complex functions [30]. In a backpropagation neural network, data is transmitted from inputs to outputs through hidden layers. This study employs a hierarchical recurrent neural network (RNN), whose structure shares similarities with feedforward networks and which generates dynamic responses to input time series data. However, time series typically contain long-term trends, cyclical patterns, and random noise, making predictions challenging for traditional RNNs. Long Short-Term Memory (LSTM) networks, as a specially designed recurrent architecture, are capable of capturing long-range dependencies that conventional RNNs struggle to handle [31].

LSTM networks share a similar layered architecture with MLPs, both containing input, hidden, and output layers. However, LSTMs are distinguished by the incorporation of memory units within the hidden layers. Each memory unit is governed by three gates: forget gate (f_t), input gate (i_t), and output gate (O_t). Figure 6 illustrates the basic structure of an LSTM unit. The core component of an LSTM is the cell state, represented as C_t , which carries information across time steps and is passed from one iteration to the next. The current cell state C_t is updated by combining the previous cell state C_{t-1} with the current input signal x_t , modulated by the three gates. These gates employ a specific design that dynamically regulates the flow of information—deciding what to retain or discard—through a combination of sigmoid and tanh neural network layers, followed by pointwise multiplication operations [10].

The state update of an LSTM cell is governed by a set of mathematical formulas, where w and b indicate the weight coefficients and biases, respectively. The first step involves the forget gate, which takes the current input x_t and the previous hidden state h_{t-1} as inputs. A sigmoid layer determines the extent to which information from the previous cell state C_{t-1} should be discarded. The mathematical definition of the forgetting gate f_t is as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (4)$$

Among these, σ represents the sigmoid function, which serves as the activation function for the gates.

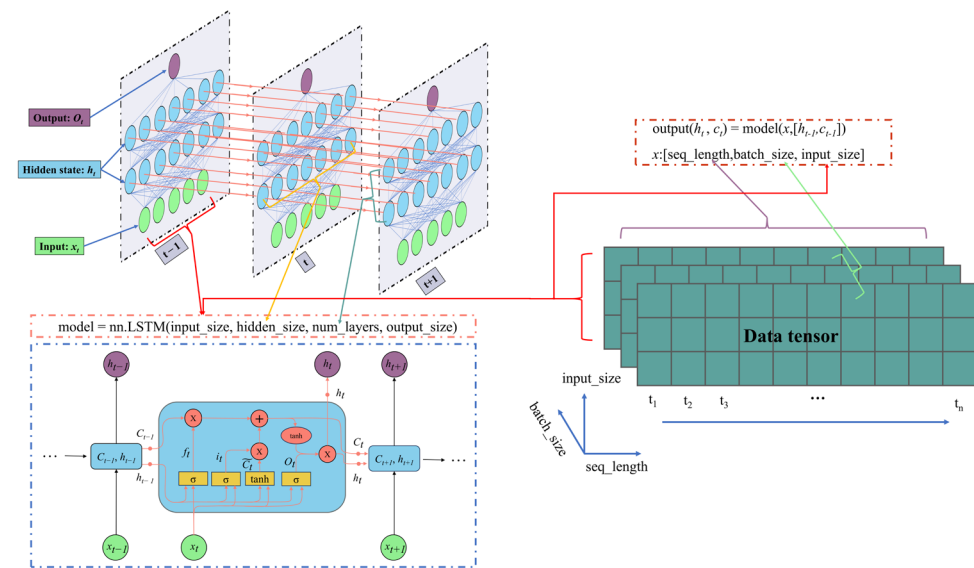


Figure 6. Basic architecture of LSTM.

The following step is to define the new input required to update the LSTM cell state. This process includes two parts: first, the input gate controls which information will be refreshed; second, a Tanh activation layer produces the potential state vector $\tilde{C}_t$, representing the new information potentially appended to the unit's status. Finally, the outputs from both are combined to refresh the current unit status. The detail is expressed by:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (5)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (6)$$

Subsequently, the previous cell state C_{t-1} is updated to the current cell state C_t . This is expressed as follows:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (7)$$

Lastly, output layers will generate data based on our cell state, obtained by employing the following expression.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (8)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (9)$$

The training of the LSTM model is a continuous iteration of the aforementioned process, learning to update parameters (w) and the bias (b) to minimise the cost function between training values and output values [32].

3.2. Model Implementation

Artificial neural networks provide a powerful computational tool for handling complex problems, exhibiting adaptive behaviour when processing complex and noisy information. Their outstanding nonlinear characteristics, robust stability, high tolerance for errors, and self-learning capabilities make them particularly well-suited for nonlinear problems such as photovoltaic power generation forecasting [33,34].

Based on the analyses in Section 2.2, a two-step feature selection strategy was implemented to determine the input variables. Initially, a Pearson correlation analysis was conducted across all candidate meteorological factors. Variables were pre-screened us-

ing a baseline threshold of $|R_{xy}| > 0.1$, which yielded 12 qualified features. To reduce computational complexity, avoid model over-fitting, and maintain a parsimonious model structure, the inputs were further refined by selecting only the top two positively and top two negatively correlated factors. Consequently, direct normal irradiance ($R_{xy} = 0.46$) and temperature ($R_{xy} = 0.43$) were chosen as the leading positive inputs, while humidity ($R_{xy} = -0.28$) and cloud cover ($R_{xy} = -0.26$) were selected as the primary negative inputs.

The training of LSTM neural networks primarily employs gradient descent-based optimisation algorithms, which compute the gradient of the loss function through automatic differentiation. Solar photovoltaic power generation prediction models primarily rely upon two critical parameters: network depth and neuron count. The Mean Squared Error was employed for performance evaluation metric in this study.

The optimal network architecture was selected through a structured search process using cross-validation to avoid arbitrary estimation. We evaluated specific hyperparameter ranges: LSTM layers (1 to 3), hidden units per layer (16 to 128), and learning rate (0.001 to 0.1). The results show that a configuration with 2 hidden layers, 32 hidden units, and a learning rate of 0.001 offers the best trade-off. It achieves the highest validation accuracy while minimizing both computational overhead and overfitting risks. Further increasing the network depth or neuron density provided no meaningful performance gains.

To prevent overfitting in the recurrent neural network model, a dropout strategy was employed with a dropout probability of 0.20 set in the study. The LSTM selected the ReLU function as its activation function and employed the ADAM optimiser, which significantly accelerates the model's learning speed. Specifically, the input sequence length (look-back window) is configured as 22, representing the 22 daily effective generation operational hours (7:00 a.m.–6:00 p.m.). To maintain structural consistency and smooth gradient transitions across different days, a Full-batch training approach was implemented. The maximum epoch number is set to 200. Table 3 lists the detailed configuration of the LSTM parameters applied in this study.

Table 3. Model construction.

