ISSN: 2710-0723

A Hybrid Machine Learning Framework for Day-Ahead PV Power Forecasting with NWP Integration and Spatial Downscaling

Zishuo Liu*, Zezheng Han

School of International Education, Hebei University of Technology, Tianjin 300401, China

*Corresponding author: zishuoliuu@163.com

Abstract

Accurate forecasting of photovoltaic (PV) power is essential for secure grid operation, but remains difficult due to meteorological uncertainty and the coarse spatial resolution of numerical weather prediction (NWP) data. To address these challenges, this study develops a hybrid framework that integrates physical modeling with machine learning. The PVWatts model is first applied as a physical baseline. Building on this, historical power records, temporal statistical features, and NWP data are incorporated into a LightGBM regression model for day-ahead forecasting. Comparative t-tests confirm that NWP features significantly improve predictive accuracy, while scenario partitioning using global horizontal irradiance (GHI) further enhances model adaptability. To resolve spatial mismatch, a Random Forest–based downscaling method is proposed, reconstructing high-resolution meteorological features from site-level data. These downscaled features are integrated into the forecasting model, yielding higher accuracy, robustness, and performance under diverse operating conditions. Overall, this study establishes a multi-level PV forecasting system that combines physical mechanisms and data-driven learning. The proposed approach achieves strong adaptability and reliable accuracy, providing effective support for renewable energy integration, smart grid scheduling, and PV plant operation.

Keywords

Spatio-temporal Feature Integration; LightGBM; Random Forest; Multi-source Data Fusion.

1. Introduction

Photovoltaic (PV) generation, as a clean and efficient renewable energy technology, has been widely deployed and rapidly developed worldwide in recent years [1]. However, compared with conventional power sources, PV generation is inherently intermittent and highly volatile, being significantly affected by solar irradiance, weather conditions, and geographical location [2]. These factors introduce considerable uncertainty into forecasting [3]. In addition, the presence of temporal variability-such as intraday fluctuations and weekly or monthly periodicity-further limits the accuracy and flexibility of short-term prediction [4]. Consequently, improving the precision of PV power forecasting has become a critical issue that must be addressed [5].

Accurate day-ahead forecasting plays a vital role in ensuring the reliability of renewable energy integration, enabling grid operators to schedule generation plans, balance power demand, and mitigate risks associated with variability. Especially under high renewable penetration scenarios, an accurate forecast directly influences dispatching efficiency, reserve management, and market operation stability. Therefore, research into advanced forecasting models is of great importance for promoting large-scale PV adoption and achieving carbon neutrality goals.

ISSN: 2710-0723

In recent years, both physical models and data-driven methods have been applied to improve forecasting accuracy [6]. Physical models can reflect energy conversion mechanisms but often fail under complex meteorological conditions, while machine learning methods capture nonlinear relationships but depend heavily on data quality [7]. To overcome these limitations, hybrid frameworks integrating physical principles with data-driven learning have become an emerging research trend [8].

To achieve high-accuracy short-term forecasting for PV power plants, several approaches have been explored, including physics-based modeling, historical data-driven methods, numerical weather prediction (NWP) integration, and spatial downscaling techniques [9]. However, challenges remain regarding the multi-scale spatio-temporal variability of PV output and the spatial mismatch between NWP data and actual plant locations. Addressing these issues, this study develops a hybrid forecasting framework that combines physical modeling, machine learning, and downscaling to improve adaptability under dynamic meteorological conditions.

In this study, the Kivikko Photovoltaic Power Plant in Finland is selected as the case study. Focusing on the uncertainty arising from multi-scale spatio-temporal features in PV power generation, this work integrates physical modeling, historical data, NWP information, and spatial downscaling techniques to construct a multi-level, high-precision forecasting framework. The proposed model leverages the PVWatts baseline for physical estimation, employs LightGBM regression to learn complex nonlinearities, and introduces Random Forest–based downscaling to refine spatial meteorological resolution, thereby achieving improved robustness and forecasting accuracy.

