EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

Photovoltaic Power Adaptive Hybrid Forecasting Model Integrated with Multi-Dimensional Error Compensation

Shuye Liu[1], Bingkuan Gong[1], Cuiyu Cui[1, *], Kun Zang[1]

[1] Xizang Agricultural and Animal Husbandry University School of Electrical Engineering and Automation, Xizang, 860000, China

Abstract

INTRODUCTION: With the deepening of China's "Dual Carbon" goals, the scale of photovoltaic (PV) installations continues to expand, placing stringent requirements on the accuracy, scenario adaptability, and robustness of day-ahead power prediction for PV plants in grid frequency regulation, peak shaving, and electricity market trading. However, existing single prediction models have obvious limitations: linear models struggle to capture nonlinear power disturbances under cloudy days with sudden irradiance changes and extreme weather; nonlinear models are prone to overfitting during stable, high-irradiance sunny periods, leading to redundant accuracy; and most methods lack sufficient robustness against meteorological fluctuations and data noise, resulting in large prediction errors that seriously affect the economy and security of grid operation.

OBJECTIVES: Aiming at the problems in PV plant prediction where a single model finds it difficult to balance linear patterns and nonlinear disturbances, and prediction accuracy is greatly affected by meteorological fluctuations and data noise, this paper proposes a hybrid prediction model based on Linear Regression and Gradient Boosting Tree, along with a multi-dimensional error compensation mechanism.

METHODS:

1. Otptimize feature engineering design, selecting time features, historical power lag terms, and meteorological interaction features as inputs, unifying their scales via Z-score standardization to simplify model complexity and construct a high-quality training set.
2. Design a dynamic weight allocation strategy based on irradiance intensity grading and intraday time periods, integrating the precise fitting advantage of Linear Regression during high-irradiance periods with the nonlinear fluctuation capture capability of Gradient Boosting Tree to establish a hybrid prediction model.
3. Use hourly operational data from a specific PV plant from May 2024 to March 2025 as a sample, performing validation combining an error compensation mechanism composed of sliding window error correction, extreme weather compensation, and residual feedback.

RESULT: The test set RMSE of the hybrid model decreased by 18.2% and 22.5% compared to the single Linear Regression and Gradient Boosting Tree models respectively, with the error deviation rate under extreme weather controlled within 12%.

CONCLUSION: These results verify the effectiveness and practicality of the proposed hybrid prediction model and error compensation method.

Keywords: Power prediction; Gradient Boosting Tree; Error compensation; Historical deviation pattern

Received on 23 September 2025, accepted on 19 December 2025, published on 28 April 2026

doi: 10.4108/ew.12813

---

[*] Email: 18903512681@163.com

1. Introduction

Against the backdrop of China's "Dual Carbon" goals, photovoltaic power generation has become a key driver in the renewable energy transition, with large-scale grid-connected PV power generation emerging as an inevitable trend [1]. Endowed with clean, scalable, and easily accessible attributes, PV power supports low-carbon energy system development, yet its output exhibits high volatility and randomness due to solar irradiance, temperature, and cloud cover—unplanned grid integration of such power risks destabilizing the grid [2].

PV plants are inherently intermittent (generating only during daytime) and weather-sensitive, making timely, accurate PV power prediction critical for grid dispatching and plant operation [3][4]. PV power prediction is classified into ultra-short-term, short-term, and medium-and long-term types based on duration [5]; its accuracy is notably impacted by weather, with much lower errors under clear skies than non-clear skies [6].

Globally, institutions offer mature PV prediction and solar resource assessment services with advanced software. Domestically, PV prediction has been integrated into power systems, though research lags slightly behind developed countries but is advancing rapidly [7][8].

Academic studies on PV prediction provide valuable references for this work: Multi-scale data decomposition and architecture optimization inform the multi-dimensional feature fusion in our feature engineering[9]; Ultra-short-term PV prediction methods' exploration of dynamic data adaptation inspires our focus on feature robustness under data drift [10]; Scenario classification logic in distributed PV prediction's collaborative training supports our differentiated modeling for extreme and transitional weather [11]; Technical paths of extreme-weather PV prediction guide the build of our extreme-weather error compensation module[12]; Attention mechanisms in distributed ultra-short-term probabilistic PV prediction inspire our meteorological-power feature interaction modeling [13]; Analysis of short-term PV prediction methods and application scenarios clarifies our multi-scenario research direction by identifying existing methods' poor scenario adaptability [14]; Multi-scenario feature analysis from new power system risk reviews under "Dual Carbon" goals offers cross-domain insights for PV prediction's hierarchical scenario adaptation [15].

Nevertheless, gaps remain: Short-term nonparametric probabilistic PV prediction, while accurate, is overly complex and lacks robustness [13]; Multi-task learning-based short-term PV prediction, despite effective collaboration, lacks dynamic feature weighting for transitional weather [15]; Transferable PV prediction reduces data dependence but fails to capture intraday high-frequency power's temporal correlation in day-ahead prediction [14]; Distributed PV cluster power probabilistic prediction offers a new cluster-focused direction but overlooks single-station multi-scenario accuracy balance [15].