Model	Hyper Parameters	
LSTM	Input size	5
	Number of neurons	32
	Number of hidden layers	2
	Learning rate	0.001
	Input sequence length	22
	Batch size	Full-batch
	Output layer	1
	Loss function	Mean absolute error
	Optimizer	Adam
	Epochs	200

3.3. Time Series Cross-Validation

Cross-validation is a widely used statistical method for evaluating the performance of machine learning models. As a resampling technique, it partitions the dataset into multiple training and test subsets. By aggregating prediction accuracy metrics across these subsets, researchers can obtain a comprehensive assessment of the model's generalisation ability [35].

When training machine learning model on time series data, special attention must be paid to the temporal order of the sequential data; it is invalid to use future observations to predict past values. Therefore, when partitioning a dataset into training and test sets, it is essential to preserve the chronological order. Moreover, test sets should be drawn

from periods subsequent to the training set to avoid look-ahead bias. In this study, the sliding window cross-validation method was applied for model evaluation. This technique evaluates time series models by moving a fixed-size training-test window progressively along the timeline, thereby providing an unbiased estimate of model performance. The window length was strategically set to utilize 60 days of continuous historical data to forecast the subsequent half-month (15 days), balancing the capture of recent temporal dynamics against the exclusion of distant historical noise. Figure 7 illustrates the partitioning procedure of sliding window cross-validation. During each validation iteration, the sample sizes of the training and test datasets remain constant: the training set includes 1320 ($60 \times 22 = 1320$) samples, and the test set contains 330 ($15 \times 22 = 330$) samples. Performance metrics are recorded for each iteration and subsequently aggregated to estimate the overall forecast accuracy.

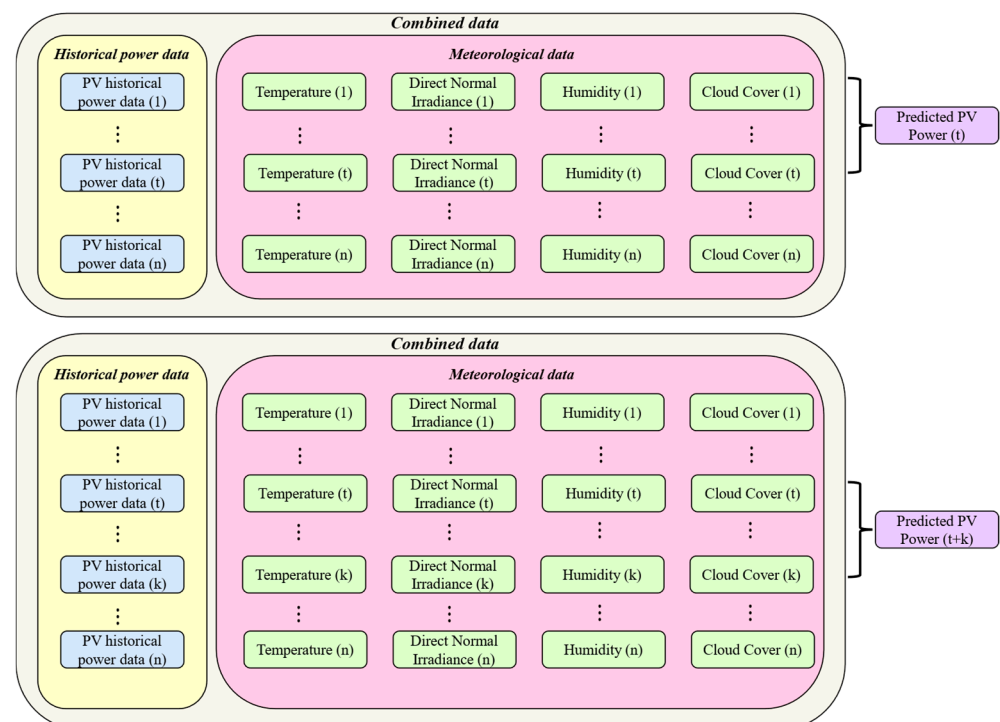


Figure 7. Construct the dataset using the sliding window cross-validation method.

Table 4 illustrates the division of the time series dataset for training and testing the LSTM forecasting model designed for different time periods. The training set comprises 60 days, while the test set consists of 15 days. This configuration allows the model to capture trends from the past two months of data, thereby predicting photovoltaic power generation for the next half-month.

Table 4. LSTM prediction models for different test datasets.

Subset	Training Dates	Test Dates
Attempt 1	1 June to 31 July 2025	1 August to 15 August 2025
Attempt 2	16 June to 15 August 2025	16 August to 31 August 2025
Attempt 3	1 July to 31 August 2025	1 September to 15 September 2025
Attempt 4	16 July to 15 September 2025	16 September to 30 September 2025

3.4. Performance Evaluation

This study uses four different evaluation metrics to measure the predictive performance of the model: Mean Absolute Error, Root Mean Square Error, Mean Absolute

Percentage Error, and Coefficient of Determination. These evaluation metrics evaluate the predictive ability of the model in multiple dimensions.

1. Mean absolute error (MAE)

It represents the average distance between the measured value and the predictive model, calculated as follows:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |p_{act} - p_{pred}| \quad (10)$$

2. Root mean square error (RMSE)

It calculates the average deviation of the predicted values from the actual values using the following formula:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N (p_{act} - p_{pred})^2}{N}} \quad (11)$$

3. Mean absolute percentage error (MAPE)

This metric measures the model's prediction accuracy as a percentage and is calculated as follows:

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^N \left| \frac{p_{act} - p_{pred}}{p_{act}} \right| \quad (12)$$

4. Coefficient of determination (R^2)

It represents the correlation of actual values with forecasted values, calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^N (p_{act} - p_{pred})^2}{\sum_{i=1}^N (p_{act} - \bar{p})^2} \quad (13)$$

where N represents the sample size used for statistical evaluation, $\bar{p}$ represents the mean of the actual values, p_{act} represents the actual value, and p_{pred} indicates the predicted value.

4. Results and Discussion

In this section, the experimental results of applying deep learning techniques to photovoltaic power generation forecasting are presented. Two methods are adopted: single-feature input (historical data) and multi-feature input (combination of historical data and meteorological information). Following the results, the discussion focuses on the core performance metrics of this method, particularly those related to the artificial neural network model.

Figure 8 shows a comparison between observed daily electricity generation and the model's forecast values, alongside the baseline persistence model. Results are presented for a specific day in August. The shadowed area in the figure represents the 95% confidence interval. It can be seen that the vast majority of actual data points remain in the 95% forecast range for most of the day, whereas the persistence model exhibits noticeable time lags and larger deviations from the actual data. The 95% confidence interval of the multi-input model (Figure 8b) is narrower compared to the single-input model (Figure 8a), indicating higher predictive accuracy and robustness. The results show that both models' predictions are highly consistent with the actual data and significantly outperform the traditional persistence model, but the multi-input LSTM model displays pronounced advantages over the single-input LSTM model.