2. Model Development and Solution Analysis

2.1. Development of a Day-Ahead PV Power Forecasting Model with Integrated NWP Data

To further enhance the accuracy and robustness of day-ahead photovoltaic (PV) power forecasting, this study develops an integrated forecasting model that incorporates multi-source and multi-scale spatio-temporal features, with a particular focus on numerical weather prediction (NWP) information. The model fuses historical power outputs, NWP meteorological variables, statistical incremental features, and periodic time indicators, while employing the LightGBM regressor to capture nonlinear patterns in power fluctuations, thereby enabling efficient prediction of the next 24-hour PV power generation [10].

Specifically, the procedure begins by collecting the historical power output sequences from the PV plant along with NWP-based meteorological data. Both datasets are aligned by timestamp to form a multivariate joint sample set. Next, multi-order lag features and rolling statistical measures are defined. As in the previous model, Equation (1) denotes the construction of lagged features, while Equations (2) and (3) describe the rolling statistical quantities. In addition, to account for abrupt weather changes and dynamic system responses, meteorological and power increment features are further introduced.

$$delta\_power_t = P_t - P_{t-1} \tag{1}$$

$$delta\_GHI_t = GHI_t - GHI_{t-1} \tag{2}$$

$$delta\_temp_t = temp\_air_t - temp\_air_{t-1} \tag{3}$$

ISSN: 2710-0723

In terms of temporal periodic features, hourly, weekly, and monthly variables are introduced, and sine–cosine transformations are applied to model the intraday cyclic patterns.

$$sin\_hour_t = \sin\left(\frac{2\pi \cdot hour_t}{24}\right) \tag{4}$$

$$cos\_hour_t = \cos\left(\frac{2\pi \cdot hour_t}{24}\right) \tag{5}$$

Finally, the target variable is defined as the PV power output at the corresponding timestamp 24 hours in the future, which serves as the prediction objective for the day-ahead forecasting task.

$$target_t = P_{t+24} \tag{6}$$

For model training and optimization, the LightGBM (Light Gradient Boosting Machine) regression framework is employed. The training process aims at stage-wise minimization, progressively refining the decision tree ensemble so that the root mean square error (RMSE) between predicted and observed outputs is minimized. To ensure generalization capability and suppress overfitting, an early stopping mechanism is introduced: training is automatically terminated if the RMSE on the validation set remains below the predefined threshold of 0.03 for 50 consecutive iterations, with the model parameters corresponding to the best performance preserved.

For performance evaluation, multiple complementary metrics are adopted to comprehensively assess the validity and effectiveness of the forecasting model.

2.2. Random Forest–Based PV Power Forecasting Model with Downscaled NWP Features

For the day-ahead power forecasting problem of photovoltaic plants, this study proposes a downscaling prediction approach based on Random Forest regression, fully integrating numerical weather prediction (NWP) information with temporal periodic features from a practical application perspective. Specifically, the method takes the actual historical power output of the PV plant as the target variable, while multiple meteorological and temporal factors are used as input parameters to establish a nonlinear mapping relationship for forecasting future power generation.

In particular, seven features are selected as model inputs: global horizontal irradiance, air temperature, relative humidity, wind speed, hour of the day, day of the week, and month.

The feature vector is expressed as:

$$\mathbf{X}_t = [GHI_t, temp\_air_t, humidity_t, wind\_speed_t, hour_t, dayofweek_t, month_t] \tag{7}$$

The processed samples are divided into training and testing sets in chronological order, with the first 40% used for training and the remaining 60% reserved for testing. Within the training set, the nonlinear relationships between the input features and power output are learned. A Random Forest regressor is then employed to integrate multiple decision trees, thereby mapping the input features to the target variable in a nonlinear manner. In this study, the Random Forest is configured with 100 trees and a maximum tree depth of 10 [11].

ISSN: 2710-0723

$$\hat{y}_t = \frac{1}{N}\sum_{k=1}^{N} h_k\left(\mathbf{X}_t\right) \tag{8}$$

where, h_k denotes the K-th regression tree, and $\hat{y}_t$ represents the predicted power output. By training on different subsets of features and samples, each tree enhances the overall generalization capability of the model and improves robustness under abnormal conditions.

Finally, the trained model is applied to the testing set for power forecasting. Six evaluation metrics are employed to assess the accuracy and goodness-of-fit of the prediction results: root mean square error (E_{rmse}), mean absolute error (E_{mae}), mean error (E_{me}), correlation coefficient (r), prediction accuracy (C_r), and qualification rate (Q_R).