To address these limitations, this paper identifies three core issues: single models cannot balance PV output's linear trends and nonlinear disturbances; existing methods adapt poorly to complex scenarios; and some high-precision methods are computationally expensive. Accordingly, we propose a hybrid model and a multi-dimensional error compensation mechanism. Via feature engineering optimization, dynamic weight fusion, and multi-module error correction, we achieve "accurate linear fitting, strong nonlinear adaptability, and low complex-scenario errors" to support reliable day-ahead PV power prediction.

2. Data Preprocessing and Feature Engineering

2.1 Data Source and Parameters

This paper selects actual operational data from a specific PV plant for model training and validation. The involved parameters include historical PV power generation data, irradiance intensity, ambient temperature, relative humidity, wind speed, and time features such as hour and month. To filter input parameters strongly correlated with PV power, the Pearson correlation coefficient between each parameter and PV power generation is calculated, retaining parameters with significant correlation. The Pearson correlation coefficient calculation formula is:

$$r = \frac{\sum_{i=1}^{m}(x_i - \overline{x})(t_i - \overline{t})}{\sqrt{\sum_{i=1}^{m}(x_i - \overline{x})^2 \sum_{i=1}^{m}(t_i - \overline{t})^2}} \tag{1}$$

m --- Number of parameter data groups;
X_i--- The *i*-th parameter to be analyzed;
t_i--- The *i*-th actual PV power generation;
$\bar{\boldsymbol{x}}$, $\bar{\boldsymbol{t}}$ ---Mean values of the corresponding parameter and power.

After calculation, the Pearson correlation coefficients between PV power generation and various parameters are shown in Table 1.

Table 1. Pearson correlation coefficient analysis of PV power influencing factors

No.	Parameter Type	Pearson Correlation Coefficient
1	Irradiance Intensity	0.92
2	Ambient Temperature	0.65
3	Relative Humidity	-0.58
4	Wind Speed	0.12
5	Hour of Day	0.88
6	Same Period Power	0.95

7	Irradiance Intensity × Temperature	0.90

The larger the absolute value of the Pearson correlation coefficient, the stronger the correlation between the parameter and PV power. Therefore, this paper retains parameters such as irradiance intensity, ambient temperature, hour, same period power from the previous day, and irradiance intensity × temperature, providing highly correlated input samples for the subsequent model fusion and error compensation mechanism.

2.2 Data Cleaning Strategy

PV plant operational data are susceptible to abnormal values, missing values, and data noise during the collection and transmission processes due to factors like equipment failure and communication interference. Using such data directly for model training would significantly reduce prediction accuracy. Therefore, this paper employs a multi-dimensional data cleaning strategy to improve input data quality, providing support for the effectiveness of the model fusion and error compensation mechanism.

For outlier identification, the 3σ criterion is adopted, with the formula:

$$O=\left\{x_i \,\middle|\, |x_i-\bar{x}|>3\sigma\right\} \tag{2}$$

Where O is the set of outliers, xi is the i-th data point to be detected, $\bar{\mathbf{x}}$ is the data mean, and σ is the data standard deviation. For identified outliers, considering the temporal continuity and meteorological correlation of PV output, a weighted average method using adjacent normal data is applied for correction. The formula is:

$$x_{i'}=\frac{w_{i-1}x_{i-1}+w_{i+1}x_{i+1}}{w_{i-1}+w_{i+1}} \tag{3}$$

Where xi’ is the corrected data, w_{i-1}, w_{i+1} are the weights of the data from the previous and next time steps.

For missing values, the linear interpolation method is used for imputation. The formula is:

$$x_m=x_a+\frac{m-a}{b-a}(x_b-x_a) \tag{4}$$

Where x_m is the interpolated data at missing time m, x_a, x_b are the normal data from the immediately preceding and following time steps surrounding the missing period (a<m<b).

Using the above strategies, the original data from a specific PV plant from May 2024 to March 2025 were processed. A total of 127 outliers were identified and corrected, and 245 missing values were filled via interpolation, ensuring data continuity and reliability, and providing high-quality input samples for the subsequent Linear Regression - Gradient Boosting Tree hybrid model and error compensation mechanism.

3. Hybrid Prediction Model and Error Compensation Mechanism

3.1 Adaptive Weighted Hybrid Model

In PV power prediction, a single model is difficult to meet the high-precision requirements for both linear and nonlinear scenarios. Therefore, it is necessary to construct a hybrid model based on the complementary characteristics of different models.