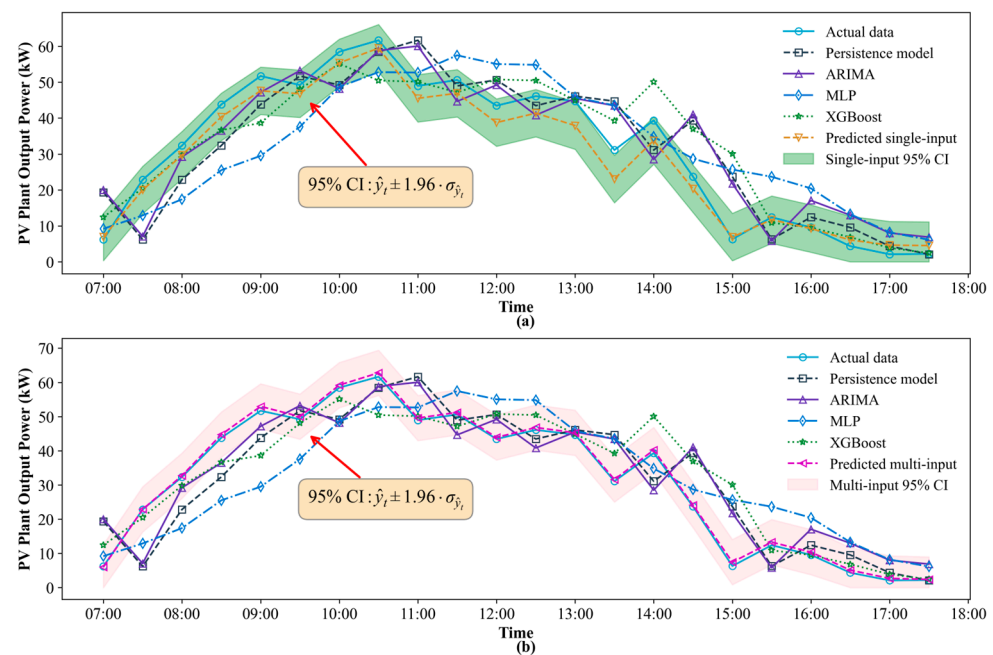


Figure 8. Comparison of actual values with predictions from different models (8 August), (a) single-input model prediction, (b) multi-input model prediction.

In Figure 8a, the traditional MLP model underestimates the power output during the morning ramp-up phase between 08:00 and 10:00, creating a significant vertical gap from the actual data line. Although ARIMA tracks the overall trend better than MLP, it fails to capture the sudden power drops, such as the sharp dip observed at 15:00. The tree-based XGBoost model exhibits a different issue, displaying unstable and highly fluctuating predictions during the afternoon period. For instance, at 14:00, XGBoost generates a false peak that deviates completely from the downward trend of the actual power generation.

In contrast, the proposed multi-input LSTM model (Figure 8b) effectively addresses these tracking errors. By incorporating multivariate meteorological features, the proposed model smoothly follows the continuous fluctuations throughout the day, especially during peak generation hours (10:00–12:00) and the volatile afternoon transitions. While the single-input LSTM curve occasionally drifts near the boundaries of its wider green confidence interval, the multi-input model's prediction curve stays closely aligned with the actual data core inside a narrower pink interval. This visual evidence confirms that the proposed method successfully captures both temporal dependencies and complex multi-variable interactions, making it far more reliable than traditional statistical or shallow machine learning frameworks.

Table 5 presents the power output evaluation indicators of each cross-validation dataset. The analysis shows that both LSTM models demonstrate robust performance across all evaluation criteria, consistently outperforming the baseline persistence model by a large margin. However, the multi-input LSTM significantly surpasses the single-input LSTM in almost all metrics, its error metrics are lower, and the R^2 value is closer to 1. Although the single-input LSTM achieved its optimal performance in attempt 4, its performance was still far inferior to the multi-input model, while still remaining superior to the corresponding persistence model. In Attempt 1, the multi-input LSTM attained peak performance, with a minimum MAE of 1.4749 and an R^2 as high as 0.9864, whereas the persistence model yielded its worst R^2 score of 0.735 in the same attempt.

As detailed in Table 5, a clear performance hierarchy emerges when comparing the proposed approach against ARIMA, MLP, and XGBoost across Attempts 1 to 4. Traditional statistical ARIMA and shallow machine learning models MLP provided stable but limited

predictive accuracy, with R^2 scores generally constrained between 0.78 and 0.86. The tree-based ensemble method, XGBoost, exhibited high sensitivity to dataset fluctuations. In Attempt 3, XGBoost underperformed significantly, yielding a high MAE of 15.4826 kW and a low R^2 score of 0.2319. This drop is primarily driven by sudden, severe weather fluctuations. Regression trees rely on discrete data splitting, making them less capable of capturing continuous, highly volatile meteorological shifts and resulting in localized forecasting failures.

Table 5. Performance comparison of different forecasting models across four test attempts.

Subset	Algorithm	MAE	MAPE (%)	RMSE	R^2 Score
Attempt 1	Persistence model	9.11	33.06	11.86	0.7350
	ARIMA	8.35	35.67	11.28	0.7818
	MLP	8.24	37.91	10.40	0.8146
	XGBoost	9.56	45.03	12.90	0.7145
	LSTM single-input	3.61	12.21	7.47	0.9042
	LSTM multi-input	1.47	7.95	2.82	0.9864
Attempt 2	Persistence model	7.20	29.64	9.49	0.8180
	ARIMA	6.26	44.22	8.46	0.8664
	MLP	6.59	45.52	8.49	0.8652
	XGBoost	8.86	44.68	12.45	0.7102
	LSTM single-input	3.12	29.09	5.05	0.9525
	LSTM multi-input	1.69	11.33	3.62	0.9755
Attempt 3	Persistence model	9.81	38.59	12.35	0.6550
	ARIMA	7.88	69.25	10.62	0.7971
	MLP	7.17	62.60	10.83	0.7887
	XGBoost	15.48	123.03	20.66	0.2319
	LSTM single-input	3.67	30.76	4.96	0.9558
	LSTM multi-input	1.92	10.60	3.38	0.9795
Attempt 4	Persistence model	7.36	32.28	10.12	0.7760
	ARIMA	6.25	42.94	9.17	0.8262
	MLP	6.27	34.57	8.86	0.8376
	XGBoost	9.29	54.37	12.94	0.6539
	LSTM single-input	3.59	34.77	5.44	0.9388
	LSTM multi-input	1.58	15.29	2.81	0.9836

In contrast, the standard single-input LSTM utilized historical time-series patterns to effectively improve prediction accuracy, achieving R^2 scores above 0.90 across all scenarios. The proposed multi-input LSTM model further enhanced this performance, consistently delivering the lowest error rates and the highest goodness-of-fit. For instance, in Attempt 3, the multi-input LSTM reduced the RMSE to 3.3765 kW, outperforming ARIMA (10.6180 kW), MLP (10.8346 kW), and XGBoost (20.6571 kW) by a significant margin.

The overall results show that introducing multivariate meteorological feature inputs can significantly enhance predictability and stability in LSTM models for photovoltaic power generation forecasting tasks, demonstrating a clear advantage over traditional naive, statistical, and machine learning approaches.