3. Results

3.1. Results of the Day-Ahead PV Power Forecasting Model with Integrated NWP Data

Based on the proposed day-ahead photovoltaic (PV) power forecasting model that integrates NWP information with multi-scale spatio-temporal features, the program implementation yields the results in Figure1.

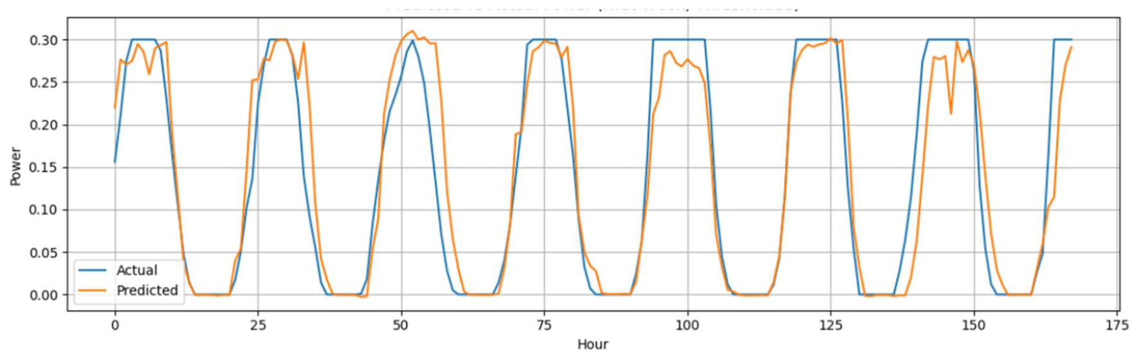

Figure 1. Comparison between predicted and actual PV power output over the first 168 hours.

As shown in Figure 1, the predicted results for the first 168 hours of the testing set exhibit a high degree of overlap with the actual power outputs. Overall, the model effectively captures the periodic fluctuation patterns, demonstrating strong forecasting capability under regular weather conditions. However, in certain special cases-such as abnormal weather events or equipment failures-noticeable deviations between the predicted and actual outputs can still be observed. In general, the forecasting curve provides an accurate representation of the actual operating output of the plant, successfully predicting the overall power fluctuation trends and exhibiting strong dynamic tracking capability.

Figure 2 presents the feature importance ranking based on gain, from which the following observations can be made:

The feature "power output at the same time on the previous day" (lag_1d) has the greatest impact on the forecasting results, far exceeding other factors. This indicates that the strong daily periodicity plays a dominant role in PV power forecasting within the studied model. Features such as "power output at the same time on the previous 2–3 days" (lag_2d, lag_3d), hour of the day, and global horizontal irradiance (GHI) also exhibit relatively high importance. This suggests that the model is effective in capturing both historical inertia effects and

ISSN: 2710-0723

environmental influences. Other statistical features and periodic variables contribute less to the model's predictive power. Nevertheless, they still provide complementary support, particularly for improving the model's performance in forecasting under abnormal fluctuation scenarios.

