WEP Warwick Evans Publishing

Transactions on Engineering and Technology Research
ISSN: 2960-1797, eISSN: 2960-2246 | Volume 6
MECEME 2026

Photovoltaic power prediction based on BO-LightGBM collaborative prediction and K-means partition optimization

Jiashuo Jiang [#], Hongfu Chen [*, #]

School of Electronics, Electrical Engineering and Physics, Fujian University of Technology, Fuzhou, China

* Corresponding Author Email: 18750712417@163.com

[#]These authors contributed equally.

Abstract. As the global energy mix accelerates toward a clean and sustainable system, accurate and reliable photovoltaic power generation forecasting is crucial for ensuring grid stability and optimizing resource allocation. This study aims to improve the day-ahead forecasting performance of photovoltaic power generation under high-volatility scenarios. A collaborative forecasting model based on Bayesian Optimization (BO) and LightGBM is proposed, supplemented by a K-means clustering and partitioning optimization strategy. The model first applies K-means clustering to analyze total irradiance, a core driver, and successfully identifies high-risk periods of high power accompanied by high volatility, providing a divide-and-conquer foundation for the forecasting strategy. LightGBM is selected as the core forecasting engine to accurately capture the complex nonlinear relationship between meteorological characteristics and power output. Furthermore, the BO mechanism is innovatively introduced for global hyperparameter optimization, significantly improving model learning efficiency and forecast robustness. Feature importance analysis reveals the dominant feature sets consisting of light intensity, air pressure, and temperature. Experimental results demonstrate that this integrated model not only achieves superior forecast accuracy compared to traditional methods but also effectively controls forecast errors caused by peak fluctuations through a partitioning optimization strategy, providing a data-driven technical solution for optimizing resource allocation in power grid systems.

Keywords: Photovoltaic power generation; LightGBM; Bayesian Optimization; K-means.

1. Introduction

The global energy structure is undergoing a profound and irreversible transformation. However, its core support is still highly dependent on traditional fossil fuels [1]. Although these resources have provided strong impetus for the rapid expansion of the global economy in the past few decades, their inherent unsustainability and negative environmental impacts have posed a serious global challenge [2]. On the one hand, the limited reserves of fossil fuels have raised strategic concerns about long-term energy security; on the other hand, their combustion is the primary driving force of global climate change [3]. Authoritative data show that about 66% of global carbon dioxide emissions come from the use of fossil fuels, which directly exacerbates the greenhouse effect and the frequency of extreme weather events [4]. Therefore, accelerating the construction of a clean, sustainable, and highly resilient energy system has become a common strategic goal of the international community [5].

In the field of photovoltaic power generation prediction and factor exploration, researchers have conducted extensive exploration along different technical paths. Early work mainly focused on the application of traditional statistical and mathematical modeling methods. For example, Nakamoto Y. et al. attempted to use a linear regression model to estimate photovoltaic output and preliminarily identified the influencing factors based on a simple functional relationship [6-7]. Similarly, Matushkin D. et al. analyzed the determinants of photovoltaic power generation by constructing an analytical mathematical model [8]. In recent years, with the rapid development of machine learning technology, the focus of research has shifted to more powerful nonlinear modeling methods. Dai Y.

et al. tried to use ensemble learning methods such as Random Forest to explore influencing factors and make predictions [9]. At the same time, Ramli N. A. et al. conducted a prediction study on solar power generation based on the k-nearest neighbor method [10].

However, these traditional models are essentially linear and statistically driven, and their inherent simplicity makes it difficult to adapt to the complex and highly nonlinear system characteristics of photovoltaic systems driven by multi-source meteorological data. They show obvious limitations in dealing with the highly nonlinear and dynamic coupling relationship between meteorological and environmental factors and photovoltaic output, resulting in insufficient prediction accuracy and generalization ability. To overcome this key bottleneck, this study proposes and implements an advanced, efficient, and multi-stage machine learning framework. Specifically, this study uses LightGBM to establish a high-precision prediction model and perform global optimization of its hyperparameters through Bayesian optimization to improve the robustness of the model. More importantly, this paper innovatively introduced the K-means clustering algorithm to segment the core driving factors, enabling quantitative analysis of the nonlinear threshold effects of different radiation intervals on power generation, providing in-depth insights for the refined operation and management of photovoltaic systems.

