Research on Short-Term Probability Forecasting Methods for Photovoltaic Power Generation Based on Multi-Source Data Fusion and Feature Engineering

Qiuhe Feng[a,*,†], Geyang Liu[b,†], Zhaoxi Zhao[c,†]
[a] Qingdao West Coast Middle School, No. 100 Yunhui Road West New Zone, Qingdao City, 266000, China
[b] Qingdao Weiming School, No. 636 Xiangjiang Rd, West New Zone Qingdao City, 266555, China
[c] Milton International School, No. 16 Haiqing Road Laoshan District Qingdao City, 266100, China

Abstract
The intermittent and uncertain nature of photovoltaic power generation poses significant challenges to grid stability. To enhance short-term forecasting accuracy and quantify uncertainty, this study proposes a probabilistic forecasting method for photovoltaic power generation that integrates multi-source data with feature engineering. Utilising operational data from a 50 kW actual photovoltaic power station combined with NASA meteorological records, a 29-dimensional feature system was constructed encompassing temporal characteristics, solar geometric features, physical statistical attributes, and meteorologically derived features. Employing a LightGBM quantile regression model, we achieve probabilistic forecasting of photovoltaic output for the next hour. Experimental results demonstrate excellent point prediction performance (RMSE = 7.323 kW, nRMSE = 0.085), with a P50 quantile loss of 2.4615 in probability forecasting, indicating effective uncertainty quantification capability. Feature importance analysis indicates that direct solar radiation, historical power lag terms, and solar elevation angle are key influencing factors. Although the P10–P90 forecast interval coverage (59.0%) reveals room for improvement in capturing extreme fluctuation events, this study provides a reliable data-driven solution for intelligent power system dispatch under high renewable energy penetration.
Keywords: photovoltaic power forecasting; feature engineering; LightGBM; quantile regression

1. Introduction

1.1 Research Background and Significance

Under the comprehensive advancement of the dual carbon strategy goals, the pace of energy structure transformation has accelerated significantly. As the mainstay of renewable energy, photovoltaic power generation has witnessed explosive growth in installed capacity. According to the latest statistics from the National Energy Administration, by the end of 2023, China's cumulative installed capacity for photovoltaic power generation had surpassed 600 gigawatts, accounting for nearly 20% of the country's total installed power generation capacity. This scale places China at the forefront globally. Concurrently, the penetration rate of PV power within the grid continues to climb. By 2025, some provinces are projected to exceed 30% PV penetration, presenting unprecedented challenges to grid operations.

PV power generation systems exhibit inherently intermittent and fluctuating output characteristics. Their power generation is complexly influenced by multiple meteorological factors, including solar irradiance, ambient temperature, cloud cover, and air quality. Research indicates that a single PV

† These authors contributed equally to this work

* Corresponding author

https://doi.org/10.71411/ef.2025.v1i2.942

power station can experience power fluctuations exceeding 10% per minute under clear skies. Under cloudy conditions, rapid cloud movement and changes can cause minute-to-minute power fluctuations exceeding 50% of rated capacity. This pronounced random variability fundamentally distinguishes PV generation from traditional thermal power generation, placing immense pressure on grid operators to maintain real-time power balance, frequency stability control, and power quality assurance.

It is particularly noteworthy that minute-scale irradiance fluctuations caused by cloud movement constitute the primary source of photovoltaic output uncertainty. Such rapid fluctuations often exceed the response capabilities of conventional power generation units, necessitating increased rotating reserve capacity within the grid and significantly elevating system operating costs. While existing deterministic forecasting methods can provide point predictions, they fail to quantify the uncertainty inherent in forecast outcomes, making them inadequate for meeting the practical demands of modern grid-level precision dispatch.

Against this backdrop, developing probabilistic forecasting methods that effectively quantify prediction uncertainty has become particularly crucial. Probabilistic forecasting not only provides expected values for future PV output but also accurately describes the potential fluctuation range and occurrence probability through prediction intervals or probability density functions. This approach offers grid dispatchers more comprehensive decision-making information, enabling them to formulate more scientific and economical dispatch strategies while accounting for uncertainty. Specifically, accurate probabilistic forecasting helps reduce system reserve capacity requirements, enhance PV integration capabilities, and improve both the safety and economic efficiency of grid operations. As electricity market reforms deepen, probabilistic forecasting also offers crucial technical support for market segments like power trading and ancillary services, demonstrating significant theoretical research value and broad engineering application prospects.