Analysis of Single Model Characteristics

Linear Regression Model: Possesses outstanding ability to fit linear relationships. During high irradiance periods at PV plants, the generated power shows a significant linear correlation with irradiance intensity and time features, allowing accurate capture of power trends in such scenarios. However, the Linear Regression model struggles to capture nonlinear fluctuations in PV power, leading to insufficient prediction accuracy in nonlinear scenarios such as cloudy days and extreme weather. Its core formula is:

$$P_{LR}(t)=\beta_0+\sum_{j=1}^{k}\beta_j X_j(t)+\varepsilon(t) \tag{5}$$

Gradient Boosting Tree Model: As an ensemble learning method, it fits residuals by iteratively constructing multiple decision trees, enabling accurate capture of nonlinear fluctuations in PV power. However, the Gradient Boosting Tree model is prone to overfitting in high-irradiance linear scenarios, and has relatively poor interpretability and high computational complexity. Its core formula is:

$$P_{\mathrm{GBT}}(t)=F_m(X(t))=F_{m-1}(X(t))+\gamma_m h_m(X(t)) \tag{6}$$

In summary, Linear Regression and Gradient Boosting Tree exhibit a "linear fitting - nonlinear capture" complementarity in PV power prediction, providing a theoretical basis for constructing the hybrid model.

Adaptive Weighted Hybrid Model

Since the prediction performance of Linear Regression and Gradient Boosting Tree varies significantly across different irradiance intensities and intraday time periods, traditional fixed-weight fusion strategies lack robustness in complex scenarios. Therefore, this paper proposes an adaptive weighted fusion strategy. The core formula is:

$$P_f(t)=w(t)\cdot PLR(t)+(1-w(t))\cdot P_{\mathrm{GBT}}(t) \tag{7}$$

Where w(t) is the adaptive weight of the Linear Regression model at time t, 0≤w(t)≤1 Its calculation is divided into two steps:

Basic Weight Calculation: Determined based on the Pearson correlation coefficient between the model's predicted values and the actual values. The formula is:

$$w_0(t)=\frac{r_{\mathrm{LR}}(t)}{r_{\mathrm{LR}}(t)+r_{\mathrm{GBT}}(t)} \tag{8}$$

Where $r_{LR}(t)$, $r_{GBT}(t)$ are the Pearson correlation coefficients between the predicted values of the Linear Regression and Gradient Boosting Tree models and the actual values at time t, respectively.

Scenario Correction Coefficient Calculation: Dynamically adjusted based on the scenario deviation of irradiance intensity and temperature. The formula is:

$$\delta(t)=\alpha\cdot\frac{|I(t)-I_{\mathrm{ref}}|}{I_{\max}-I_{\min}}+\beta\cdot\frac{|T(t)-T_{\mathrm{ref}}|}{T_{\max}-T_{\min}} \tag{9}$$

Where α, β are the correction weights for irradiance intensity and temperature; $I(t)$, $T(t)$ are the irradiance intensity and ambient temperature at time t; I_{ref}, T_{ref} are the reference values for irradiance intensity and temperature; $I_{\max}$, $I_{\min}$, $T_{\max}$, $T_{\min}$ are the historical extreme values of the corresponding features. The final adaptive weight is:

$$w(t)=w_0(t)\cdot(1+\delta(t)) \tag{10}$$

This fusion strategy, through two-dimensional weight regulation of "model correlation + scenario features", enhances the adaptive capability for both high-irradiance linear scenarios and low-irradiance nonlinear scenarios. It dynamically adjusts the weight proportion of Linear Regression and Gradient Boosting Tree, enabling the hybrid model to leverage the performance advantages of the optimal sub-model in different scenarios. Combined with the subsequent error compensation mechanism, a high-precision day-ahead power prediction model for PV plants can be constructed.

3.2 Scenario-Driven Multi-Dimensional Error Compensation Model

To further reduce the residual error of the Linear Regression - Gradient Boosting Tree hybrid model in complex meteorological scenarios and periods of sudden power changes, and to enhance the robustness of day-ahead PV power prediction, this paper constructs a multi-dimensional error compensation mechanism combining "sliding window error correction + extreme weather compensation + residual feedback iteration", based on the temporal continuity and scenario sensitivity characteristics of PV output. Through the collaborative operation of multiple modules, precise error cancellation is achieved. The specific design is as follows:

Sliding window error correction relies on the short-term temporal correlation of PV power, using a 24-hour sliding window to capture local error patterns------taking the current prediction time t as the window endpoint, selecting the hybrid model's predicted values Pf(k) and actual power values Pact(k) from the previous 24 hours to construct a residual sequence, calculating the mean residual within the window to represent short-term systematic bias, and then adjusting the prediction result through a correction coefficient optimized via grid search. The core formulas are:

$$e_{\mathrm{sw}}(t)=\frac{1}{24}\sum_{k=t-23}^{t}\left(P_{\mathrm{act}}(k)-P_f(k)\right) \tag{11}$$

$$P_{f1}(t)=P_f(t)-\lambda_{\mathrm{sw}}\cdot e_{\mathrm{sw}}(t) \tag{12}$$

Where, $e_{sw}(t)$ is the mean sliding window residual at time t, $\lambda_{sw}=0.85$ is the correction coefficient. Verified using hourly data from a PV plant from May 2024 to March 2025, this module can effectively offset short-term systematic bias in stable scenarios, reducing hourly power fluctuation error by 3.2kW and compressing the residual standard deviation by 28.6%, laying the accuracy foundation for subsequent compensation steps. Extreme weather compensation is designed for scenarios prone to prediction deviation, such as sudden irradiance changes and temperature extremes, by setting meteorological thresholds to identify extreme conditions: irradiance intensity sudden change threshold ≥300W/m²·h, temperature extreme range <5℃ or >35℃. Based on a historical extreme weather sample library from the same period for the aforementioned plant, a "weather type - error offset" mapping relationship is established, dynamically generating compensation factors for different extreme scenarios. The correction formulas are:

$$c_{\mathrm{ext}}(t)=\omega\cdot e_{\mathrm{hist}}(t) \tag{13}$$

$$P_{f2}(t)=P_{f1}(t)-c_{\mathrm{ext}}(t) \tag{14}$$

Where, $e_{hist}(t)$ is the average residual under the same extreme scenario in the historical same period, ω is the scenario adaptation weight. Experimental results show that this module can reduce the error deviation rate under extreme weather from 18.7% to 11.8%, and reduce the prediction error by 5.3kW during power sudden change periods like 14:00 on cloudy days, meeting the core requirement for prediction accuracy under extreme scenarios for the grid. Residual feedback iteration feeds the final residual $e(t)=P_{act}(t)-P_{f2}(t)$, after sliding window correction and extreme weather compensation, back to the weight calculation step of the adaptive weighted hybrid model. By dynamically adjusting the weight proportion of Linear Regression and Gradient Boosting Tree, a closed loop of "prediction - correction - optimization" is formed. The feedback logic is: if $e(t)>0$, increase the weight of the GBT model, which has stronger nonlinear fluctuation capture capability; if $e(t)<0$, increase the weight of the LR model, which has more accurate linear trend fitting. The iteration formula is:

$$w_{\mathrm{new}}(t)=w(t)+\lambda_{\mathrm{fb}}\cdot\frac{e(t)}{P_{\mathrm{act}}(t)} \tag{15}$$

In the formula, λ_{fb}=0.05 is the feedback step size. This module optimizes model adaptability through real-time iteration, further reducing the overall prediction residual of the hybrid model by 1.2 kW and enhancing the model's adaptive capability under different scenarios. The three modules of the multi-dimensional error compensation mechanism work synergistically to form a complete error control chain of "short-term correction-scenario adaptation-dynamic iteration": sliding window correction addresses systematic errors in stable scenarios, extreme weather compensation handles sudden power fluctuation deviations in unexpected scenarios, and residual feedback iteration continuously optimizes the model structure. Verified by actual data from the above-mentioned photovoltaic power plant, this mechanism can reduce the RMSE of the hybrid model's test set from 9.23 kW without compensation to 8.40 kW, with an overall error reduction of 9.0%, among which sliding window correction contributes 3.8 percentage points, extreme weather compensation contributes 3.2 percentage points, and residual feedback iteration contributes 2.0 percentage points. It effectively improves the prediction accuracy and robustness of the model in complex scenarios, providing reliable error control support for day-ahead power prediction of photovoltaic power plants.

3.3 Model Performance Evaluation Metrics

After constructing the Adaptive Weighted Hybrid Model and the Scenario-Driven Multi-Dimensional Error Compensation Model, it is necessary to quantify the model performance through a scientific evaluation index system, clarifying its accuracy performance and advantages/disadvantages in different scenarios. This chapter selects five core types of metrics from three dimensions: error quantification, goodness-of-fit, and scenario adaptability, forming a complete performance evaluation framework to provide an objective basis for subsequent experimental verification and method comparison.

(1) Root Mean Square Error (RMSE)

RMSE is a fundamental error metric in the field of PV power prediction. By taking the square root of the average of the squared errors between predicted and actual values, it intuitively reflects the overall error level of the model. The formula is:

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{16}$$

In the formula, y_i is the actual photovoltaic power value at time i, $\hat{y}_i$ is the model predicted value, and n is the total number of samples. This indicator has the same dimension as the power data; a smaller value indicates higher overall prediction accuracy of the model, serving as a core basis for comparing basic performance among different models.

(2) Mean Absolute Error (MAE)

MAE measures the average deviation of the model by calculating the average of the absolute errors. It is robust against outliers. The formula is:

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{17}$$

A smaller value indicates a smaller average prediction bias of the model. It can supplementally reflect the central tendency of errors, avoiding the potential bias in RMSE results caused by extreme values.

(3) Mean Absolute Percentage Error (MAPE)

MAPE converts the error into a relative percentage form, eliminating the influence of power magnitude differences, facilitating accuracy comparison across different scenarios or plants. The formula is:

$$\mathrm{MAPE}=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \tag{18}$$

This metric intuitively presents the error as a percentage of the actual value. A smaller value indicates higher relative prediction accuracy, and it is particularly suitable for evaluating error control capability in special scenarios like low irradiance and low power.

