Available online at www.HighTechJournal.org

HighTech and Innovation Journal

ISSN: 2723-9535

Vol. 5, No. 4, December, 2024

Research on Power Consumption Data Prediction of Distributed Photovoltaic Power Station

Junfeng Yao [1], Chun Xiao [1, 2*], Junbo Hao [3], Xiaoxia Yang [1]

[1] *State Grid Shanxi Marketing Service Center, Taiyuan, Shanxi, 030032, China.*

[2] *Taiyuan University of Technology, Shanxi 030024, China.*

[3] *State Grid Shanxi Integrated Energy Service Co.,LTD., Taiyuan, Shanxi, China.*

Received 10 May 2024; Revised 30 October 2024; Accepted 09 November 2024; Published 01 December 2024

Abstract

At present, the construction of distributed photovoltaic power stations in China lacks systematic and comprehensive preliminary planning; The construction cost exceeded the estimated estimate. After the completion of the project economic benefits cannot reach the expected income, project operating costs exceed expectations and other problems. In order to solve these problems, it is urgent to reasonably forecast the electricity consumption data of distributed photovoltaic power stations. Therefore, in order to solve these problems, a reliable model is established to predict the electricity consumption data of distributed photovoltaic power stations, and the indirect prediction method is used to forecast, that is, the irradiance of medium and long-term time scales is predicted by historical meteorological data, and then the system electricity consumption data is obtained. Among them, the model used is the Long short-term memory (LSTM) neural network model. Under the effect of this model, the electricity consumption data prediction of distributed photovoltaic power stations is carried out. The result shows that the MAPE of monthly prediction is 3.5%, and the annual prediction is 1.1%, which has ideal prediction accuracy and can achieve better prediction effect. This indirect forecasting method breaks the shackles of traditional forecasting methods, avoids the problems of data collection and other aspects, and is a new development trend and the performance of scientific and technological progress, which is conducive to the development of distributed photovoltaic power stations.

Keywords: Distributed Photovoltaic Power Station; Forecast; Electricity Consumption Data; LSTM.

1. Introduction

In order to alleviate the worsening climate and environmental problems, the development of renewable resources has become a top priority. Large-scale use of renewable energy has become one of the important measures to adjust the energy structure, ensure energy security, strengthen environmental protection, and achieve sustainable development. Most of the world's renewable energy comes from solar energy, and many researchers in the world continue to study solar energy development technology. In recent years, new technologies on solar energy development and utilization have developed rapidly, and related industries with solar energy as the core technology have become one of the fastest growing industries [1-3]. Therefore, photovoltaic power generation (PV) has been widely concerned in the world for its characteristics of pollution-free, renewable, low cost and mature technology. In this regard, Charbonnier et al. (2024) studied the household power data generator (HEDGE) to generate actual data of photovoltaic power generation with the

* Corresponding author: xiaochun@sx.sgcc.com.cn

http://dx.doi.org/10.28991/HIJ-2024-05-04-05

help of this tool, and then understand the electricity consumption during the period through the analysis of the data. In the end, the feasibility of the tool is tested, and the results show that HEDGE can quickly generate the required energy data series without being affected by contour size and cluster, etc., and has high application value and practicability [4].

In the same year, Liang et al. (2024) studied a comprehensive system for the development of solar drive, which is composed of solar photovoltaic, compressed air energy storage, CAES storage and other modules, which can better reduce the cost of electricity at night. In order to test the feasibility of the system design, an empirical study has been conducted, and the results show that under given conditions, the system can reasonably control the expenditure cost, have objective economic benefits, and reduce the loss of the photovoltaic system, which has certain feasibility [5]. With the increasing installed capacity of photovoltaic power generation, the variability and uncertainty of photovoltaic power generation output have caused great obstacles to the stable operation of the power system. Photovoltaic power generation output for different time scales will show variability, and it is difficult to predict. Domestic scholars Qiu et al. (2024) proposed a photovoltaic power generation prediction method based on improved variational mode decomposition and ensemble learning, aiming at the problem of power generation prediction performance caused by non-stationary photovoltaic power generation data. In order to test the feasibility of the method, a quantitative study was carried out, and the results showed that the mean square error, mean absolute error and determination coefficient values of the proposed method for predicting photovoltaic power generation on the open data set were 0.223, 0.338, and 0.9797, respectively, which had higher prediction accuracy and smaller error compared with other methods. It can be used in electricity prediction research [6].