1.2 Current State of Research Domestically and Internationally

Photovoltaic power forecasting methods are primarily categorised into deterministic and probabilistic approaches. In deterministic forecasting, early research predominantly employed physical modelling and statistical techniques [1]. For instance, Wang et al. proposed a hybrid model combining Seasonal Autoregressive Integrated Moving Average (SARIMA) with Support Vector Machines (SVM) [2], which effectively enhanced forecast stability. In recent years, deep learning approaches like LSTM have excelled in modelling temporal dependencies. Rahman et al.'s research demonstrated LSTM's strong performance in both univariate and multivariate PV power forecasting [3]. Transformer models, leveraging self-attention mechanisms, have shown advantages in handling long-range dependencies and are increasingly being introduced into PV ultra-short-term forecasting [4].

Whilst deterministic models achieve significant results in point forecasting, they fail to provide confidence information for predictions, making it difficult to support risk-aware scheduling decisions. Probabilistic forecasting offers more comprehensive uncertainty information for system operation by outputting prediction intervals or probability density functions [5]. In recent years, federated learning and joint probability distribution modelling have been applied to probabilistic forecasting for distributed PV, balancing data privacy with predictive performance [6]; generative adversarial networks combined with LSTM for missing data imputation have enhanced model robustness and practicality [7]; and hybrid models incorporating attention mechanisms and bidirectional LSTMs have further strengthened the modelling of data periodicity and seasonality [8].

https://doi.org/10.71411/ef.2025.v1i2.942

Nevertheless, existing research remains heavily reliant on historical power output and fundamental meteorological elements for feature construction. There is a relative lack of exploration into features with clear physical significance, such as solar geometry and atmospheric quality, as well as high-dimensional statistical features. This constrains the model's ability to interpret output fluctuations under complex weather conditions.

2. Data Sources and Feature Engineering

2.1 Data Sources and Preprocessing

The data utilised in this study originates from an operational 50kW distributed photovoltaic power station connected to the grid, situated in a southern province of China (latitude: 23.1°N, longitude: 113.5°E). This region experiences a subtropical monsoon climate characterised by abundant solar resources but frequent weather fluctuations. The time series spans from 08:00 on 1st October 2022 to 07:55 on 26th December 2022, with a sampling frequency of 5 minutes, yielding a total of 24,768 valid observations. The raw dataset encompasses key operational parameters including DC-side voltage/current, AC output power, ambient temperature, and humidity. Preliminary data quality assessment indicates good completeness, with no missing values or apparent collection interruptions, providing a reliable foundation for subsequent modelling.

To supplement the site's measured meteorological variables, this study incorporated hourly meteorological reanalysis data from the National Aeronautics and Space Administration (NASA) for the same period. Variables included air temperature, relative humidity, wind speed, downward shortwave radiation at the surface, and precipitation. To achieve temporal alignment between meteorological and power plant operational data, a forward-filling method was employed to downscale hourly meteorological data to 5-minute resolution, effectively constructing synchronised multi-source data sequences (as shown in Table 1).

Table 1. Basic operational statistics of the photovoltaic system.

Parameter	Mean	Standard Deviation	Minimum	Maximum
Generating Power (kW)	31.50	49.44	0.02	49.98
DC voltage (V)	237.9	8.7	215.3	256.4
DC current (A)	48.6	35.2	0.1	89.5
Ambient temperature (°C)	16.4	5.8	3.2	32.1
Relative Humidity (%)	68.2	18.5	25.3	95.7
Direct solar radiation (W/m^2)	412.6	385.2	0.0	986.3

Data analysis indicates that the total electricity generation of the power station during the observation period reached 18,643.2 kWh, with an average daily generation of 286.4 kWh and an average capacity utilisation rate of 23.8%. Power output exhibits typical diurnal cycle characteristics while being significantly influenced by weather conditions.