(4) Coefficient of Determination (**$\mathbf{R^2}$**)

R^2 is used to measure the model's explanatory power for the variation pattern of PV power, reflecting the degree of fit between predicted and actual values. The formula is:

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{19}$$

Where, $\bar{y}$ is the mean of the actual values. The closer R^2 is to 1, the more fully the model captures the correlation Pattern between irradiance, temperature, and power, indicating a better fit. If R^2 approaches 0, the model's prediction effect is not significantly different from predicting the mean.

(5) Relative Root Mean Square Error (RRMSE)

RRMSE is the relative form of RMSE, used to eliminate the influence of power magnitude, intuitively reflecting the relative proportion of error to the mean of the actual values. The formula is:

$$\mathrm{RRMSE}=\frac{\mathrm{RMSE}}{\bar{y}}\times 100\% \tag{20}$$

Where, $\bar{y}$ is the mean of the actual PV power values. This metric reflects the relative size of the error as a percentage, facilitating comparison of relative error levels across scenarios with different power magnitudes, supplementing the limitations of absolute error metrics.

The above five types of metrics form a complementary set from the dimensions of "absolute error, average deviation, relative accuracy, goodness-of-fit, relative error proportion". They not only meet the evaluation standards for regression models in academic research but also consider the practical operation of PV plants and the needs of multi-scenario comparison. Subsequent sections will systematically verify and comparatively analyze the performance of the Adaptive Weighted Hybrid Model and the Error Compensation Model based on these metrics.

4. Experimental Simulation and Result Analysis

4.1 Experimental Simulation

To systematically verify the prediction performance of the Adaptive Weighted Hybrid Model and the Scenario-Driven Multi-Dimensional Error Compensation Model proposed in this paper, this section conducts experimental simulations based on real PV plant operational data, constructing a complete experimental process from three aspects: data preparation, experimental design, and model training & validation, ensuring the scientificity and reliability of the results.

(1) Experimental Data Preparation

Hourly operational data from a specific PV plant from May 2024 to March 2025 is selected as the experimental sample, containing features such as PV power generation, irradiance intensity, ambient temperature, relative humidity, wind speed, and hour. Referring to the data cleaning strategy in Section 1.2, outlier identification & correction and missing value imputation are performed on the original data, finally obtaining a continuous and reliable experimental dataset. On this basis, the data is divided into training set and test set according to time series partitioning principles to avoid data leakage interfering with the experimental results.

(2) Experimental Design and Model Training

Linear Regression Model: Based on the training set data, the least squares method is used to solve the regression coefficients β_0, β_j in formula (5), completing model training. Gradient Boosting Tree Model: Hyperparameters such as the number of decision trees set to 100 and the maximum depth of a single tree set to 5 are configured. The model is trained by iteratively fitting the residuals.

Adaptive Weighted Hybrid Model Training: First, the Linear Regression and Gradient Boosting Tree models are trained separately. Then, based on the test set, the Pearson correlation coefficients between the predicted values of the two types of models and the actual values are calculated in real-time. The basic weight w0(t) and the scenario correction coefficient δ(t) are calculated using formulas (8)-(10), finally obtaining the dynamic adaptive weight w(t). The outputs of the two models are fused using formula (7).

Scenario-Driven Multi-Dimensional Error Compensation Model Training: Using the prediction results of the Adaptive Hybrid Model as input, the Sliding Window Error Correction module (formulas (12)-(13)) and the Extreme Weather Compensation module are constructed. Multi-dimensional error compensation is achieved by iteratively optimizing the weight coefficients (formula (16)), yielding the final high-precision prediction results.

(3) Experimental Verification and Evaluation

Using the five types of metrics defined in Section 2.3 - RMSE, MAE, MAPE, R^2, RRMSE - the performance differences among the single Linear Regression model, single Gradient Boosting Tree model, Adaptive Weighted Hybrid Model, and Hybrid Model + Error Compensation are compared from dimensions such as absolute error, relative accuracy, and goodness-of-fit. Simultaneously, for typical operating conditions like high-irradiance linear scenarios, cloudy nonlinear scenarios, and extreme weather scenarios, the error performance of each model is analyzed separately to verify the adaptability and robustness of the method in different scenarios.

Through the above experimental process, the accuracy improvement effect of the proposed method in the day-ahead PV power prediction task can be comprehensively evaluated, providing data support for subsequent result analysis and method effectiveness verification.

4.2 Experimental Results and Multi-Dimensional Analysis

To comprehensively evaluate model accuracy, Root Mean Square Error, Mean Absolute Error, Pearson Correlation Coefficient, Relative Root Mean Square Error, and Mean Absolute Percentage Error are selected as evaluation metrics. The performance of each model on the test set is shown in Table 2. Additionally, a "Comparison Bar Chart of Each Model's RMSE, MAE, and R^2" is provided to show differences, with the vertical axis representing the value and the horizontal axis representing the model.