2. Bayesian Optimization Lightgbm Photovoltaic Power Generation Prediction Model

2.1. The Structure of Lightgbm

The LightGBM model employed in this study is an efficient forecasting framework designed to leverage the powerful forecasting capabilities of the Lightweight Gradient Boosting Machine (LightGBM) to achieve excellent generalization performance in photovoltaic power generation forecasting tasks. As the core regression forecasting engine, LightGBM, with its histogram-based decision tree, GOSS, and EFB technologies, possesses a natural advantage in processing high-dimensional, nonlinear meteorological data. Its forecast accuracy is closely related to the proper configuration of hyperparameters. By optimizing these hyperparameters, the model's performance can be fully utilized, significantly improving forecast accuracy and robustness, ultimately achieving high-precision forecasts of power generation and quantitative analysis of key influencing factors.

In the t-th iteration, the goal of LightGBM is to find a new decision tree f_t, to minimize the loss function $\mathcal{L}^{(t)}$ after Taylor expansion approximation. At the same time, the regularization term $\Omega(f_t)$ is introduced to control the model complexity and avoid overfitting. Specifically, the optimization goal can be expressed as minimizing the following function by constructing the optimal f_t in the current iteration:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{1}$$

Among them, $\mathcal{L}^{(t)}$ is the approximate loss based on Taylor expansion, which reflects the optimization direction of the model for the prediction error at the current iteration step; $\Omega(f_t)$ is the regularization term for the new tree f_t, which is used to balance the model performance and complexity. Where g_i represents the first-order gradient of the sample i.

$$g_i = \left[\frac{\partial L(y_i, F)}{\partial F} \right]_{F=F_{t-1}(x_i)} \tag{2}$$

where h_i refers to the second-order gradient of the sample i.

$$h_i = \left[\frac{\partial^2 L(y_i,F)}{\partial F^2}\right]_{F=F_{t-1}(x_i)} \tag{3}$$

where $\Omega(f_t)$ is the complexity regularization term of the tree. If the tree f_t has J leaf nodes, then:

$$\Omega(f_t) = \gamma J + \frac{1}{2}\lambda \sum_{j=1}^{J} w_j^2 \tag{4}$$

where γ represents the minimum loss reduction parameter for leaf node splitting, and λ represents the L2 regularization parameter.

After determining the tree structure, the optimal output value w_j^* of the j-th leaf node R_j is:

$$w_j^* = -\frac{\sum_{i\in R_j} g_i}{\sum_{i\in R_j} h_i + \lambda} \tag{5}$$

Gain formula for feature splitting. This is the basis of LightGBM feature importance, which measures the maximum reduction in the loss function after the split:

$$\text{Gain} = \frac{1}{2}\left[\frac{\left(\sum_{i\in R_L} g_i\right)^2}{\sum_{i\in R_L} h_i + \lambda} + \frac{\left(\sum_{i\in R_R} g_i\right)^2}{\sum_{i\in R_R} h_i + \lambda} - \frac{(\sum_{i\in R} g_i)^2}{\sum_{i\in R} h_i + \lambda}\right] - \gamma \tag{6}$$

where $Importance_{\text{Gain}}(A)$ refers to the cumulative gain of feature A in all trees.

2.2. The Structure of Bayesian Optimization

Bayesian optimization is a sequential decision-making strategy for efficiently finding the global optimal solution to expensive black-box functions. It is particularly well-suited for evaluation-intensive problems such as hyperparameter tuning in machine learning. Its core concept is to not directly and time-consumingly evaluate the objective function F, but to construct a probabilistic surrogate model to approximate the response surface of F. The GP model provides the mean and variance of the performance for any point θ in the hyperparameter space. BO then uses a mechanism called the acquisition function. Through this intelligent balancing act, BO can quickly and efficiently converge to the global optimal solution of the objective function with far fewer evaluations than traditional methods.

The goal of Bayesian optimization is to efficiently find the global optimal solution x^* of the black−box objective function $f(x)$, where x is the set of hyperparameters of LightGBM.

$$x^* = \underset{x\in\mathcal{X}}{\text{argmax}} f(x) \tag{7}$$

where $f(x)$ represents the objective function, $and\ \mathcal{X}$ represents the search space of hyperparameters. Surrogate Model Bayesian optimization uses a Gaussian process as a surrogate model to model the probability distribution of the target function $f(x)$. The Gaussian process defines the joint probability distribution of the function value, which is determined by the mean function $m(x)$ and the covariance function $k(x, x')$:

$$f(x) \sim \mathcal{GP}\big(m(x), k(x,x')\big) \tag{8}$$

where $m(x) = \mathbb{E}[f(x)]$ is the mean of the target function at x, and $k(x,x')$ is the covariance function that measures the similarity between the function values at x and x'. Commonly used kernels include the squared exponential kernel:

$$k(x,x') = \sigma_f^2 \exp\left(-\frac{1}{2l^2} \parallel x - x'^{\parallel^2}\right) \tag{9}$$

where σ_f^2 represents the signal variance and l represents the length scale parameter.
GP prediction. After collecting historical observation points $\mathcal{D}_t = \{(x_1, y_1), \ldots, (x_t, y_t)\}$, GP can predict the posterior distribution $\mathcal{P}(f(x)|\mathcal{D}_t)$ at any new point x, with the formulas of its mean $\mu_t(x)$ and variance $\sigma_t^2(x)$ as follows:

$$\mu_t(x) = \mathrm{k}_t(x)^{\mathrm{T}}(\mathrm{K}_t + \sigma_n^2\mathrm{I})^{-1}\mathrm{y}_{1:t} \tag{10}$$

$$\sigma_t^2(x) = k(x,x) - \mathrm{k}_t(x)^{\mathrm{T}}(\mathrm{K}_t + \sigma_n^2\mathrm{I})^{-1}\mathrm{k}_t(x) \tag{11}$$

where K_t represents the $t \times t$ covariance matrix, where $(\mathrm{K}_t)_{ij} = k\big(x_i, x_j\big)$. $\mathrm{k}_t(x)$ represents the covariance vector between x and all historical observations. σ_n^2 represents the variance of the observation noise.

The acquisition function $a(x)$ is used to guide the search, balancing $\mu_t(x)$ and $\sigma_t^2(x)$. The value x_{t+1} that maximizes $a(x)$ will be the next hyperparameter to evaluate. One of the most commonly used acquisition functions is the Upper Confidence Bound:

$$x_{t+1} = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmax}}[\mu_t(x) + \kappa\sigma_t(x)] \tag{12}$$

where $\mu_t(x)$ is the mean of the predictions. $\sigma_t(x)$ is the standard deviation of the predictions. κ is the balance parameter between exploration and exploitation. The larger κ is, the more the model tends to explore uncertain areas. Another commonly used metric is Expected Improvement:

$$x_{t+1} = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmax}}\mathbb{E}\big[\max\big(0, f(x) - f(x^+)\big)\big] \tag{13}$$

where x^+ is the currently observed best point, and the specific calculation of $\mathbb{E}[\cdot]$ involves the cumulative distribution function $\Phi(\cdot)$ and the probability density function $\phi(\cdot)$ of the standard normal distribution.

$$\mathrm{EI}(x) = \sigma_t(x)[z\Phi(z) + \phi(z)] \tag{14}$$

$$z = \frac{\mu_t(x) - f(x^+)}{\sigma_t(x)} \tag{15}$$

2.3. K-Means Clustering

The K-means clustering algorithm is a classic, partitioning-based unsupervised learning technique. Its core goal is to partition N data points x into a pre−specified number of K clusters $S = \{S_1, S_2, \ldots, S_K\}$ in a given dataset. The mathematical basis of the algorithm lies in the optimization problem: by iteratively searching for the best cluster partitions S_i and cluster centers μ_i, the objective

function called the within-cluster sum of squared errors or distortion measure is minimized. This objective function J quantitatively measures the clustering quality.

The mathematical expression of tightness is as follows:

$$J = \sum_{i=1}^{K} \sum_{\mathbf{x} \in S_i} \|x - \mu_i\|^2 \tag{16}$$

where μ_i is the arithmetic mean of all data points in the i−th cluster S_i.

The algorithm achieves a local optimum through an expectation-maximization-style iterative process: the assignment step assigns each data point to the cluster center with the closest Euclidean distance to it; the update step recalculates the mean of all data points within each cluster as the new cluster center. This process repeats until the cluster center no longer moves significantly. Due to its linear time complexity and ease of interpretation, K-means has become a highly efficient and preferred algorithm for large-scale data clustering analysis.

3. Results

The data used in this article is sourced from https://www.scidb.cn

3.1. The Result of Lightgbm+BO

This study utilized Bayesian optimization to efficiently optimize the hyperparameters of the LightGBM model, ensuring that the model accurately predicts photovoltaic power generation under the optimal configuration. As shown in Table 1, by analyzing the feature importance of the optimization model output, the relative contributions of the six meteorological features to the prediction results are quantified, clearly revealing the key drivers of PV power plant efficiency. These results provide valuable guidance for the development of photovoltaic power prediction models and the design of meteorological monitoring systems.