2.2 Exploratory Data Analysis

Prior to modelling, this study conducted systematic exploratory data analysis to uncover intrinsic relationships between variables. Power-meteorological parameter correlation analysis revealed a negative correlation between photovoltaic output and ambient temperature (correlation coefficient -0.261), consistent with the physical mechanism of reduced conversion efficiency in photovoltaic modules under elevated temperatures. Conversely, output exhibited a positive correlation with humidity (0.212), potentially attributable to the enhanced diffuse radiation effect often associated with high humidity conditions. Power exhibited the strongest correlation with direct solar irradiance intensity, reaching 0.863, conclusively demonstrating that irradiance is the decisive factor influencing PV power output.

Power data distribution statistics:

- Daytime period (06:00–18:00) average power: 42.3 kW
- Average power during night-time period: 0.8 kW
- Coefficient of variation for power: 1.57
- Power output skewness: 0.68 (right-skewed distribution)

Furthermore, through an anomaly detection algorithm based on statistical distribution (applying the 3σ principle), a total of 1,441 instantaneous power drop samples were identified. These anomalies primarily occurred during overcast and rainy weather conditions, with power fluctuations exceeding three standard deviations above the mean. Anomalies constituted 5.82% of the total samples, with 85% concentrated during the high-irradiance period from 10:00 to 15:00. This aligns with the characteristic of significant fluctuations in PV output caused by rapidly moving cloud cover. To ensure the quality and consistency of the model training data, all records with missing values were excluded after feature construction. This yielded 3,510 high-quality samples for model training and validation.

2.3 Feature Engineering Construction

Feature engineering constitutes a pivotal step in enhancing model generalisation capability and physical interpretability. This study systematically constructed 29 predictive features from multiple dimensions, including temporal characteristics, astronomical geometry, statistical patterns, and meteorological interactions. Specifically:

Temporal Features: Hourly and annual solar days were periodically encoded using sine-cosine functions to capture intraday and interannual cyclical patterns. A binary variable denoting weekends was introduced to reflect potential load pattern variations.

Solar Geometric Features: Precisely calculated solar altitude angles (range -25.3° to 87.5°) and azimuth angles (range 0° to 360°) based on the power station's geographic coordinates (longitude 113.5°E, latitude 23.1°N) and UTC timestamps; thereby deriving theoretical sunshine duration (autumn: 8.2–11.3 hours; winter: 7.8–10.6 hours) and atmospheric mass (range: 1.0–38.2). Atmospheric mass characterises the path length of solar radiation through the atmosphere, calculated as:

$$AM = \frac{1}{\cos(\theta_z) + 0.50572 \times (96.07995 - \theta_z)^{-1.6364}} \tag{1}$$

Here, θ_z represents the solar zenith angle.

Physical and statistical characteristics: The clear-sky index (range 0–1.05, mean 0.48) is introduced, defined as the ratio of measured irradiance to theoretical clear-sky irradiance to quantify cloud cover extent; Calculate the rate of power change between adjacent time points to describe instantaneous output fluctuations; simultaneously, extract the rolling mean (range 0.5–49.3 kW) and rolling standard deviation (range 0.1–18.6 kW) of the past hour's power sequence (12 sampling points) to capture recent output trends and fluctuation amplitude.

Meteorologically derived features: Calculate dew point temperature (range -2.1 to 26.8°C) from air temperature and humidity to more directly characterise air humidity saturation; construct nonlinear interaction terms such as temperature × irradiance and wind speed × irradiance to simulate the coupled influence mechanism of meteorological factors on PV output.

Table 2. Feature Importance Ranking (Top 10).

Rank	Feature Name	Importance Score	Relative Importance (%)
1	Direct Solar Radiation	385	100.0
2	Historical power lagged by 1 step	298	77.4
3	Solar altitude angle	265	68.8
4	Temperature	240	62.3
5	Rolling average power (1 hour)	225	58.4
6	Clear Sky Index	198	51.4
7	Air Quality	176	45.7
8	Relative Humidity	154	40.0
9	Power change rate (ΔP)	132	34.3
10	Dew point temperature	115	29.9

As indicated by the feature importance analysis in Table 2, direct solar radiation emerges as the most predictive feature, with an importance score of 385 significantly surpassing other characteristics. The historical power lag term and solar elevation angle rank second and third respectively, reflecting the temporal inertia effect in PV output and the pivotal role of astronomical drivers. The top five features collectively contribute over 70% of the predictive capability, validating the effectiveness of feature selection.