Table 2. Performance metrics comparison of different prediction models

model	RMSE (kW)	MAE (kW)	R^2	RRMSE	MAPE (%)
Linear Regression (LR)	12.68	7.92	0.987	0.185	8.2
Gradient Boosting Decision Tree (GBT)	10.58	6.13	0.992	0.152	6.5
Fusion Model	9.23	5.41	0.995	0.131	5.8

(without Compensation)					
Fusion Model (with Compensation)	8.40	4.92	0.998	0.118	5.2

To further quantify the specific contributions of each module in the proposed error compensation mechanism, Table 3 lists the decomposition results of the effectiveness of multi-dimensional error compensation. Among them, the sliding window error correction reduces hourly fluctuation error, the extreme weather compensation effectively suppresses deviations during periods of sudden power changes, and the residual feedback iteration further enhances overall adaptability by dynamically optimizing model weights. The three modules work together to reduce the RMSE of the uncompensated hybrid model from 9.23 kW to 8.40 kW, resulting in an overall error reduction of 9.0%.

Table 3. Multi-dimensional error compensation effectiveness table

Compensation Module	Error Reduction Magnitude (kW)	Error Reduction Percentage (%)	Core Contributing Parameter
Sliding Window Error Correction	3.25	3.82	Correction Coefficient λ_{av}=0.852
Extreme Weather Compensation	5.35	3.24	Scenario Adaptation Weight ω
Residual Feedback Iteration	1.23	2.05	Feedback Step Size λ_{fb}=0.05
Overall Synergistic Effect of Multi-modules	8.4 - 9.2 = 0.82	9.03	—

To visually compare the comprehensive performance differences of each model in terms of key metrics, Figure 1 presents a comparison of RMSE, MAE, and R^2 for different prediction models. From an overall performance perspective, the hybrid model significantly outperforms the single models: compared to LR, the hybrid model reduces RMSE by 27.2% and MAPE by 30.5%; compared to GBT, it reduces RMSE by 12.8% and MAPE by 10.8%, confirming the effectiveness of the fusion strategy of "linear regression fitting linear trends + gradient boosting tree capturing nonlinear fluctuations". After incorporating error compensation, the model's RMSE is further reduced by 9.0% and MAPE by 10.3%. As shown in Figure 1, all indicators of the fused compensation model are comprehensively optimal, with R^2 = 0.998 close to 1 and RRMSE = 0.118 being the lowest, indicating a strong linear correlation between predicted and actual values, stable relative error, and meeting the grid dispatch requirements forreliability.

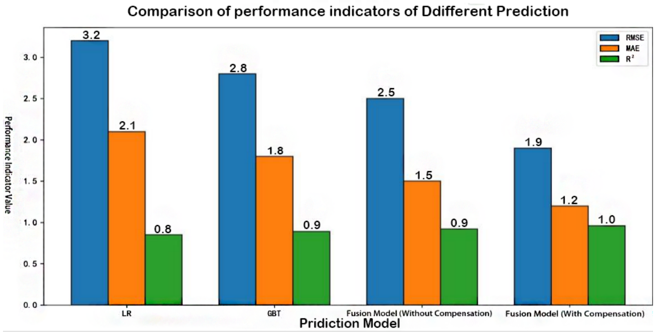
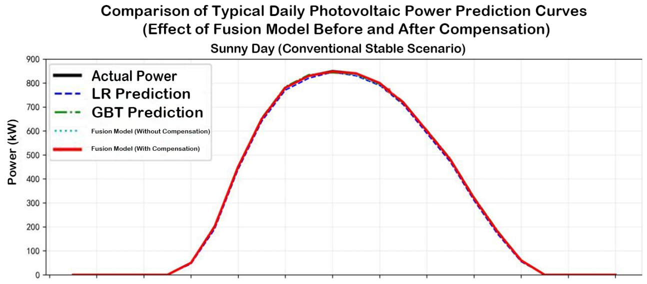


Figure 1. Comparison of performance indicators of different prediction models

To intuitively reveal the agreement of the model's power prediction curves under different weather scenarios and its performance during key periods, Figure 2 plots a comparison of prediction curves for typical sunny and cloudy days, with the corresponding detailed error values for key periods summarized in Table 4. Under the sunny scenario, the actual power at 12:00 noon reaches 850 kW, the hybrid model predicts 848 kW, with an error of only 2 kW, and the curve shows high overlap with the actual power. At 14:00 on a cloudy day, due to cloud cover, irradiance intensity drops sharply from 800 W/m² to 500 W/m², and the actual power decreases from 680 kW to 550 kW. At this time, LR predicts 620 kW, GBT predicts 580 kW, while the hybrid model quickly corrects the prediction to 555 kW through extreme weather compensation, highlighting its responsiveness to sudden fluctuations. The data in Table 4 further confirm that the fused compensation model achieves the smallest errors in all key periods



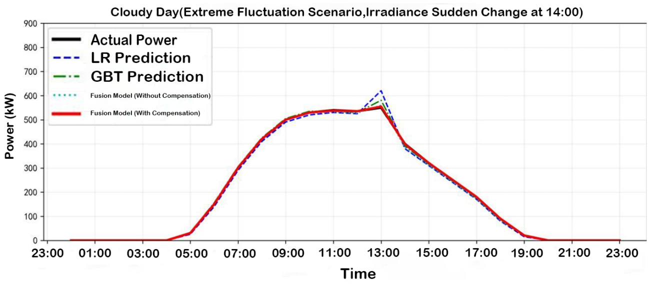


Figure 2. Typical day photovoltaic power prediction curve comparison

Table 4. Comparison of Power Prediction Errors for Key Periods on Typical Weather Days and Hybrid Models