Yu et al. (2022) proposed a prediction model based on neural computing to solve the problem of decentralization of power supply system caused by the continuous integration of renewable energy in power network. The robustness of the model is tested by simulation experiments. The results show that the Relative Mean Absolute Error (RMAE) of the model can reach 2.5% in summer and 0.5% in winter. Over the full year, by using a reduced mean of input class characteristics, RMAE for PV and wind can reach 1.7% and 4.9%, respectively. The implementation of this method provides strong support for the integration of regional renewable energy, and also provides a new solution to the volatility and uncontrollability of renewable energy [7]. In order to promote the development of photovoltaic power generation forecasting, Lu et al. (2020) proposed a research on regional power grid photovoltaic power generation forecasting based on support vector machine. A quantitative study was conducted, and the results showed that the prediction accuracy of the experimental group was 92.37%, that of the prediction method based on the improved firefly algorithm was 85.43%, and that of the prediction method based on the grey correlation and sparrow optimization algorithm was 74.66%, confirming the feasibility of the proposed method [8]. Most of the existing literature on photovoltaic power generation forecasting aims at solving its uncertainty problem. The basic principle of photovoltaic power generation prediction is to build a model to fit the relationship between the input characteristic data and the power output of the photovoltaic power generation system combined with the characteristics of the photovoltaic power generation system and geographical location, so as to realize the photovoltaic power generation prediction [9, 10]. In summary, it is not difficult to find that the prediction of photovoltaic power generation mostly uses ground weather station data, satellite image data, photovoltaic system operation data and numerical weather forecast data, while the research on irradiance is relatively few, and the prediction time is short. In order to improve the research in this aspect, this paper will take the light amplitude as the prediction index of photovoltaic power consumption data. The correlation analysis is carried out, and the final prediction results are given to test its feasibility.

2. Distributed Photovoltaic Power Station

Energy is the cornerstone of photovoltaic power generation, energy as the main source, with a stepped structure of energy utilization system [11-13]. Distributed photovoltaic power generation technology has become an inevitable trend in the development of global energy technology due to its characteristics of high energy efficiency, strong reliability and low environmental pollution. At present, China has only conducted preliminary exploration of distributed photovoltaic power generation technology in Beijing, Shanghai and Guangdong. So far, more than 40 power stations have been built, such as the Beijing Olympic Games Media Village, Shanghai Pudong International Airport, Guangzhou University Town and Sichuan Hope Group's deep blue and green energy Center, but relevant research is still blank. Statistics show that at present, China has built a gas distributed power supply, mainly for gas power generation, with an installed capacity of about 54×105 kW. Clean and pollution-free photovoltaic distributed power generation is different from traditional intensive power generation, it is a production capacity method to build small-scale generator sets near the client to meet the needs of customers as the main goal, while connecting the surplus electricity to the grid and cooperating with the existing grid system to obtain government subsidies [14, 15].

At present, the technical system of domestic distributed photovoltaic power generation projects has been relatively mature, running well, and has a good momentum of development [16-18]. Photovoltaic modules convert solar energy into electricity by exploiting the photovoltaic effect of semiconductor materials. The solar energy received by photovoltaic modules usually refers to the amount of radiation actually received by the panel surface, and the measure of radiation is usually expressed by irradiance. Irradiance is defined as the solar radiant energy received by a photovoltaic module per unit area in a unit time period. Irradiance is a direct meteorological factor that determines the final output of

electric energy in photovoltaic power generation systems. Since the amount of solar radiation received by the ground is greatly affected by clouds and particles in the air during transmission, the irradiance is random and intermittent, Resulting in photovoltaic power generation system efficiency is not guaranteed. Under the influence of solar irradiance, This kind of power generation efficiency will also show periodic changes. In general, when the light intensity is higher, the greater the amount of radiation, the greater the photovoltaic power consumption data. After determining the area of the photovoltaic module, its output of electrical energy is proportional to the irradiance.

3. Prediction of Electricity Consumption Data by Algorithm Model

3.1. Model Overview

3.1.1. Multiple Regression Forecasting Model

Regression algorithm is the basis of the development of most artificial intelligence algorithms, and the model is simple and practical, suitable for most systems [19]. As far as regression algorithm model is concerned, the input parameters are different, and its definition will be different, such as multiple regression, unary regression, etc. At the same time, different spatial distribution will have different definitions, such as nonlinear regression, linear regression and so on. Unitary linear regression is a model that takes the main influencing factors as independent variables. In practical applications, because the dependent variables are generally affected by multiple factors, two or more parameters of independent variables will be used, So multiple regression can solve the problems in practice. If there is a linear relationship between different independent variables and dependent variables, then the analysis can be called multiple linear regression. If there is a nonlinear relationship between the two, then this analysis can be called multiple nonlinear regression, and the expression of multiple regression is shown in Equation 1.