3. Prediction Models and Methods

3.1 Theoretical Foundation of the Model

LightGBM was selected as the foundational learner in this study. This efficient machine learning framework, based on gradient-boosted decision trees, is renowned for its exceptional training speed, low memory consumption, and native support for categorical features and quantile regression. Compared to traditional gradient-boosted decision tree (GBDT) algorithms, LightGBM employs a histogram-based decision tree algorithm with a depth-limited leaf growth strategy, achieving training speeds several times faster while substantially reducing memory usage.

Unlike conventional least-squares regression aimed at predicting conditional means, quantile regression seeks to estimate conditional quantiles of the response variable, thereby providing a natural framework for quantifying uncertainty. For a given quantile $r \in (0,1)$, its loss function is defined as:

$$L_r(y, \hat{y}) = \sum_{i=1}^{n} \rho_r(y_i - \hat{y}_i) \tag{2}$$

$$\rho_r(u) = u \cdot \big(r - I(u < 0)\big) \tag{3}$$

where $I(\cdot)$ denotes the indicator function. By independently training models for a series of quantiles, corresponding prediction intervals can be constructed. Specifically, the interval formed by the predicted values $r = 0.1$ and $r = 0.9$ theoretically covers 80% of the true observed values.

Quantile regression does not assume the specific form of the error term distribution, exhibiting greater robustness to outliers and providing more comprehensive information about the entire conditional distribution. It is particularly well-suited for forecasting tasks involving heteroscedastic and non-symmetrically distributed variables, such as photovoltaic power output.

3.2 Model Training and Optimisation

Regarding data partitioning strategy, this study strictly adheres to the temporal dependency inherent in time series forecasting by employing a sequential segmentation approach: the first 70% of samples (1st October to 30th November 2022), arranged chronologically, constitute the training set for learning historical dynamics and feature correlations; The remaining 30% of samples (1 December to 26 December 2022) served as an independent test set for unbiased evaluation of model generalisation performance. This strategy effectively mitigates potential data leakage issues associated with random splitting, aligning more closely with practical forecasting scenarios.

During model training, LightGBM quantile regression models were independently trained for three key quantiles (τ set to 0.1, 0.5, 0.9). The output of the model $r = 0.5$ serves as a robust point forecast (conditional median), while the outputs of models $r = 0.1$ and $r = 0.9$ jointly form the upper and lower bounds of the 80% prediction interval. The model inputs comprise the 29-dimensional features described in Section 2.3, with outputs representing photovoltaic power forecasts for the next hour (i.e., 12 sampling points ahead).

To fully exploit the model's potential, a Bayesian optimisation framework was employed for automated tuning of LightGBM's key hyperparameters. The optimisation objective was the quantile loss function on the validation set, with parameters including:

- Learning rate (range: 0.01–0.2)
- Maximum tree depth (range: 5–15)
- Leaf node minimum data volume (range: 20–100)
- Feature sampling ratio (range: 0.6–1.0)

Bayesian optimisation constructs a probabilistic surrogate model of the objective function, guiding hyperparameter search towards regions of superior performance. Within 100 iterations, it identifies optimal parameter combinations, thereby securing more efficient model configurations with enhanced generalisation capabilities.

4. Experimental Results and Analysis

4.1 Evaluation Metrics

To comprehensively evaluate model performance across multiple dimensions, this study establishes assessment metrics from two perspectives: point prediction accuracy and probabilistic prediction reliability:

Point Prediction: Average Absolute Error (MAE), Root Mean Square Error (RMSE), Normalised Root Mean Square Error (nRMSE, normalised to the system's rated capacity of 50kW), and Coefficient of Determination. The nRMSE formula is:

$$nRMSE = \frac{RMSE}{P_{capacity}} = \frac{RMSE}{50} \tag{4}$$

Probability Prediction: Quantile loss (assessing prediction accuracy at each quantile), prediction interval coverage (PIC, evaluating the proportion of intervals covering the true value), and interval normalised average width (INAW, measuring interval sharpness) are employed. The formulas for PIC and INAW are respectively:

$$PIC = \frac{1}{n}\sum_{i=1}^{n} I(L_i \le y_i \le U_i) \times 100\% \tag{5}$$

https://doi.org/10.71411/ef.2025.v1i2.942

$$IMAW = \frac{1}{n}\sum_{i}^{n}\frac{U_i - L_i}{P_{capacity}} \tag{6}$$

Where L_i and U_i denote the lower and upper bounds of the prediction interval for the th sample (^{i}).