Figure 9 presents the comparative results of photovoltaic power generation forecasting among various calculation methods across different attempt, including the proposed multi-feature input LSTM algorithm, persistence model, ARIMA, MLP, and XGBoost. Under stable weather conditions, the predicted curves of the proposed LSTM model are tightly congruent with the actual observations, demonstrating a much higher tracking accuracy than the traditional baseline methods and proving its significant practical value for routine dispatch.

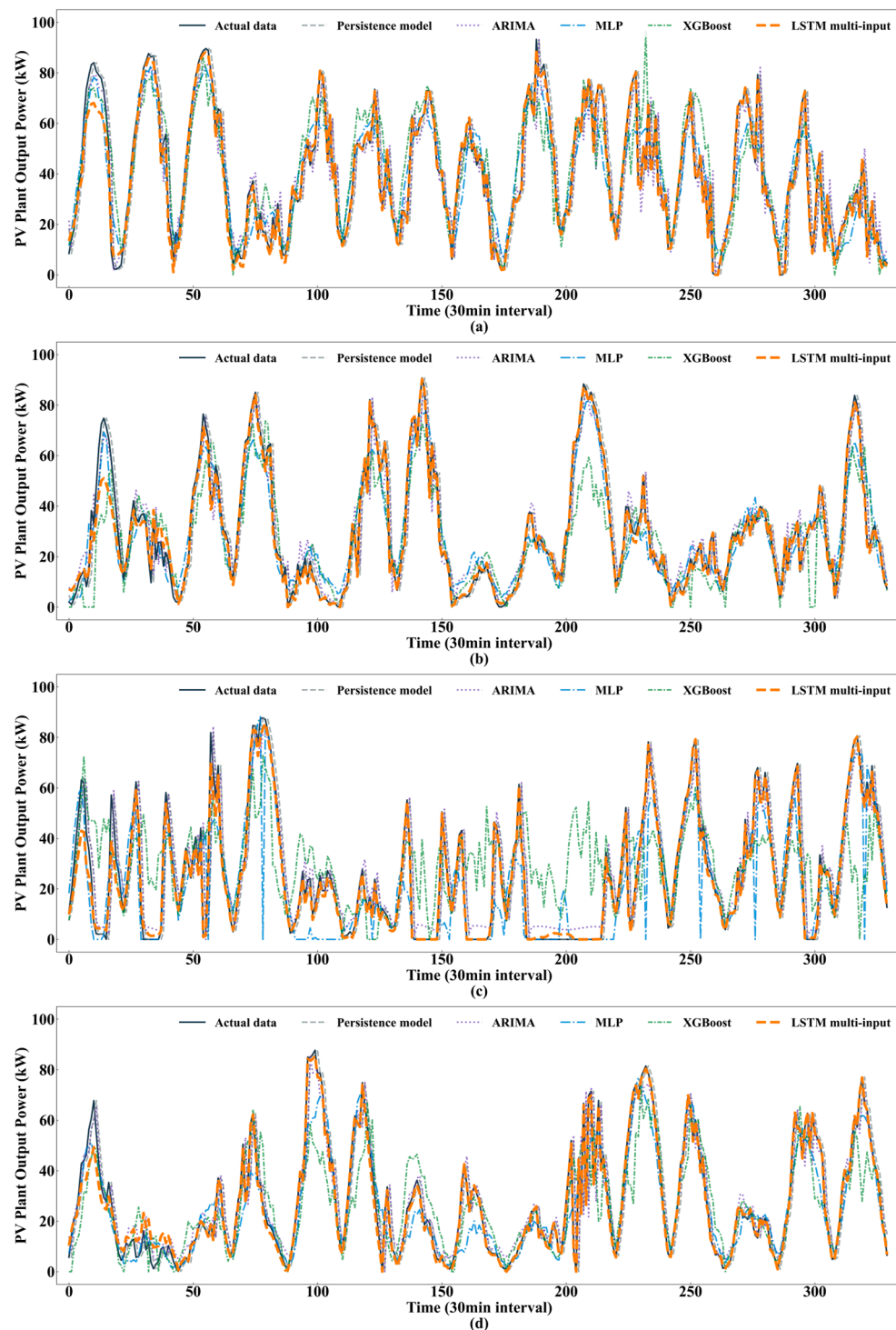


Figure 9. Comparison of predicted and actual PV power generation using different forecasting models across four validation sets: (a) Attempt 1 (1 June–31 July 2025), (b) Attempt 2 (16 June–15 August 2025), (c) Attempt 3 (1 July to 31 August 2025), (d) Attempt 4 (16 July–15 September 2025).

However, during periods of extreme weather or rapid meteorological transitions, the prediction error all compared methods increases noticeably, with machine learning models like XGBoost showing redundant, high-frequency oscillations. This performance degradation primarily stems from two underlying mechanisms: first, the severe sparsity of extreme event samples in the historical training dataset restricts the model's extrapolation capability; second, sudden weather shifts introduce abrupt, highly non-linear variations

that the sequential memory of the standard LSTM architecture struggles to capture instantly without a time lag.

This weather dependence represents a critical limitation, particularly for grid operators who require high certainty during volatile conditions. To mitigate this restriction, subsequent model refinements may explore integrating physics-informed constraints into the neural network to enrich the model's exposure to rare, high-impact weather scenarios. Consequently, while the current integrated model significantly improves the baseline predictability of photovoltaic power generation, addressing these extreme-weather challenges remains paramount to maximizing its economic value for solar power plants.

Figure 10 presents a comprehensive regression analysis comparing the proposed multi-input LSTM model with several baseline models, including the Persistence model, ARIMA, MLP, and XGBoost, across four validation sets. The multi-input LSTM model (Figure 10(a5,b5,c5,d5)) consistently achieves the highest prediction accuracy, with outstanding R^2 values of 0.9864, 0.9755, 0.9795, and 0.9836, respectively. In validation set 1, the LSTM model's predictions align almost perfectly with the actual values (Figure 10(a5)), whereas the MLP (Figure 10(a3), $R^2 = 0.8146$) and XGBoost (Figure 10(a4), $R^2 = 0.7145$) display significant scatter, particularly in the 20–80 kW range. For validation set 2, although all models capture the general trend, they show different levels of deviation at high power outputs above 60 kW. The XGBoost model (Figure 10(b4)) suffers from a severe underestimation under these peak conditions, while the proposed LSTM model (Figure 10(b5)) maintains a much tighter cluster around the diagonal reference line.

In the validation set 3, the performance gap between the models becomes even more pronounced under fluctuating conditions. The XGBoost model performs poorly (Figure 10(c4)), with data points scattered widely across the entire power range, yielding a low R^2 of only 0.2319. Similarly, the MLP model (Figure 10(c3)) shows a visible dispersion ($R^2 = 0.6376$) in the medium-to-low power range of 40–60 kW. In contrast, the multi-input LSTM model (Figure 10(c5)) successfully suppresses this scattering, keeping the points closely aligned with the fit line and achieving a high R^2 of 0.9795. Under the stable weather conditions of validation set 4, all models show improved performance, but the multi-input LSTM (Figure 10(d5), $R^2 = 0.9836$) still clearly outperforms the MLP (Figure 10(d3), $R^2 = 0.8376$) and XGBoost (Figure 10(d4), $R^2 = 0.6539$). These results confirm that the proposed multi-input LSTM model possesses superior stability and generalization ability across diverse meteorological scenarios.