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_j X_{ji} \quad (1)$$

where $i = 1,2,\cdots,n, j = 1,2,\cdots,k$. According to the actual irradiance, the least square method is selected as the loss function, and the square error between the predicted irradiance value and the actual irradiance value is calculated by fitting the historical meteorological data and the historical irradiance data.

3.1.2. Persistent Prediction Model

Different from multiple regression model, the input parameters and equation formula of the persistent prediction model are fixed, and the input parameters mainly include irradiance data from historical monitoring of local weather stations, real-time meteorological data and theoretical calculation data [20]. The solution formula is as follows:

$$p_{t+1} = a + b \cdot T \cdot G_{t+1} + c \cdot G_{t+1} + d \cdot G_{t+1}^2 \quad (2)$$

where, the empirical coefficients *a*, *b*, *c*, and *d* can be obtained by means of fitting. The prediction of total radiation should be completed on the basis of time series model and radiometric empty model. This article uses the persistence model as the baseline mode.

3.1.3. Support Vector Regression Machine

Unlike traditional learning methods, SVM solves the optimal value by minimizing the structural risk function. The algorithm uses the generalized error rate on the test set to find the boundary dependent on VC dimension while reducing the error rate of the model. Under the condition that the label is separable, it is necessary to make the second term as small as possible while ensuring that the previous term is equal to 0. Because SVM has good generalization ability, it can better solve classification problems even if it does not have relevant knowledge in other fields.The central idea is to first determine the nonlinear contrast rays, and then project the original data into the high-latitude feature space to obtain the best classification results, and then expand the separation boundary between samples. At the conceptual level, a support vector can be regarded as the data point closest to the decision plane, and the location of the data point can be used to determine the location of the optimal classification hyperplane. For the irradiance prediction problem, the relationship between irradiance and meteorological factors is usually nonlinear. In order to better fit the nonlinear trend on the training data set, it is necessary to deal with the linear regression problem and integrate it into the nonlinear regression problem, and the realization of both needs the support of kernel function. In order to better deal with nonlinear problems, the Gaussian radial basis kernel function will be selected in this chapter, as follows:

$$\mathrm{K}(x, x_i) = exp\left(\frac{x - x_i}{\sigma^2}\right) \quad (3)$$

Usually linear SVR contains multiple linear equations, for a given sample x,y, there is the following formula:

$$\mathrm{y} = \mathrm{f(x)} = \langle \mathrm{w}, \mathrm{x} \rangle + \mathrm{b} \quad (4)$$

where $\mathrm{w} \in \mathrm{R}^{\mathrm{a}}$ is the weight of the vector and $\mathrm{b} \in \mathrm{R}$ is the constant quantity. The loss function of SVR meets the structural risk minimization principle, and the parameters of the optimal solution of SVR must meet the conditions that make Equation 5 take the minimum value.

$$R[f] = \int (y - f(x))^2 P(x, y) dxdy \quad (5)$$

Since the probability distribution function $P(x, y)$ in the above formula is unknown, formula 6 is usually used instead of the structural risk function to solve the minimum Be worth.

$$\emptyset(w, \xi^*, \xi) = \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{1}(\xi_i^* + \xi_i) \quad (6)$$

In the formula, the fixed value of the given penalty factor is represented by c; The lower limit of the relaxation variable is represented by ξ_i^*; The upper limit of the relaxation variable is represented by ξ_i. The constraints of Equation 7 are as follows:

$$\begin{cases} y_i - (\langle w, x_i \rangle + b) \leq \varepsilon + \xi_i \\ (\langle w, x_i \rangle + b) - y_i \leq \varepsilon + \xi_i^* \\ \xi_i^*, \xi_i \geq 0 \end{cases} \quad (7)$$

In order to solve the above constraints, the Lagrange multiplication algorithm can be constructed to obtain the following Equation 8.

$$\begin{cases} min\frac{1}{2}\sum_{i,j=1}^{1}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j \rangle + \sum_{i=1}^{1}\alpha_i(\varepsilon - y_i) + \sum_{i}^{1}\alpha_i^*(\varepsilon + y_i) \\ \sum_{i=1}^{1}(\alpha_i^* - \alpha_i) = 0 \end{cases} \quad (8)$$

In summary, the support vector regression machine equation can be obtained, as shown in Equation 9.

$$f(x) = \sum_{i}^{1}(\alpha_i^* - \alpha_i)(x_i, x) + b \quad (9)$$

When constructing a nonlinear SVR,The nonlinear problem should be transformed into a linear problem, and the SVR function in this case is shown in Equation 10.

$$f(x) = \sum_{i}^{1}(\alpha_i^* - \alpha_i)K(x_i, x) + b \quad (10)$$

The SVM model prediction process is shown in Figure 1.