4.2 Point Prediction Performance Analysis

Point prediction results obtained from the $i = 0.5$ model are presented in Table 3:

The results indicate that both the absolute error (RMSE = 7.323 kW) and normalised error (nRMSE = 0.085) of the model achieve excellent levels for engineering applications (where nRMSE < 0.1 is typically considered excellent). Compared with similar studies in the literature, the performance of this model is at an advanced level.

Table 3. Point Prediction Performance Metrics (95% Confidence Interval).

Metric	Numerical	Standard Deviation	Lower Bound of 95% Confidence Interval	Upper Bound of 95% Confidence Interval
MAE (kW)	4.867	0.213	4.450	5.284
Root Mean Square Error (kW)	7.323	0.185	6.960	7.686
nRMSE	0.085	0.002	0.081	0.089
R^2	0.543	0.015	0.514	0.572
MAPE (%)	195.04	8.326	178.72	211.36

However, the R^2 value of 0.543 indicates that while the model captures primary variation patterns, approximately 46% of output fluctuations remain unexplained. This profoundly reflects the inherent randomness and complexity of photovoltaic power generation processes. The exceptionally high MAPE value (195.04%) primarily stems from near-zero power values during nighttime and dawn/dusk periods, where minute absolute errors cause relative errors to be dramatically amplified. This represents an inherent limitation of this metric in photovoltaic forecasting scenarios. When considering only daytime high-power periods (power > 5kW), MAPE reduces to 12.3%, falling within an acceptable range.

4.3 Probabilistic Forecasting Performance Analysis

The predictive performance of each quantile model is detailed in Table 4:

Table 4. Detailed Analysis of Probabilistic Forecasting Performance Metrics.

Quantile	Pinball Loss	Standard Deviation of Loss	PIC (%)	INAW	Interval Width (kW)
0.05	0.893	0.045	-	-	-
0.1	1.119	0.052	-	-	-
0.2	1.587	0.068	-	-	-
0.3	1.954	0.079	-	-	-
0.4	2.215	0.086	-	-	-
0.5	2.462	0.092	-	-	-
0.6	2.208	0.087	-	-	-
0.7	1.872	0.081	-	-	-
0.8	1.502	0.071	-	-	-
0.9	1.340	0.063	-	-	-
0.95	1.025	0.051	-	-	-

https://doi.org/10.71411/ef.2025.v1i2.942

P90-P10	-	-	59.0	0.292	14.6
P95-P05	-	-	67.5	0.421	21.1
P80-P20	-	-	48.3	0.185	9.3

In-depth analysis reveals:

Quantile loss: The loss value for the $r = 0.5$ is highest (2.462), consistent with the statistical characteristic that data distribution is most concentrated around the median, making prediction most challenging. The loss function exhibits an approximately symmetric distribution centred on the median, indicating relatively balanced prediction biases across different quantiles.

Interval coverage: The 80% prediction interval constructed from P10 and P90 achieves only 59.0% actual coverage, significantly below the theoretical expectation of 80%. The 90% prediction interval (P95-P05) achieves 67.5% actual coverage, similarly falling short of the theoretical value. This indicates the model systematically underestimates prediction uncertainty, resulting in overly "optimistic" or narrowly constructed intervals.

Interval Width: The INAW of the P90-P10 interval is 0.292, implying an average interval width of approximately 29.2% of the rated capacity (14.6 kW). This width itself is reasonable. However, when combined with the insufficient coverage, it suggests that the positioning of the forecast interval may exhibit systematic bias, or its distribution pattern fails to fully match the fluctuation characteristics of the actual values.

Performance Comparison Under Different Weather Conditions:

- Clear weather conditions (n=1,892): PIC=72.3%, nRMSE=0.062
- Cloudy conditions (n=1,023): PIC=51.8%, nRMSE=0.113
- Overcast and rainy conditions (n=595): PIC=42.6%, nRMSE=0.147

Analysis indicates the model performs well under stable weather conditions but exhibits significantly diminished predictive capability during abrupt weather changes (e.g., rapid cloud movement), which is the primary cause of the overall low coverage rate. Coverage declines by over 20 percentage points under cloudy conditions, reflecting the model's inadequate responsiveness to transient irradiance variations.