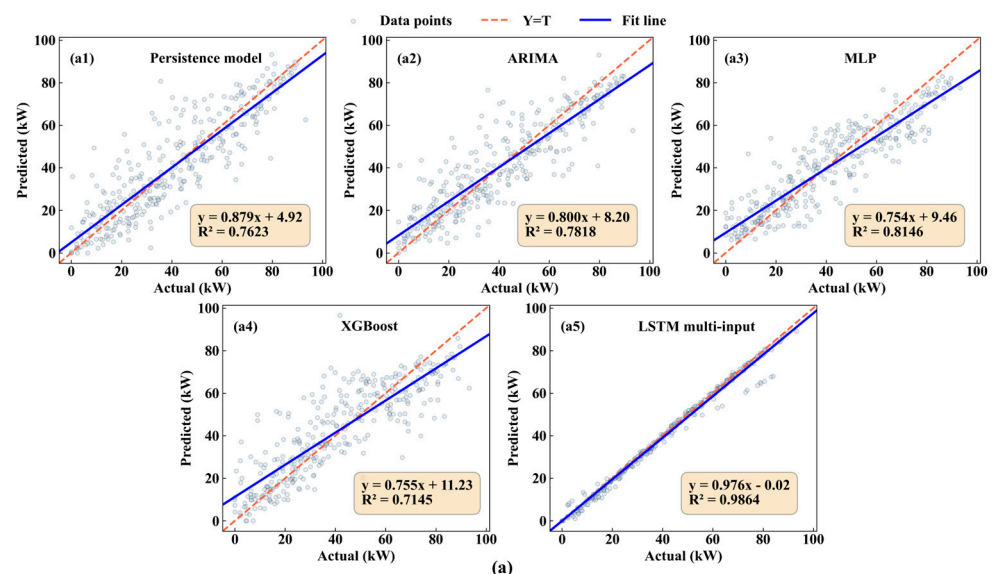


Figure 10. Cont.

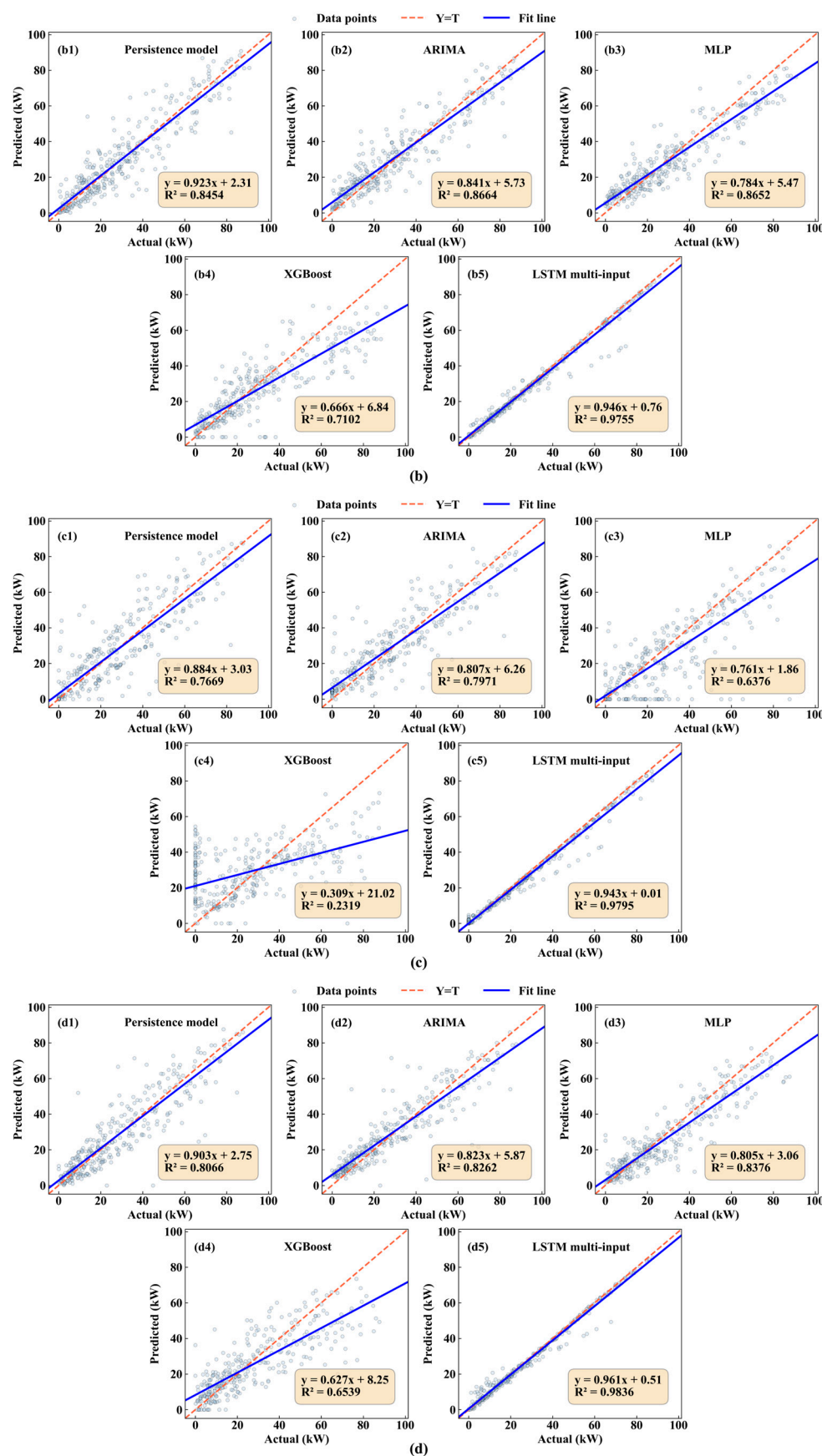


Figure 10. Comparison of regression analysis among different forecasting models across four validation sets: (a) Attempt 1 (1 June–31 July 2025), (b) Attempt 2 (16 June–15 August 2025), (c) Attempt 3 (1 July to 31 August 2025), (d) Attempt 4 (16 July–15 September 2025).