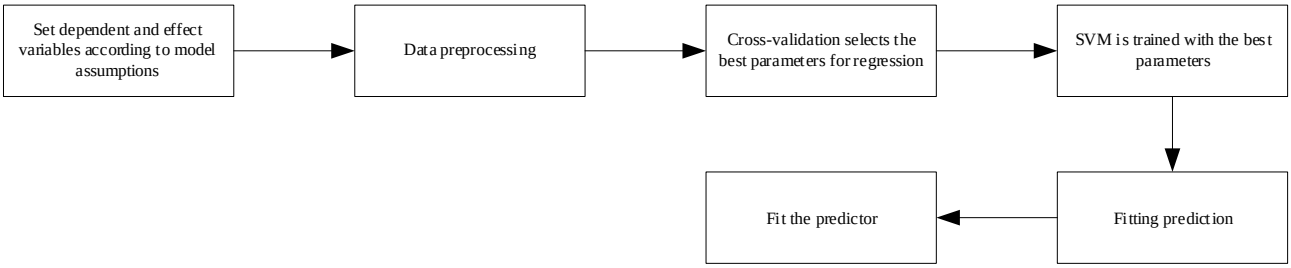


Figure 1. SVM model prediction process

3.2. Irradiance Prediction Model Based on Long- and Short-Term Memory Neural Network

LSTM neural network has long-term memory function, which can deeply explore the long-term dependency relationships and trends of limited data samples. It can also solve the problem of recurrent neural networks (RNNs) losing their ability to perceive distance moments due to the disappearance of gradients during training. LSTM can solve this kind of problem. Specifically, special neurons are used for permanent memory, while long-term relationships are captured, extending the service life of information and giving full play to the depth of computation. LSTM neural network can deeply evaluate the long-term dependence and trend relationship of limited data samples, and is suitable for medium and long term irradiance prediction of limited data samples. According to the historical irradiance data and selected key meteorological factors data, Based on LSTM, the irradiance prediction model is constructed based on the medium and long term timeline. The formula is expressed as follows:

$$\big(W(t+1),\ W(t+2),\cdots,W(t+m)\big)=F\big(W(t),W(t-1),\cdots,W(t-n),x(t),x(t-1),\cdots x(t-n)\big) \quad (11)$$

where, $W(t+1),W(t+2),\cdots,W(t+m)$ represents the predicted irradiance data; $W(t),W(t-1),\cdots,W(t-n)$ represents the current and previously actually measured irradiance data values, n is determined by traversing the sample set of experimental data, m is 12; $x(t),x(t-1),\cdots x(t-n)$ stands for current and past actual measured key meteorological factor data. On the long-term scale, the accuracy of numerical weather prediction is less satisfactory, using the meteorological influence factor data of the numerical weather forecast as the model input of the forecast period will bring large prediction errors. Therefore, The input to the irradiance prediction model is current and previous months irradiance data as well as key meteorological impact factor data, and the output is the monthly irradiance of the following year.

3.2.1. Recurrent Neural Network

As far as previous neural networks are concerned, information processing lacks linkage, focusing on information processing at the current moment and lacking memory ability. Recurrent neural networks (RNNs), on the other hand, retain information from the present moment for the next moment of information processing.It has a certain memory ability and provides a simpler processing channel for all kinds of information processing. However, this can not meet the power consumption data prediction in this paper, because the long time span and small sample data limit the processing performance of the network. It is difficult to solve the problem of long period dependence.

3.2.2. Forecast Process

First you'll enter the oblivion door level. The forgotten information can be calculated through the gate layer. In the long-term irradiance prediction task, the current irradiance prediction needs to rely on the node data of the previous time period of the same time series. At the same time, the gate layer can read the current input information x_t and the output information h_{t-1} of the previous layer, and after processing by the sigmoid function, obtain the corresponding value f_t, the size of which is between 0 and 1, and then f_t will be transmitted to the current unit state C_{t-1} in real time.The actual meaning of f_t is as follows: "1" means that all states are reserved; "0" means all forgotten states, so the corresponding expression is shown in Equation 12.