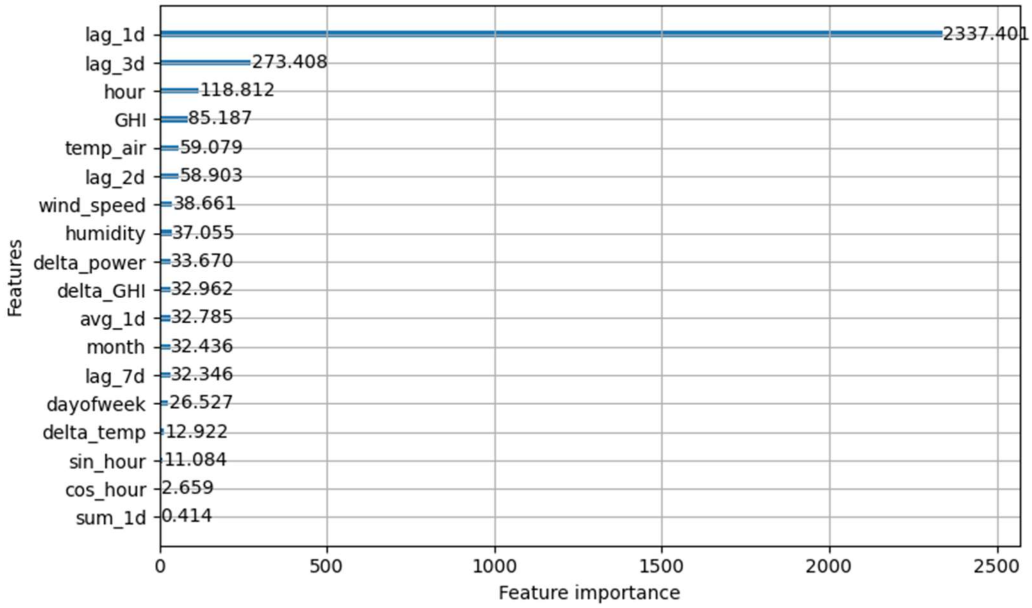

Figure 2. illustrates the relative importance of input features in the forecasting model, where the ranking is measured by the gain criterion

Table 1. Training log of the forecasting model

Training rounds	Training RMSE	Validation RMSE
100	0.1175410	0.0742321
200	0.0847516	0.0564512
300	0.0725906	0.0531192
400	0.0656070	0.0528876
500	0.0606002	0.0529310
600	0.0567723	0.0528315
700	0.0535781	0.0527113
800	0.0508582	0.0525628
900	0.0484073	0.0525264
1000	0.0458897	0.0524958

After introducing the early stopping mechanism, the training process was recorded. As shown in Table 1, with the increase in training iterations, the root mean square errors (RMSE) of both the training and validation sets gradually decreased, reaching 0.0458897 and 0.524958, respectively, at the 1000th iteration. At this point, the RMSE of the validation set tended to stabilize, indicating that the model achieved good generalization capability. This suggests that the training process was reasonable and that model performance did not deteriorate due to overfitting as the number of iterations increased.

To further evaluate whether the proposed NWP-integrated model can effectively improve forecasting accuracy, its prediction performance was compared with that of the baseline regression model developed in Section 2, which incorporated only historical power features and statistical temporal characteristics. Specifically, the global horizontal irradiance (GHI) was

ISSN: 2710-0723

used as the grouping criterion, and the testing samples were divided into four intervals: low, medium-low, medium-high, and high GHI. For each interval, the mean absolute error (MAE) of the baseline model (without NWP features) and the integrated model (with NWP features) was compared. The error improvement rate was then calculated as the difference between the baseline error and the integrated model error.

Table 2. Statistical t-test comparison with the baseline model

GHI Interval	Sample Size	Baseline Mean Error	Integrated Mean Error	Error Improvement	*p*-value
Medium-high	1058	0.049989	0.048333	0.001656	4.417273e−01
High	1065	0.048203	0.041069	0.007135	3.144808e−04
Medium-low	1056	0.012445	0.009473	0.002972	8.560413e−95
Low	1079	0.011812	0.001841	0.009971	0.000000e+00

According to the results of the t-test, which are shown in the Table 2, the NWP-integrated model shows statistically significant improvements in the high, medium-low, and low GHI intervals, where the p-values are all less than 0.05. In these cases, the mean error of the integrated model is consistently lower than that of the baseline model developed in Problem 2, confirming that the incorporation of NWP features effectively enhances forecasting accuracy.

Furthermore, scenario-based partitioning was performed using global horizontal irradiance (GHI) to improve predictive performance. In medium-high and high GHI scenarios, prioritizing the use of machine learning models with integrated NWP features yields significant accuracy gains. In medium-low GHI scenarios, selectively incorporating NWP features can moderately reduce forecasting errors. In low GHI scenarios, the benefits of NWP features are limited; hence, considering cost-effectiveness, their use can be minimized.

Adopting different strategies under varying meteorological conditions facilitates a more intelligent adjustment of PV power forecasting accuracy. To further validate the rationality of the proposed scenario partitioning scheme, a Random Forest model was applied to perform point-by-point forecasting of 15-minute interval power outputs during the last week of February for each GHI interval, and the results were visualized accordingly.

From Figure 3, the following observations can be made:

In the low-GHI interval, power outputs are generally low, with predicted and actual values highly concentrated in the lower range. The fitting status is good, deviations are minimal, and the inclusion of NWP features has little impact.