Figure 11 displays the error distribution histograms and normal fitting curves for the five forecasting models across the different validation datasets. Compared with the baseline models, the proposed multi-input LSTM model exhibits a significantly narrower error distribution, indicating a highly concentrated prediction accuracy. In validation set 1, while the baseline models like XGBoost (Figure 11(a4)) show a wide dispersion with a standard deviation (Std) of 12.8, the error of the multi-input LSTM model is tightly clustered within the range of ± 5 kW, yielding a much smaller Std of 2.6 (Figure 11(a5)). Under this set, the LSTM model has an error mean of -1.0 , with 192 out of 330 samples falling in the extremely narrow range of -1 to 1 kW, though a slight tendency to underestimate peak values remains. A similar performance superiority is observed in validation set 2, where the LSTM model achieves an average error of -0.9 kW and a Std of 3.5 (Figure 11(b5)), significantly outperforming the Persistence model (Figure 11(b1), Std = 9.1) and XGBoost (Figure 11(b4), Std = 12.0).

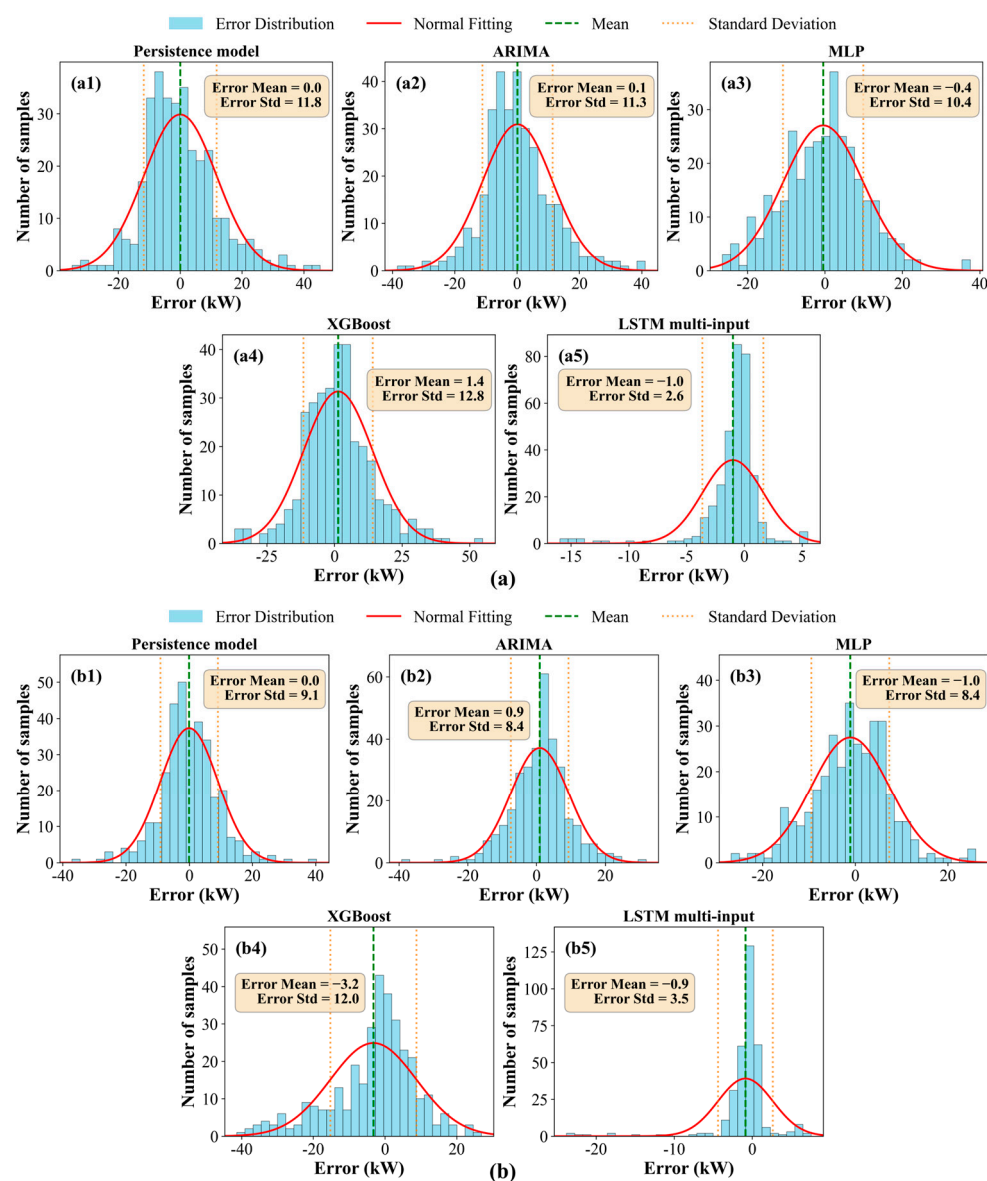


Figure 11. Cont.