$$f_t=\sigma\big(W_f[h_{t-1},x_t]+b_f\big) \quad (12)$$

In the formula, The vector in parentheses is the sum of the vectors; The weight matrix is represented by W_f; sigmoid is represented by σ; The offset term is represented by b_f; Second, enter the door layer. The gate layer has two functions: one is to determine the input value of σ; Second, it has the creation function, which can complete the creation of new candidate vectors and implant them into the input value of tanh. The specific calculation formula of LSTM is shown as follows.

$$i_t=\sigma(W_i[h_{t-1},x_t]+b_i) \quad (13)$$

$$\tilde{C}_t=tanh(W_c[h_{t-1},x_1]+b_c) \quad (14)$$

Among them, there are two kinds of weight matrix: one is W_i, which represents the weight matrix of the first part; The second is W_c, which represents the weight matrix of the second part; There are two offset terms, b_i and b_c.Next comes the update door layer. With the help of the gate layer, the update process of the previous cell state can be realized, and the product of all output values of the forgetting gate and the sum of the product of input gate state value and output value is the current cell state value, as shown Equation 15:

$$C_t=f_t*C_{t-1}+i_t*\tilde{C}_t \quad (15)$$

Finally, Finally, output the gate layer. Through this gate layer, the parameter information output by σ can be calculated, and the obtained result can be multiplied with tanh, then the final result can be obtained. The expression is shown as follows.

$$\begin{cases}O_t=\sigma(W_o[h_{t-1},x_t]+b_o)\\ h_t=O_t*tanh(C_t)\end{cases} \quad (16)$$

where W_o represents the weight matrix of the gate layer, and b_o represents the offset term.

4. Experiment and Results

To learn more about the model's predictions, R-square (autocorrelation coefficient) was introduced for evaluation, and the corresponding calculation formula was shown in Equation 17.

$$\mathrm{R}^2=1-\frac{\sum(\mathrm{y}-\hat{\mathrm{y}})^2}{\sum(\mathrm{y}-\bar{\mathrm{y}})^2} \quad (17)$$

The ratio of 1 minus y (the actual value) to the variance (unexplained deviation) of the regression equation to the total variance of y (the actual value), that is, the part of the fit equation that cannot be explained, So, the more the value of R-square approaches 1, the more the parsing approaches the actual value. In forecasting practice, the model that makes R-square the highest is often adopted.

4.1. Medium- and Long-Term Irradiance Prediction

In this experiment, the irradiance data and related meteorological data include the measured irradiance values and local meteorological data of the entire field from 2017 to 2023, with a sampling frequency of 15 minutes. The sampled data is calculated and processed to convert it into the required monthly mean meteorological data and irradiance data. The model is constructed on the basis of LSTM, and the performance of the model is tested by comparison, including the persistent prediction model, multiple regression prediction model and support vector machine model.Compared with these models, the predictive model was constructed by changing the input conditions and the optimal input was selected. Figure 2 shows the fitting effect of the multiple regression model on the training data.