In the medium-low GHI interval, power outputs increase compared with the low-GHI case, while the fitting quality remains satisfactory. Incorporating NWP features provides a modest improvement in prediction accuracy within this range.

In the medium-high and high GHI intervals, power outputs increase significantly and data points are more widely distributed. This indicates that under stronger irradiance conditions, the inclusion of NWP features substantially enhances the model's ability to capture large power fluctuations, thereby improving forecasting accuracy.

In summary, the day-ahead PV power forecasting model that integrates NWP information with multi-scale spatio-temporal features effectively improves prediction accuracy. Moreover, adopting different forecasting strategies across GHI scenarios allows for more rational and adaptive day-ahead forecasting of PV plant power outputs.

ISSN: 2710-0723

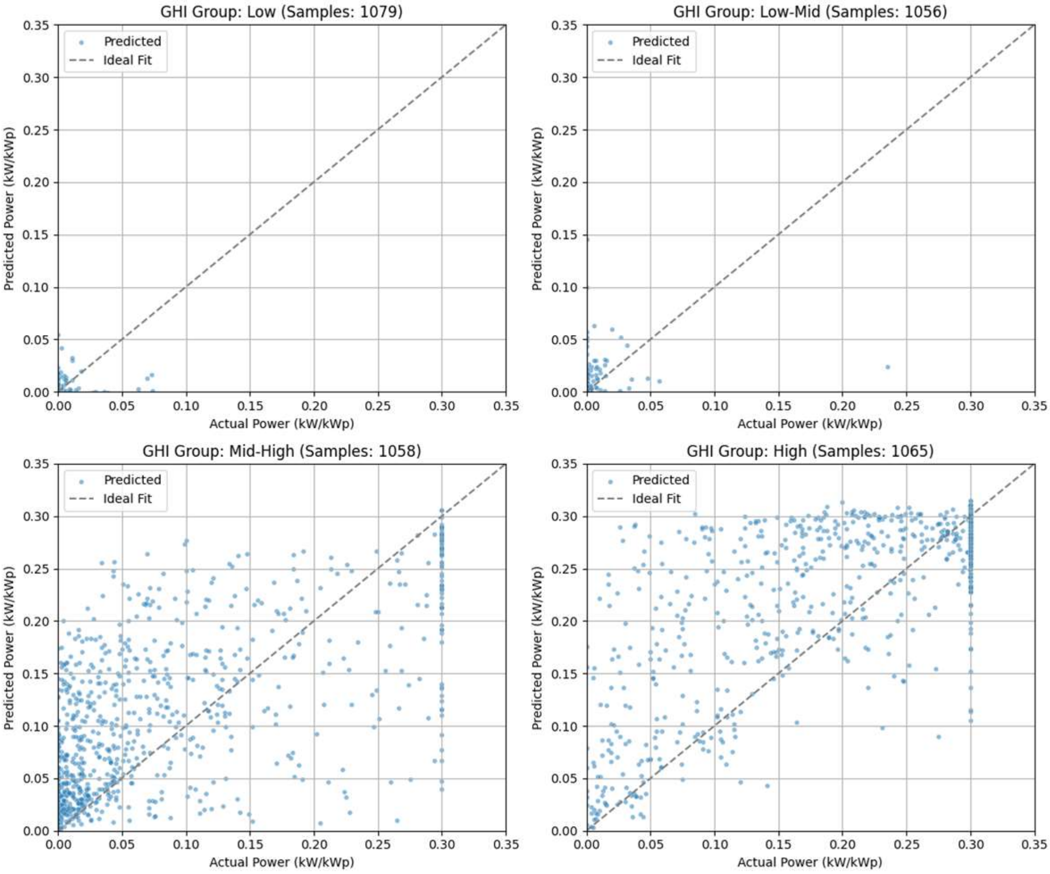

Figure 3. Fitted scatter plot of predicted vs. actual PV power output

3.2. Results of the Random Forest–Based PV Power Forecasting Model with Downscaled NWP Features

Based on the above theoretical modeling and program implementation, the following results are obtained in Figure 4.