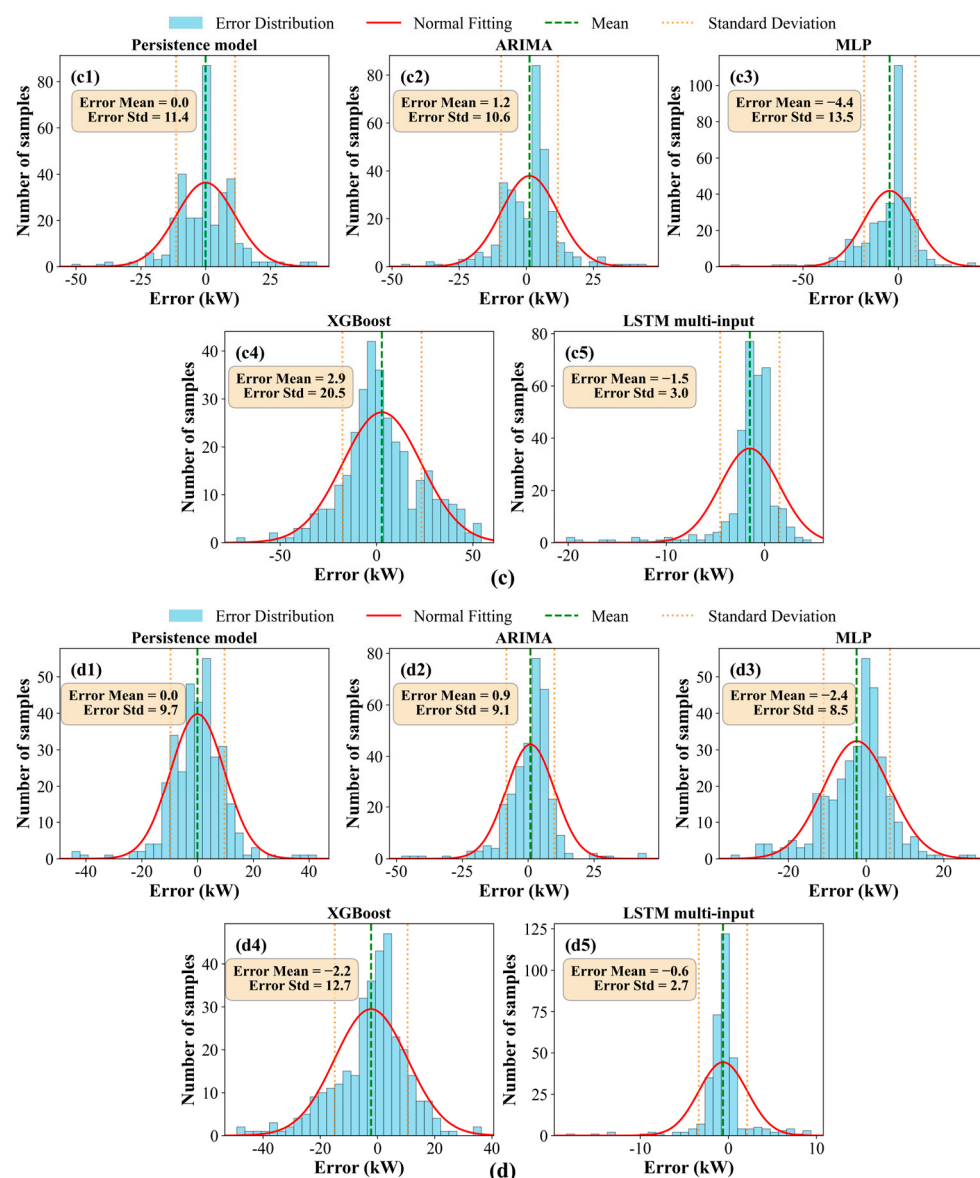


Figure 11. Comparison of prediction error distributions and normal fitting curves among different forecasting models across four validation sets: (a) Attempt 1 (1 June–31 July 2025), (b) Attempt 2 (16 June–15 August 2025), (c) Attempt 3 (1 July to 31 August 2025), (d) Attempt 4 (16 July–15 September 2025).

For validation set 3, which represents highly fluctuating meteorological conditions, the performance gap between the models becomes even more distinct. The benchmark models exhibit poor prediction accuracy, with XGBoost (Figure 11(c4)) and MLP (Figure 11(c3)) showing extremely broad error distributions with standard deviations of 20.5 and 13.5, respectively. In contrast, the multi-input LSTM model (Figure 11(c5)) successfully maintains a concentrated error profile with a Std of only 3.0, demonstrating its excellent anti-interference capability under complex conditions. In validation set 4, the error distribution of the LSTM model (Figure 11(d5)) is highly similar to that of validation set 1, with a Std of 2.7. Here, 244 out of 330 samples exhibit a prediction error of less than zero, reflecting a minor and stable underestimation trend under steady weather conditions. Overall, this detailed error analysis confirms that the proposed multi-feature input LSTM model offers substantial practical value for highly reliable half-hourly solar power forecasting.

5. Conclusions

In this study, an LSTM model was proposed for daily PV power generation forecasting. The model was validated using four months of operational data (1 June to 30 September 2025) collected from a 100-megawatt PV power plant located in Hebei Province, China. Comprehensive comparative analyses demonstrated that the proposed multi-feature LSTM model substantially outperformed single-feature baseline models across diverse weather conditions. Based on the experimental results, the key findings and their practical implications are summarized as follows:

- The multi-input LSTM model achieved its peak performance in Attempt 1, with a minimum Mean Absolute Error of 1.4057 kW and an R^2 value as high as 0.9865. For a 100 MW power plant, such a low error means that grid operators can significantly reduce the requirement for spinning reserve capacity. This enhances the overall stability of the power grid and prevents sudden voltage fluctuations.
- The error distribution analysis showed that the majority of prediction errors in stable datasets fell within a narrow range of ± 5 kW, with most errors concentrated between -1 kW and 1 kW. In practice, this high precision directly translates into substantial economic benefits. It allows power plant managers to make highly accurate day-ahead generation schedules, thereby minimizing financial penalties imposed by grid companies due to power supply imbalances.
- Although the model performed exceptionally well under stable weather conditions, its prediction error increased noticeably during sudden meteorological transitions. Regression analysis across the four validation sets indicated deviations at high power output points (>60 kW) and wider error distributions in unstable periods. This limitation arises because extreme weather samples are rare in historical data, and the standard LSTM architecture exhibits a time lag when capturing sudden non-linear variations. This underscores the need for grid operators to maintain flexible backup options during volatile weather.

While the current integrated model significantly improves the baseline predictability of PV power generation, addressing these extreme-weather challenges remains paramount. Future research will focus on incorporating physics-informed constraints and complementary AI techniques into the neural network. This will help construct more robust, adaptive, and reliable forecasting systems, maximizing both the technical and economic value of solar power plants.

Author Contributions: Conceptualization and methodology, C.H. and M.G.; software, C.H. and Z.G.; formal analysis, C.H.; writing—original draft preparation, C.H.; writing—review and editing, S.Z., X.L. and C.S.; visualization, C.H.; project administration, M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the NSFC (12502288), Guangdong Basic and Applied Basic Research Foundation (2024A1515010711), Guangdong Philosophy and Social Science Foundation (GD26YJY18), Guangdong Department of Education (2025GXJK0621, 2025KQNCX110), Guangdong Department of Science and Technology (F0100018), and BNBU research fund (R0700117-25, R0200021-25).

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy and institutional restrictions.