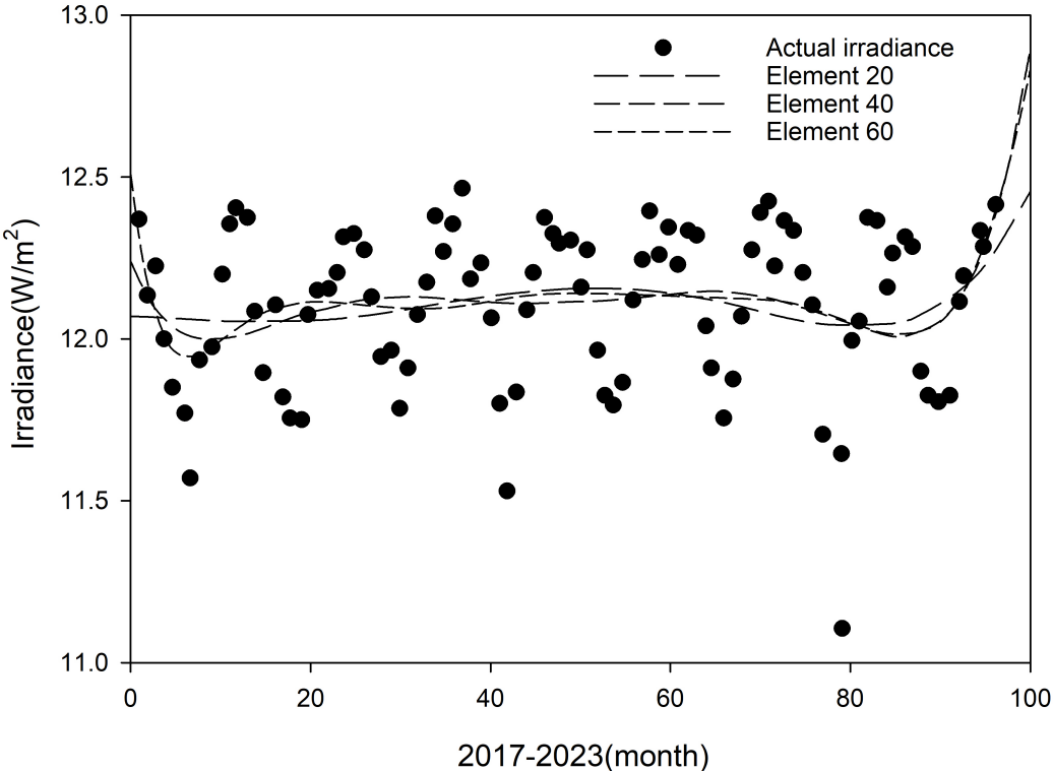


Figure 2. Multiple regression fitting

The irradiance in the figure is processed by log exponential smoothing. From the figure, we can see that if the input of the model contains only irradiance, the actual significance of the model is to fit the change trend of the actual irradiance in the time series, but the fitting effect of the multiple regression model on the training data set is not very ideal. As the number of elements of the multiple regression model increases, It does not improve the final fitting effect, However, the calculation time of the model will increase, and the complexity will increase. In the time series, the actual irradiance will have a large fluctuation phenomenon, relying on this condition alone, can not capture the correct change trend. Therefore, factors affecting the irradiance are considered in the input data. The results of persistent model fitting based on the training data set are shown in Figure 3.

Figure 3 shows the optimal fitting results of the persistent prediction model under different combinations of input data. The errors are shown in Table 1, where (a) the input irradiance plus clear sky index is the optimal input when considering a single meteorological factor, and its RMSE 0.27, the autocorrelation coefficient (R-square) was 0.12; (b) In order to consider the two meteorological factors, the input irradiance, clear sky index and sunshine duration are the optimal inputs, with RMSE 0.26 and R-square 0.21; (c) In order to consider multiple meteorological factors, the input irradiance, clear sky index, sunshine duration and cloud cover ratio were the optimal inputs, with RMSE 0.25 and R-square 0.24; (d) The RMSE is 0.26, and the autocorrelation coefficient (R-square) is 0.15, considering the input in the case of all meteorological factors.