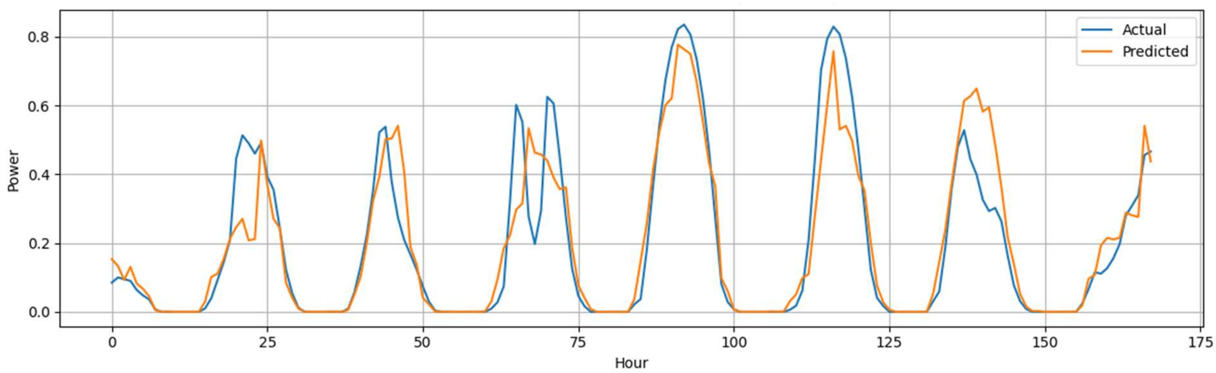

Figure 4. Time series comparison of downscaled predictions and actual values based on Random Forest regression.

As shown in Figure 4, the predicted power curve closely matches the actual observations, indicating that the model effectively captures the intraday cyclical fluctuations of PV output and accurately identifies both peak and valley values. Minor deviations appear only at certain times during early morning and late evening, where the predicted values are slightly overestimated or underestimated. However, these differences remain small, and the overall forecasting error

ISSN: 2710-0723

is minimal, with the predicted values remaining largely consistent with the actual measurements.

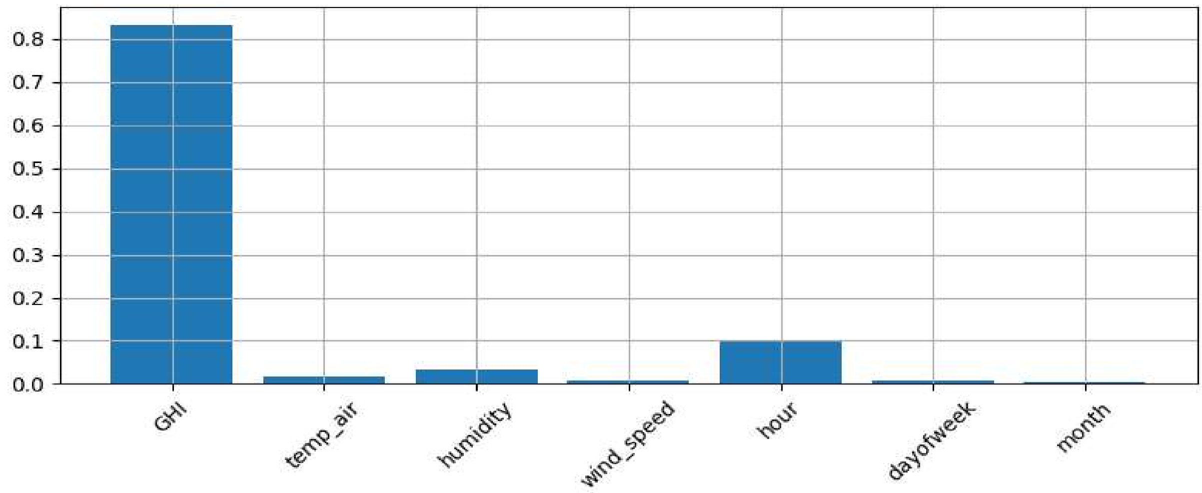

Figure 5. Feature importance ranking of the Random Forest model.

In terms of feature importance, the following observations can be made:

As shown in Figure 5, global horizontal irradiance (GHI) contributes far more than any other input variable, confirming its role as the primary factor influencing PV output forecasting. Beyond meteorological variables, the temporal periodic feature hour also exhibits relatively high importance, effectively enhancing the model's ability to capture intraday dynamic characteristics of PV output. Although the remaining features have comparatively lower weights, they still provide auxiliary benefits, particularly in identifying extreme weather conditions or capturing abnormal fluctuations in power output.