Acknowledgments: During the preparation of this manuscript, the authors used Gemini 1.5 for the purpose of improving the visual quality in Figure 2. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. IEA Renewables 2024—Analysis and Forecast to 2030. Available online: <https://www.iea.org/reports/renewables-2024> (accessed on 25 January 2026).
2. Calero, I.; Cañizares, C.A.; Bhattacharya, K.; Baldick, R. Duck-Curve Mitigation in Power Grids with High Penetration of PV Generation. *IEEE Trans. Smart Grid* **2022**, *13*, 314–329. [\[CrossRef\]](#)
3. Kanchana, W.; Singh, J.G.; Ongsakul, W. Data-Driven Hidden Solar PV and Energy Storage Capacity Estimation from the Net-Load of Active Distribution Systems. *Sustain. Energy Grids Netw.* **2025**, *44*, 101940. [\[CrossRef\]](#)
4. Sobri, S.; Koochi-Kamali, S.; Rahim, N.A. Solar Photovoltaic Generation Forecasting Methods: A Review. *Energy Convers. Manag.* **2018**, *156*, 459–497. [\[CrossRef\]](#)
5. Fan, Y.; Tang, Q.; Guo, Y.; Wei, Y. BiLSTM-MLAM: A Multi-Scale Time Series Prediction Model for Sensor Data Based on Bi-LSTM and Local Attention Mechanisms. *Sensors* **2024**, *24*, 3962. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Shen, M.; Ye, K.; Liu, X.; Zhu, L.; Kang, J.; Yu, S.; Li, Q.; Xu, K. Machine Learning-Powered Encrypted Network Traffic Analysis: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2023**, *25*, 791–824. [\[CrossRef\]](#)
7. Gupta, P.; Singh, R. Forecasting Hourly Day-Ahead Solar Photovoltaic Power Generation by Assembling a New Adaptive Multivariate Data Analysis with a Long Short-Term Memory Network. *Sustain. Energy Grids Netw.* **2023**, *35*, 101133. [\[CrossRef\]](#)
8. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaria, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *J. Big Data* **2021**, *8*, 53. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Hu, W.; Zheng, T.; Zhang, Y. Study on Carbon Emission Driving Factors and Carbon Peak Forecasting in Power Sector of Shanxi Province. *PLoS ONE* **2024**, *19*, e0305665. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Oyewola, O.M.; Ismail, O.S.; Olasinde, M.O.; Ajide, O.O. Mapping of Solar Energy Potential in Fiji Using an Artificial Neural Network Approach. *Heliyon* **2022**, *8*, e09961. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Kayri, İ. The Effects of Coolant Mass Flow Rate and Atmospheric Indicators in a PV/T System with Experimental and ANN's Models. *Sustain. Energy Grids Netw.* **2023**, *36*, 101189. [\[CrossRef\]](#)
13. Gao, Y.; Miyata, S.; Akashi, Y. Interpretable Deep Learning Models for Hourly Solar Radiation Prediction Based on Graph Neural Network and Attention. *Appl. Energy* **2022**, *321*, 119288. [\[CrossRef\]](#)
14. Choi, J.; Son, S.-W.; Lee, S.; Park, S. Advancing Global Solar Photovoltaic Power Forecasting with Sub-Seasonal Climate Outlooks. *Renew. Energy* **2024**, *237*, 121803. [\[CrossRef\]](#)
15. Wang, S.; Wang, Y.; Cheng, Y.; Sun, S.; Wang, N.; Yu, P.; Wang, S. An Improved Model for Power Prediction of PV System Based on Elman Neural Networks. In *Proceedings of the 2020 Asia Energy and Electrical Engineering Symposium (AEEES)*; IEEE: New York, NY, USA, 2020; pp. 902–907.
16. Elsaraiti, M.; Merabet, A.; Elsaraiti, M.; Merabet, A. A Comparative Analysis of the ARIMA and LSTM Predictive Models and Their Effectiveness for Predicting Wind Speed. *Energies* **2021**, *14*, 6782. [\[CrossRef\]](#)
17. Bashir, T.; Wang, H.; Tahir, M.; Zhang, Y. Wind and Solar Power Forecasting Based on Hybrid CNN-ABiLSTM, CNN-Transformer-MLP Models. *Renew. Energy* **2025**, *239*, 122055. [\[CrossRef\]](#)
18. Elsaraiti, M.; Merabet, A. Application of Long-Short-Term-Memory Recurrent Neural Networks to Forecast Wind Speed. *Appl. Sci.* **2021**, *11*, 2387. [\[CrossRef\]](#)
19. Liu, H.; Mi, X.; Li, Y. Wind Speed Forecasting Method Based on Deep Learning Strategy Using Empirical Wavelet Transform, Long Short Term Memory Neural Network and Elman Neural Network. *Energy Convers. Manag.* **2018**, *156*, 498–514. [\[CrossRef\]](#)
20. Iheanetu, K.J. Solar Photovoltaic Power Forecasting: A Review. *Sustainability* **2022**, *14*, 7005. [\[CrossRef\]](#)
21. Dhaked, D.K.; Dadhich, S.; Birla, D. Power Output Forecasting of Solar Photovoltaic Plant Using LSTM. *Green Energy Intell. Transp.* **2023**, *2*, 100113. [\[CrossRef\]](#)
22. Elsaraiti, M.; Merabet, A. Solar Power Forecasting Using Deep Learning Techniques. *IEEE Access* **2022**, *10*, 31692–31698. [\[CrossRef\]](#)
23. Sharadga, H.; Hajimirza, S.; Balog, R.S. Time Series Forecasting of Solar Power Generation for Large-Scale Photovoltaic Plants. *Renew. Energy* **2020**, *150*, 797–807. [\[CrossRef\]](#)
24. Zhang, Q.; Yan, H.; Liu, Y. Power Generation Forecasting for Solar Plants Based on Dynamic Bayesian Networks by Fusing Multi-Source Information. *Renew. Sustain. Energy Rev.* **2024**, *202*, 114691. [\[CrossRef\]](#)
25. Kumari, P.; Toshniwal, D. Deep Learning Models for Solar Irradiance Forecasting: A Comprehensive Review. *J. Clean. Prod.* **2021**, *318*, 128566. [\[CrossRef\]](#)
26. Jebli, I.; Belouadha, F.-Z.; Kabbaj, M.I.; Tilioua, A. Prediction of Solar Energy Guided by Pearson Correlation Using Machine Learning. *Energy* **2021**, *224*, 120109. [\[CrossRef\]](#)
27. Liu, X.; Liu, Y.; Kong, X.; Ma, L.; Besheer, A.H.; Lee, K.Y. Deep Neural Network for Forecasting of Photovoltaic Power Based on Wavelet Packet Decomposition with Similar Day Analysis. *Energy* **2023**, *271*, 126963. [\[CrossRef\]](#)

28. Lee, Y.H. Beyond the Shockley-Queisser Limit: Exploring New Frontiers in Solar Energy Harvest. *Science* **2024**, *383*, eado4308. [[CrossRef](#)] [[PubMed](#)]
29. Kopp, G.; Lean, J.L. A New, Lower Value of Total Solar Irradiance: Evidence and Climate Significance: FRONTIER. *Geophys. Res. Lett.* **2011**, *38*, L01706. [[CrossRef](#)]
30. Yaghoubi, E.; Yaghoubi, E.; Khamees, A.; Vakili, A.H. A Systematic Review and Meta-Analysis of Artificial Neural Network, Machine Learning, Deep Learning, and Ensemble Learning Approaches in Field of Geotechnical Engineering. *Neural Comput. Appl.* **2024**, *36*, 12655–12699. [[CrossRef](#)]
31. Xinxin, W.; Xiaopan, S.; Xueyi, A.; Shijia, L. Short-Term Wind Speed Forecasting Based on a Hybrid Model of ICEEMDAN, MFE, LSTM and Informer. *PLoS ONE* **2023**, *18*, e0289161. [[CrossRef](#)] [[PubMed](#)]
32. Kumar, I.; Tripathi, B.K.; Singh, A. Attention-Based LSTM Network-Assisted Time Series Forecasting Models for Petroleum Production. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106440. [[CrossRef](#)]
33. Cao, H.; Wang, T.; Chen, P.; Cheng, W.; Cao, Y.; Liu, Z. Solar Energy Forecasting in Short Term Based on the ASO-BPNN Model. *Front. Energy Res.* **2022**, *10*, 902486. [[CrossRef](#)]
34. Li, P.; Zhou, K.; Lu, X.; Yang, S. A Hybrid Deep Learning Model for Short-Term PV Power Forecasting. *Appl. Energy* **2020**, *259*, 114216. [[CrossRef](#)]
35. Sweet, L.; Müller, C.; Anand, M.; Zscheischler, J. Cross-Validation Strategy Impacts the Performance and Interpretation of Machine Learning Models. *Artif. Intell. Earth Syst.* **2023**, *2*, e230026. [[CrossRef](#)]

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