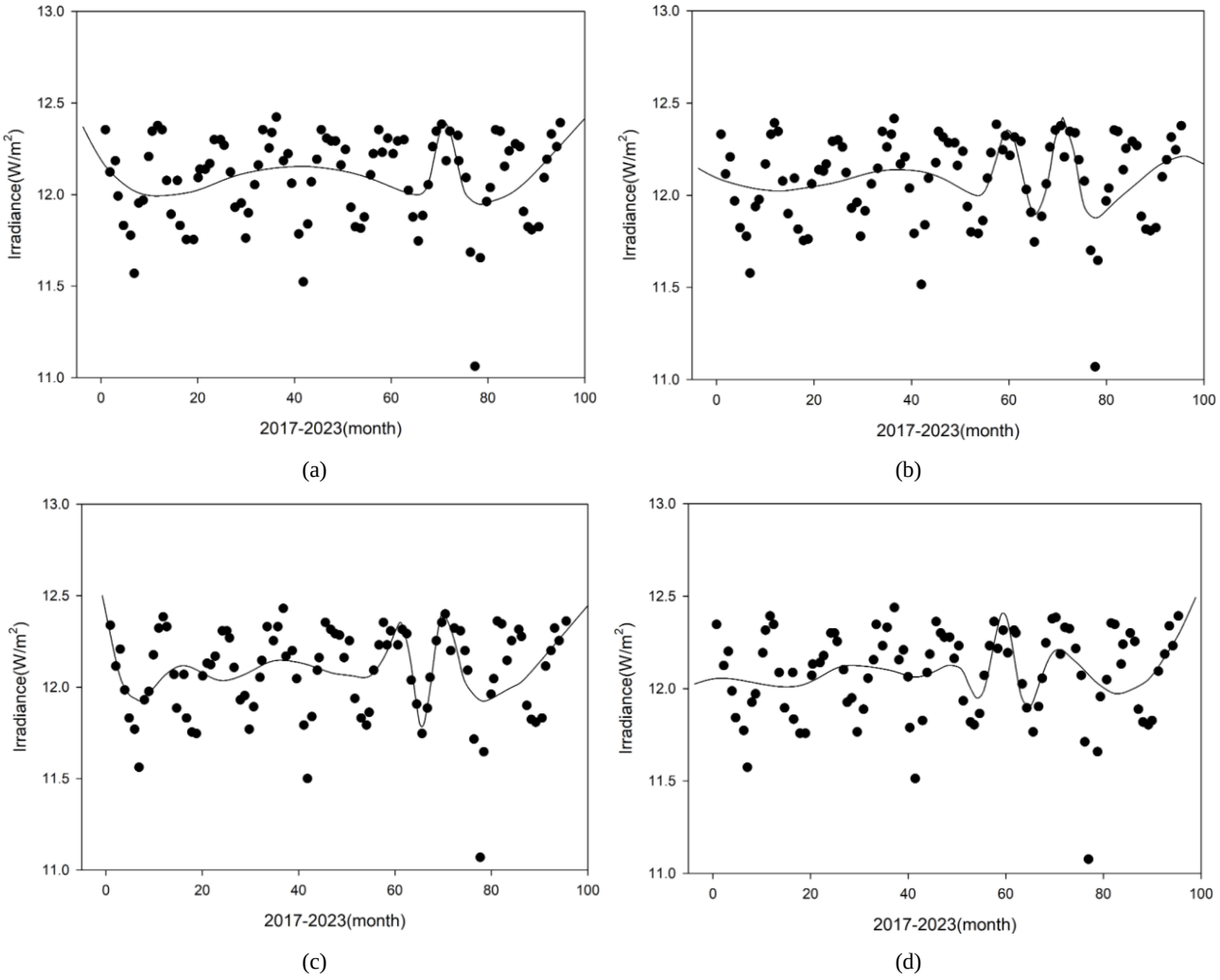
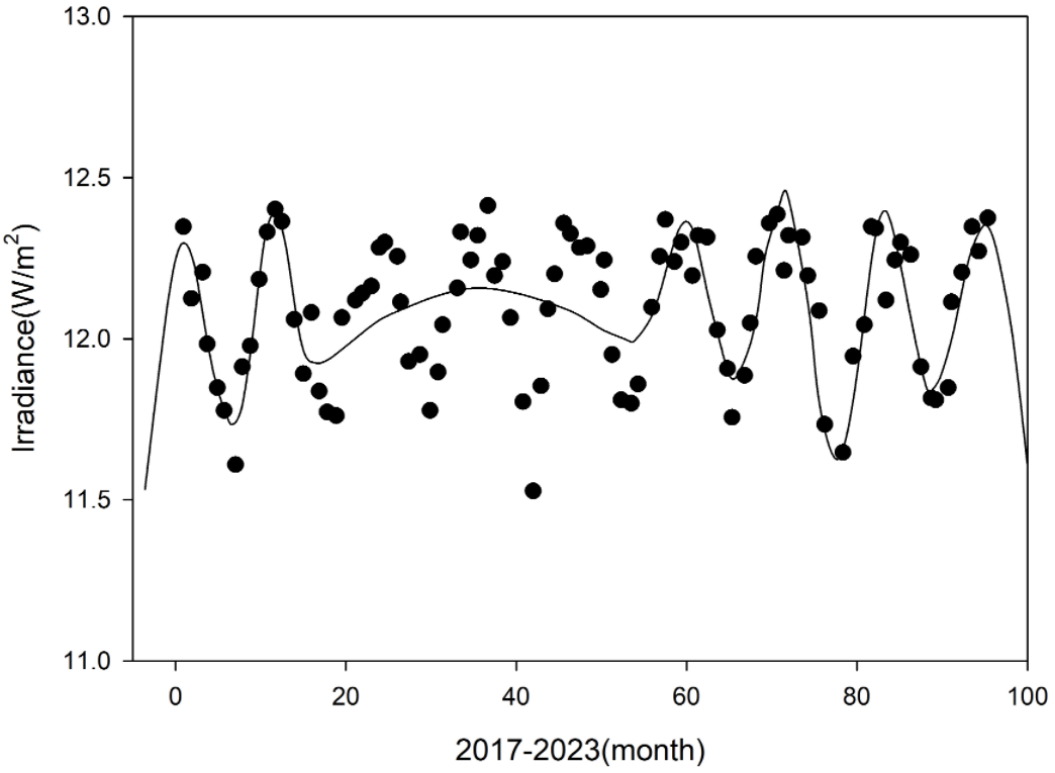