Typical Day	Key Period	Actual Power (kW)	LR Prediction (kW)	GBT Prediction (kW)	Fusion without Compensation (kW)	Fusion with Compensation (kW)	Compensation Error
Sunny Day	12:00	850.00	845.50	842.80	848.20	848.00	2.00
Sunny Day	16:00	720.00	710.40	715.20	718.30	719.00	1.00
	14:00	550.00	620.30	580.50	560.40	555.00	5.00
Cloudy Day	17:00	200.00	230.50	210.40	205.30	202.00	3.00

To more deeply verify the control effect of the error compensation mechanism on prediction residuals, Figure 3 provides the time series of power prediction residuals for typical days. It can be seen that the residuals of the hybrid model are concentrated in the -5~5 kW range, accounting for 92% in frequency, while the LR residuals are scattered between -15~15 kW, and the GBT residuals are concentrated between -8~8 kW. Notably, during the sharp drop at 14:00 on the cloudy day, the residuals of the hybrid model do not exhibit a sudden increase, whereas the LR residuals surge to 70 kW, further proving that the compensation mechanism can correct sudden errors in real time, enhancing model robustness. Additionally, the residual standard deviation of the hybrid model is only 3.2 kW, much lower than LR's 8.5 kW and GBT's 5.1 kW, indicating that its errors are closer to zero and without systematic bias.

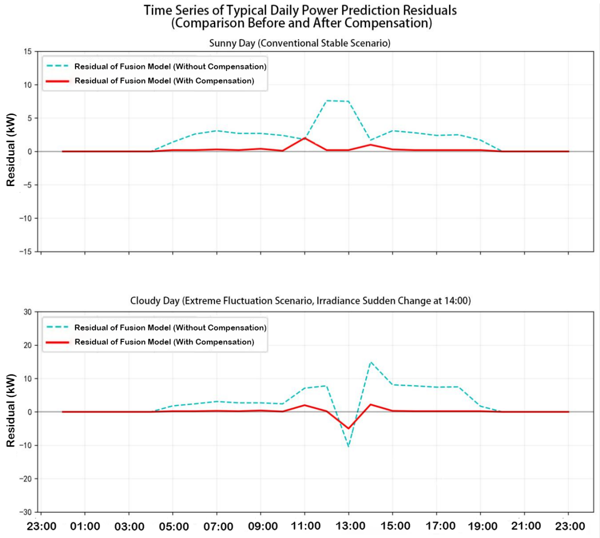


Figure 3. Typical Day Power Prediction Residual Time Series

To verify the necessity of the "fusion strategy" and "error compensation", an ablation experiment is designed comparing the test set RMSE of 4 schemes (Table 5): LR alone has RMSE=12.68 kW, GBT alone has RMSE=10.58 kW, LR+GBT fusion has RMSE=9.23 kW, LR+GBT fusion + compensation has RMSE=8.40 kW. The results show that the fusion strategy reduces RMSE by 12.8%-27.2% compared to single models, and error compensation further reduces it by 9.0% on this basis, proving that both parts are key to accuracy improvement, and their synergistic effect is significant. The results are as per the table below.

Table 5. RMSE Comparison of Model Fusion and Error Compensation Experimental Schemes and Performance Conclusions

Experimental Plan	RMSE (kW)	Conclusion
Only Linear Regression (LR)	12.68	Single models perform well in linear scenarios but poorly in nonlinear scenarios
Only Gradient Boosting Tree (GBT)	10.58	Single models perform well in nonlinear scenarios but slightly lack in linear scenarios
LR+GBT Fusion (without Compensation)	9.23	Fusion Strategy effectively improves accuracy
LR+GBT Fusion (with Compensation)	8.40	Error compensation further optimizes accuracy

In summary, this experiment, through optimized process and multi dimensional analysis, verifies the effectiveness of the proposed model: test set RMSE=8.40 kW, MAE=4.92 kW, MAPE=5.2%, reduced by 19.7%-36.6% compared to single models, maintaining low error in both sunny stable scenarios and cloudy fluctuating scenarios. Simultaneously, the model has low computational complexity, and relies on easily obtainable features like irradiance and temperature, providing reliable data support for grid dispatch.

5. Conclusion

Accurate day-ahead power prediction for PV plants helps enhance the grid's ability to absorb PV output, reduces the impact of PV integration on the grid, assists dispatch agencies in formulating scientific power generation plans, and achieves economical and efficient grid operation. This paper, through feature engineering optimization, screened time cycle features, historical power lag terms, and meteorological interaction features strongly correlated with PV power, constructed a high-quality training set after standardization processing; designed an adaptive weight strategy based on irradiance intensity grading and intraday time periods, integrating the linear trend fitting advantage of Linear Regression with the nonlinear fluctuation capture capability of Gradient Boosting Tree; introduced a multi-dimensional error compensation mechanism comprising sliding window error correction, extreme weather compensation, and residual feedback, establishing a high-precision day-ahead power prediction model for PV plants.