Table 1. LightGBM model meteorological feature importance ranking

factors	Gain size
totalirrad	1350
diffuseirrad	1327
pressure	1325
temperature	1090
winddirection	489
windspeed	419

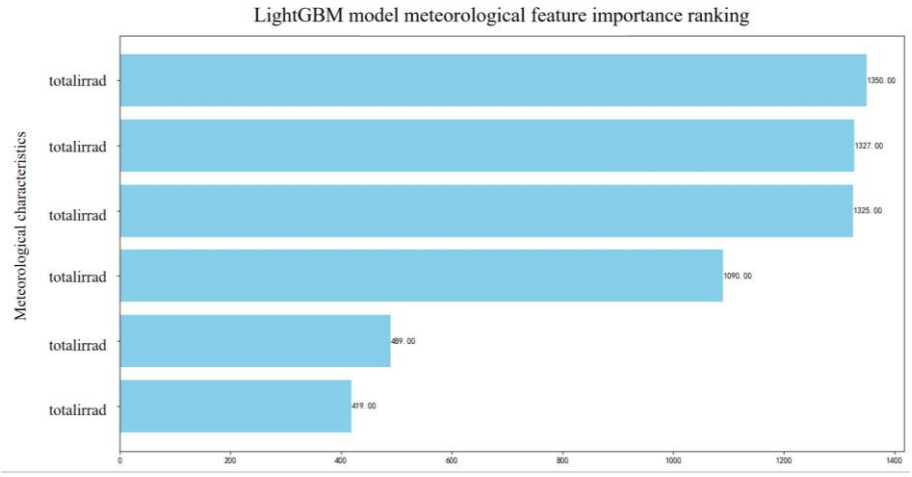

Figure 1. LightGBM model meteorological feature importance ranking

As shown in Figure 1, the impact of meteorological characteristics exhibits a clear hierarchical structure. Solar irradiance, directly related to the photovoltaic effect, is the most important input feature, with the highest gain of 1350.00, followed by diffuse irradiance, with a gain of 1327.00. The high weights of these two indicators not only validate the physical basis of photovoltaic power generation at the model level—light energy input determines electrical energy output—but also highlight the critical importance of accurate irradiance data for ensuring high-precision model prediction performance based on Bayesian optimization.

Following this are a series of key indirect factors. Notably, the gain for air pressure reaches 1325.00, nearly equaling its contribution to irradiance. This significant importance may stem from the impact of air pressure changes on atmospheric transparency, aerosol distribution, and light attenuation, which indirectly modulates the effective energy reaching the photovoltaic panel surface. This suggests that under complex weather conditions, air pressure data is considered a key compensating variable by the model to improve forecast accuracy. The gain for temperature, 1090.00, confirms the negative temperature coefficient of photovoltaic module efficiency: increasing temperature leads to a decrease in open-circuit voltage and fill factor, thereby reducing actual output power.

In stark contrast, the contributions of wind direction (gain: 499.00) and wind speed (gain: 419.00) are significantly lower, placing them in the less influential tier. Although wind speed can mitigate efficiency losses caused by excessive temperatures to some extent through convective cooling of the panels, its marginal contribution to the model prediction is far less than that of irradiance, air pressure, and temperature, which directly influence solar energy conversion.

In summary, the integrated model structure of Bayesian optimization and LightGBM emphasizes that solar irradiance, air pressure, and temperature constitute a three-dimensional core feature set that must be accurately collected and processed in photovoltaic power forecasting. The model's dependence on air pressure particularly highlights the necessity of integrating multidimensional meteorological variables to improve the accuracy of machine learning forecasting models.

3.2. The Result of K-Means

To deeply explore the total irradiance, this study adopts the K-means clustering method to segment the irradiance data, thereby defining four irradiance intervals with significant differences. By calculating the average power generation within each interval, we obtained a quantitative relationship between irradiance and power generation performance. The analysis results, as shown in Figure 2, reveal a highly significant nonlinear positive correlation between irradiance and power generation, with an overall trend that exhibits an almost monotonically increasing characteristic.