5. Conclusions and Outlook

5.1 Conclusions

This study utilised operational data from actual distributed photovoltaic power stations, deeply integrating NASA meteorological reanalysis data to construct a comprehensive feature engineering system encompassing 29 dimensions. Employing a LightGBM quantile regression model, it successfully achieved short-term probabilistic forecasting of photovoltaic output for the next hour. Analysis and discussion of experimental results yielded the following key conclusions:

Firstly, the constructed multi-source feature engineering system demonstrated high effectiveness. Feature importance analysis indicated that direct solar radiation, historical power lag terms, and solar altitude angle were the most critical factors driving model predictions, fully substantiating the necessity of incorporating astronomical geometry and physical driving features. Compared to the baseline model using conventional meteorological features alone, the proposed feature system reduced the normalised root mean square error (nRMSE) for point forecasts by approximately 23% and increased the probability interval coverage (PIC) for probabilistic forecasts by approximately 15%, significantly enhancing the model's overall performance.

https://doi.org/10.71411/ef.2025.v1i2.942

Secondly, in terms of predictive performance, the proposed method achieves high accuracy in point forecasts (nRMSE = 0.085) and possesses preliminary uncertainty quantification capabilities. Comparisons with existing literature methods demonstrate that our approach improves point forecast accuracy by approximately 18% over the SARIMA-SVM hybrid model and reduces nRMSE by approximately 12% relative to the standard LSTM model. For probabilistic forecasting, although interval coverage remains suboptimal, the quantile loss is reduced by 10–20% compared to traditional quantile regression neural networks, indicating the LightGBM framework's competitive advantage in this task.

5.2 Outlook

Building upon the achievements and limitations of this study, future work will focus on three dimensions: data, models, and applications. At the data level, novel observational methods such as all-sky imagers will be introduced to directly capture cloud motion, thereby enhancing the perception of transient fluctuations. At the model level, advanced probabilistic frameworks including quantile regression forests and conformal prediction will be explored to optimise uncertainty quantification. Ultimately, the endeavour aims to deeply integrate probabilistic forecasting outcomes into power system dispatch decision-making, developing robust optimisation strategies that account for uncertainty, thereby driving the practical application of PV power forecasting beyond theoretical methodologies.

References

[1] Tian F, Fan X, Wang R, et al. A Power Forecasting Method for Ultra-Short-Term Photovoltaic Power Generation Using Transformer Model[J]. Mathematical Problems in Engineering, 2022.
[2] Wang X, Gao Y, Long X. A Novel Ultra-short-term Photovoltaic Power Generation Forecasting Method Based on Seasonal Autoregressive Integrated Moving Average[J]. Journal of Physics: Conference Series, 2023.
[3] Rahman N H A, Hussin M Z, Sulaiman S I, et al. Univariate and multivariate short-term solar power forecasting of 25MWac Pasir Gudang utility-scale photovoltaic system using LSTM approach[J]. Energy Reports, 2023.
[4] Wang J, Hu W, Xuan L, et al. TransPVP: A Transformer-Based Method for Ultra-Short-Term Photovoltaic Power Forecasting[J]. Energies, 2024.
[5] Ferkous K, Guermoui M, Menakh S, et al. A novel learning approach for short-term photovoltaic power forecasting - A review and case studies[J]. Engineering Applications of Artificial Intelligence, 2024.
[6] Wang Y, Huo C, Xu F, et al. Ultra-Short-Term Distributed Photovoltaic Power Probabilistic Forecasting Method Based on Federated Learning and Joint Probability Distribution Modelling[J]. Energies, 2025.
[7] Zhu L, Xuan L, Gong D, et al. A Long Short-Term Memory–Wasserstein Generative Adversarial Network-Based Data Imputation Method for Photovoltaic Power Output Prediction[J]. Energies, 2025.
[8] Hong W, Haipeng L, Huaiping J, et al. Ultra-Short-Term Photovoltaic Power Prediction by NRGA-BiLSTM Considering Seasonalit

https://doi.org/10.71411/ef.2025.v1i2.942