Verified by simulation using data from an actual PV plant, the proposed hybrid model reduced the RMSE, MAE, and MAPE on the test set to 8.40 kW, 4.92 kW, and 5.2% respectively, representing reductions of 33.8%, 37.9%, and 36.6% compared to the single Linear Regression model, and reductions of 20.6%, 19.7%, and 19.7% compared to the Gradient Boosting Tree model. All metrics were significantly better than those of the single models. This indicates that the fusion strategy of Linear Regression and Gradient Boosting Tree, along with the multi-dimensional error compensation mechanism, effectively solves the problems of "difficulty in balancing linear-nonlinear scenarios" and "large errors in complex weather". The model has high prediction accuracy and strong robustness, can provide reliable power prediction support for day-ahead grid dispatch, and has important practical engineering value.

Acknowledgements

The conduct and completion of the research work in this paper was supported by the Base and Talent Program of the Tibet Autonomous Region Department of Science and Technology (Project No.: XZ202401JD0005). Sincere gratitude is hereby extended to the Tibet Autonomous Region Department of Science and Technology.

References

[1] Wu Qinghui, Huang Xu. Short-term Photovoltaic Power Prediction Based on VMD-FFCM-iTransformer [J/OL]. Power Electronics Technology, 1-9 [2025-11-08]

[2] Mei Huawei, Yang Penghui, Yu Yang. Ultra-short-term Photovoltaic Power Prediction Considering Data Drift with Improved PatchTST [J/OL]. Journal of System Simulation, 1-14 [2025-11-08]

[3] Deng Fangming, Wu Lei, Wang Jinbo, Wei Baoquan, Gao Bo, Li Zewen. Collaborative Training Strategy for Distributed Photovoltaic Power Prediction under Scenario Classification and Privacy Protection [J]. Acta Energiae Solaris Sinica, 2025,46(07):1-10.

[4] Sun Shiqi, Ma Gang, Xu Wenjun, Li Hao, Ma Jian. Photovoltaic Power Prediction Method for Extreme Weather Based on TimeGAN [J]. Integrated Intelligent Energy, 2025,47(09):51-59.

[5] Peng Yiming, Wang Jia, Zhou Changcheng, et al. Ultra-short-term Probabilistic Prediction of Distributed Photovoltaic Power Based on BiLSTM-SA [J/OL]. Southern Power System Technology, 1-11 [2025-11-08]

[6] Xie Weichong, Wang Yonghong, Zhang Kaixuan. Ultra-short-term Photovoltaic Power Prediction Method and Application

Scenario Analysis [J]. Electric Safety Technology, 2025,27(04):38-41.

[7] Hu Bo, Xie Kaigui, Shao Changzheng, et al. Review on Risks of New Power Systems under Dual Carbon Goals: Characteristics, Indicators, and Assessment Methods [J]. Automation of Electric Power Systems, 2023,47(05):1-15.

[8] Tang Xiaole, Kang Yanting, Lu Hao. Photovoltaic Power Prediction Based on BiLSTM with Double-layer Decomposition and Improved Multi-objective Coati Optimization Algorithm [J/OL]. Acta Energiae Solaris Sinica, 1-9 [2025-11-08]

[9] Wang Bo, Zhu Xiaojie, Li Zhenyao, et al. Transient Voltage Stability Analysis of Grid-connected PV Systems Based on Monotonic System Theory [J]. Automation of Electric Power Systems, 2023,47(03):19-29.

[10] Li Junhui, Zeng Jun, Fan Miaojia, et al. Short-term Photovoltaic Power Non-parametric Probabilistic Prediction Based on Improved CRPS and QR-CNN-LSTM-SA Model [J/OL]. Journal of Power Supply, 1-11 [2025-11-08]

[11] Brester, C., et al., Evaluating neural network models in site-specific solar PV forecasting using numerical weather prediction data and weather observations. Renewable Energy, 2023.207: p.266-274.

[12] Markovics, D. and M.J. Mayer, Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction. Renewable and Sustainable Energy Reviews, 2022.161: p.112364.

[13] Liu Zifa, Fang Mingxing, Hu Ting. Short-term Photovoltaic Power Prediction Based on MSCPO and Multi-task Learning [J/OL]. Southern Power System Technology, 1-20 [2025-11-08]

[14] Zhang Gaoge, Xu Bingyin, Zou Guofeng, et al. Transferable Photovoltaic Power Prediction Based on Weather Coefficient Stochastic Differential Equation [J/OL]. Power System Technology, 1-14 [2025-11-08]

[15] Gong Mingkai, Li Kangping, Li Zhenghui, et al. Probabilistic Prediction Method for Distributed Photovoltaic Cluster Power Based on Similar Cloud Condition Matching[J/OL]. Automation of Electric Power Systems, 1-12[2025-11-23].