Table 3. Error evaluation metrics of the downscaled prediction model based on Random Forest regression.

Error Evaluation Metric	Result
E_{rmse}	0.0692
E_{mae}	0.0315
E_{me}	-0.0030
r	0.9448
C_r	21.11%
Q_R	-45.80%

From Table 3, the following conclusions can be drawn:

The correlation coefficient is close to 0.95, indicating a very strong correlation and goodness of fit between the predicted and actual power outputs.

Both the root mean square error and mean absolute error are relatively low, reflecting the strong numerical approximation capability of the model and its superior performance in terms of forecasting accuracy.

The prediction accuracy and qualification rate demonstrate that the forecasting errors of the model remain within a controllable range, thereby confirming the reliability of the approach.

ISSN: 2710-0723

Overall, the proposed method achieves high-precision and robust modeling of PV power generation. Under realistic meteorological conditions, it provides high-quality forecasting results with strong accuracy and stability. This confirms the model's capability to effectively adapt to dynamic power variations under complex climatic and environmental conditions, highlighting its high practical applicability.

4. Conclusion and Outlook

For the proposed day-ahead PV power forecasting model integrating NWP information and multi-scale spatio-temporal features, several improvements over baseline models are achieved. These include the incorporation of NWP-based meteorological features to enrich the input dimension, the fusion of spatio-temporal features to better capture sequential and periodic patterns, and the use of ensemble learning to enhance both predictive accuracy and generalization capability. By introducing multi-lag historical features, rolling statistical characteristics, and temporal variables, the model demonstrates strong responsiveness to meteorological variability and maintains high forecasting accuracy under different geographical, seasonal, weather, and operational conditions.

Nevertheless, certain limitations remain. The model relies heavily on the completeness and timeliness of historical data, and its predictive performance may degrade under conditions of missing data or inconsistent sampling frequencies.

In future applications, the integration of high-resolution remote sensing data and micro-meteorological sensor measurements could further enhance the model's responsiveness to abrupt changes. Moreover, incorporating terrain information and surface parameters into machine learning–based spatial downscaling methods may enable more accurate modeling of local meteorological and irradiance variations. These improvements would support the development of a more intelligent, accurate, and generalizable PV power forecasting model.

References

[1] Das U K, Tey K S, Seyedmahmoudian M, et al. Forecasting of photovoltaic power generation and model optimization: A review[J]. Renewable and Sustainable Energy Reviews, 2018, 81(1): 912-928..

[2] Gaboitaolelwe J, Zungeru A M, Yahya A, et al. Machine learning based solar photovoltaic power forecasting: A review and comparison[J]. IEEE Access, 2023, 11: 40820-40845.

[3] Markovics D, Mayer M J. Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction[J]. Renewable and Sustainable Energy Reviews, 2022, 161: 112364.

[4] Wang K, Qi X, Liu H. Photovoltaic power forecasting based LSTM-convolutional network[J]. Energy, 2019, 189: 116225.

[5] Mayer M J, Gróf G. Extensive comparison of physical models for photovoltaic power forecasting[J]. Applied Energy, 2021, 283: 116239.

[6] Li G, Xie S, Wang B, et al. Photovoltaic power forecasting with a hybrid deep learning approach[J]. IEEE Access, 2020, 8: 175871-175880.

[7] De Giorgi M G, Congedo P M, Malvoni M. Photovoltaic power forecasting using statistical methods: impact of weather data[J]. IET Science, Measurement & Technology, 2014, 8(3): 90-97.

[8] Li P, Zhou K, Lu X, et al. A hybrid deep learning model for short-term PV power forecasting[J]. Applied Energy, 2020, 259: 114216.

[9] Li G, Wang H, Zhang S, et al. Recurrent neural networks based photovoltaic power forecasting approach[J]. Energies, 2019, 12(13): 2538.

ISSN: 2710-0723

[10] Kaaya I, Ascencio-Vásquez J. Photovoltaic power forecasting[C]// Solar Radiation: Measurement, Modeling and Forecasting Techniques for Photovoltaic Solar Energy Applications. Cambridge: Academic Press, 2022: 195.

[11] Wang K, Qi X, Liu H. A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network[J]. Applied Energy, 2019, 251: 113315.