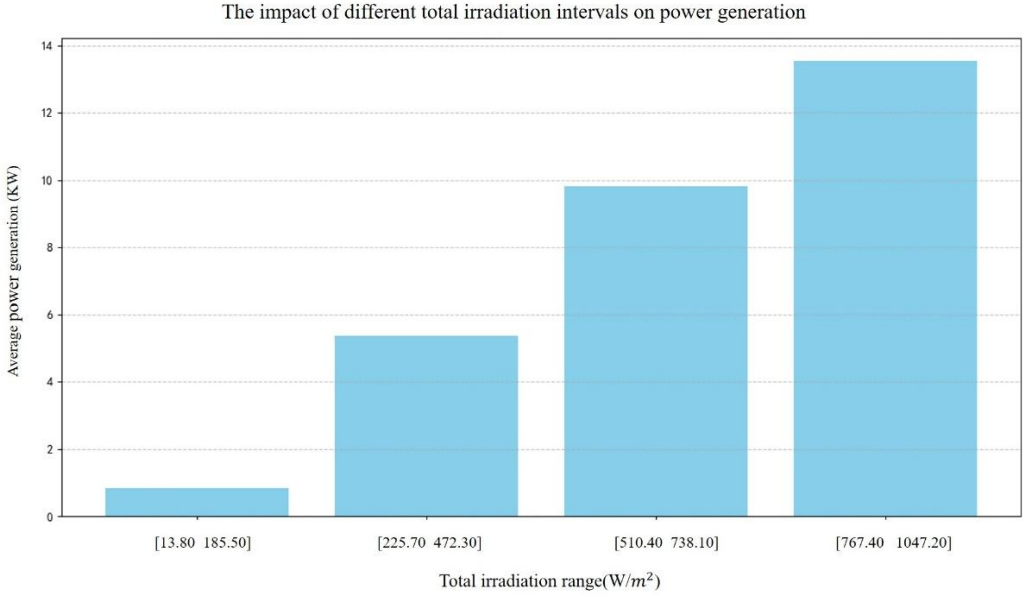

Figure 2. The impact of different total irradiation intervals on power generation

In the lowest irradiance range [13.80, 185.50] W/m², the system's average power generation is only about 0.9 kW, indicating that the system's efficiency is extremely low under weak light or shaded conditions. This initial range sets a baseline for subsequent analysis. As irradiance increases, the average power output shows a segmented stepwise growth. In the second range [225.70, 472.30] W/m², the average power rises significantly to about 5.3 kW, marking the system's transition into an effective power generation state. When the irradiance further increases to the third range [510.40, 738.10] W/m², the average power jumps again to around 9.8 kW. Finally, in the highest irradiance range [767.40, 1047.20] W/m², the average power reaches a peak of approximately 13.7 kW. This stepwise growth pattern directly quantifies the average contribution of different irradiance levels to power generation. The results of this bar chart strongly confirm that total irradiance is the most critical driving factor in determining photovoltaic output power and clearly reveal the segmented nonlinear growth characteristics of the photovoltaic system under different lighting conditions, providing key empirical evidence for performance assessment and optimization of control strategies for photovoltaic power plants.

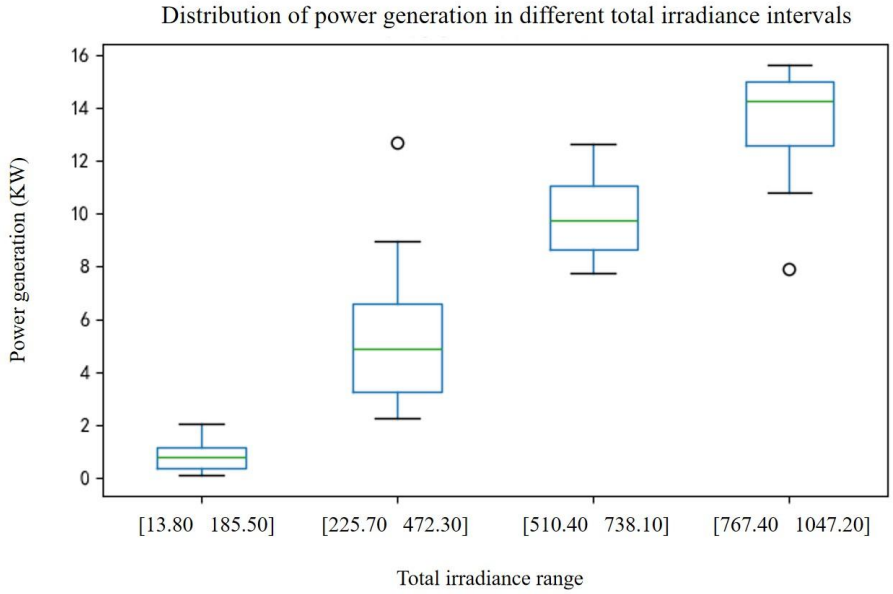

Figure 3. Distribution of power generation in different total irradiance intervals

The boxplot results are shown in Figure 3. We used boxplots to provide an in-depth visualization of the statistical distribution, dispersion, and robustness of power generation within the same time period. Unlike bar charts that only provide average values, boxplots reveal key statistical characteristics within each interval, such as the median, quartiles, extreme values, and outliers. The boxplot provides an in-depth visualization of the statistical distribution, dispersion, and robustness of power generation within the four irradiance intervals defined by K-means clustering. The plot further reveals key statistical characteristics within each interval, such as the median, quartiles (Q1 and Q3), extreme values, and outliers. Analysis of data dispersion reveals the following key findings: the median trend is consistent with the mean trend, steadily increasing with increasing irradiance intervals, confirming the role of core drivers. However, variability analysis shows that the boxplot for the lowest irradiance interval is shorter due to its smaller power base, indicating less power fluctuation.

4. Conclusions

This study aims to improve the performance of accurate power forecasting for photovoltaic power plants under high-volatility scenarios by constructing a collaborative forecasting model based on Bayesian optimization (BO) and LightGBM, supplemented by a K-means clustering divide-and-conquer strategy. By applying K-means clustering to the core driver, total irradiance, we effectively identify and isolate interval 4, where high power output is accompanied by high volatility (with a power increase exceeding 34 times), thereby overcoming the limitations of single models in high-risk

periods. The study innovatively introduces the BO mechanism for hyperparameter optimization in LightGBM, significantly improving the model's learning efficiency and robustness. Experimental results demonstrate that the integrated model achieves optimal performance for photovoltaic power forecasting, demonstrating the significant advantages of ensemble learning and intelligent optimization algorithms in mining complex meteorological features. Furthermore, the model clearly reveals the dominant feature set consisting of light intensity, air pressure (gain 1325.00), and temperature. In practical applications, this model, through its partitioning optimization strategy, helps improve the reliability of forecast results and provides a data-driven technical solution for optimizing resource allocation in power systems.

This study has three limitations: the sample size does not fully cover rare meteorological subtypes, limiting the model's generalizability; the LightGBM model is highly sensitive to the distribution of training data, potentially limiting its generalization when encountering novel meteorological patterns; and the forecasting strategy relies on static K-means clustering boundaries, making it difficult to meet real-time requirements. Future work on this photovoltaic forecasting model, which combines LightGBM with Bayesian optimization, will focus on exploring multi-site transfer learning to improve generalization capabilities, employing hybrid ensemble strategies to extract complex time series features, deepening scientific understanding through uncertainty quantification and causal analysis, and applying parallel BO to improve optimization efficiency.

References

[1] Gajdzik B, Wolniak R, Nagaj R, et al. The influence of the global energy crisis on energy efficiency: A comprehensive analysis[J]. Energies, 2024, 17(4): 947.

[2] Kjärstad J, Johnsson F. Resources and future supply of oil[J]. Energy policy, 2009, 37(2): 441-464.

[3] Filonchyk M, Peterson M P, Zhang L, et al. Greenhouse gases emissions and global climate change: Examining the influence of CO2, CH4, and N2O[J]. Science of The Total Environment, 2024, 935: 173359.

[4] Kumar A, Singh P, Raizada P, et al. Impact of COVID-19 on greenhouse gases emissions: A critical review[J]. Science of the total environment, 2022, 806: 150349.

[5] Dai Y, Wang Y, Leng M, et al. LOWESS smoothing and Random Forest based GRU model: A short-term photovoltaic power generation forecasting method[J]. Energy, 2022, 256: 124661.

[6] Nakamoto Y, Eguchi S. How do seasonal and technical factors affect generation efficiency of photovoltaic power plants[J]. Renewable and Sustainable Energy Reviews, 2024, 199: 114441.

[7] Chang R, Bai L, Hsu C H. Solar power generation prediction based on deep learning[J]. Sustainable energy technologies and assessments, 2021, 47: 101354.

[8] Matushkin D, Bosak A, Kulakovskyi L. Analysis of factors for forecasting electric power generation by solar power plants[J]. 2020.

[9] Ramli N A, Hamid M F A, Azhan N H, et al. Solar power generation prediction by using k-nearest neighbor method[C]//AIP Conference Proceedings. AIP Publishing LLC, 2019, 2129(1): 020116.

[10] Dai Y, Wang Y, Leng M, et al. LOWESS smoothing and Random Forest based GRU model: A short-term photovoltaic power generation forecasting method[J]. Energy, 2022, 256: 124661.