Figure 3. Fits different inputs based on a persistent model

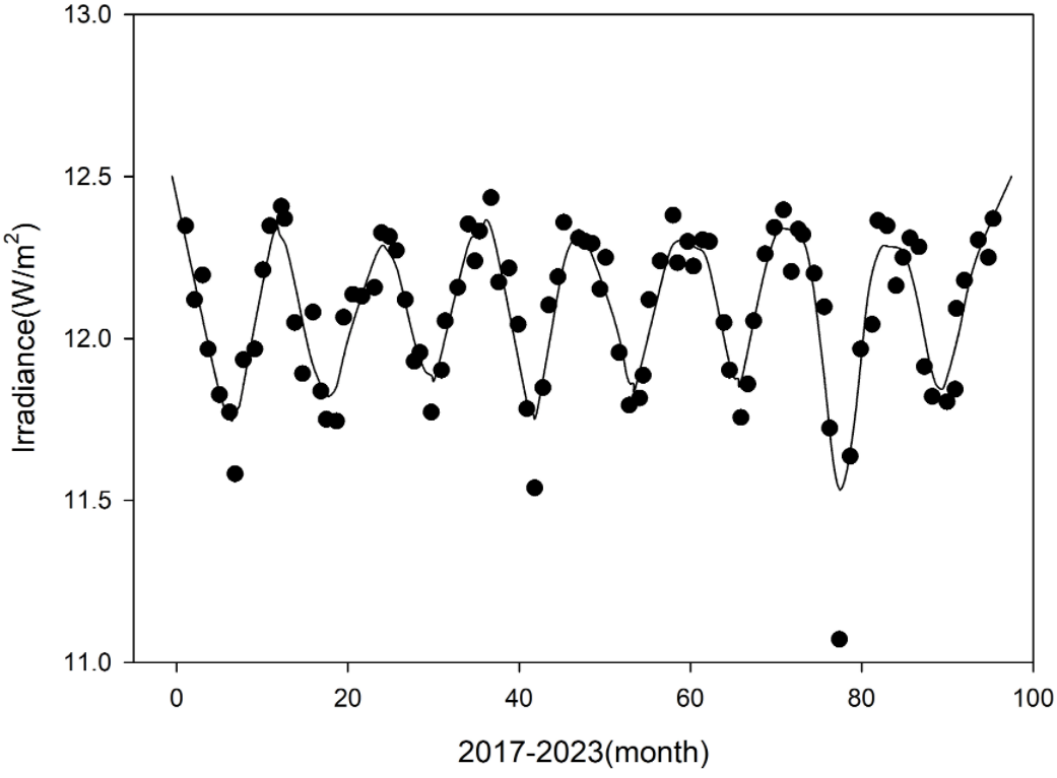
Table 1. Different input errors based on persistent models

Input combination	Irradiance + clear sky index	Irradiance + clear sky index + sunshine duration	Irradiance + clear sky index + sunshine duration + cloud cover ratio	Irradiance + total meteorological factors
RMSE	0.27	0.26	0.25	0.26
R-square	0.12	0.21	0.24	0.15

Due to the fact that the irradiance data is smoothed using the log1p index and its magnitude is compressed between 11 and 13, the RMSE calculated from the model training error cannot clearly reflect the error quality under different input combinations. Therefore, it is necessary to compare the performance of models under different input combinations based on the autocorrelation coefficient (R-square). Generally speaking, the closer the autocorrelation coefficient is to 1, the better the training effect of the model. In terms of the amount of input, because it's done on a monthly basis throughout the forecasting process, and the total amount of data is limited, most of the meteorological factors are processed by means of average. Therefore, during model training, emphasis should be placed on the selection of the feature dimension of the input data. It's not hard to see back in Figure 1, the larger the feature dimension required for model training, the model training effect is inversely proportional to the complexity, which increases and decreases. In addition, increasing the feature dimension of training data with limited sample data will lead to the deterioration of the generalization ability. According to the fitting results of the persistent model, irradiance, clear sky index, In the following comparison experiment based on the support vector machine model, irradiance, clear sky index, sunshine duration and cloud cover ratio are taken as the input of the two models. The fit of these two models on the training set is shown in Figure 4.



(a)



(b)

Figure 4. Fitting comparison between LSTM and support vector machine

It can be seen from the fitting results in Figure 4, the fitting effect of support vector machine and LSTM neural network on the training set has been significantly improved compared with multiple regression model and persistent prediction model. Since irradiance has a strong change law in time series, And the forecast time is long,LSTM neural network can deal with "long period dependence" effectively because of its unique memory module in the network structure. Table 2 shows the error table of fitting between support vector machine and LSTM neural network on training set.

Table 2. Results of fitting error between LSTM neural network and support vector machine

Selection model	Support vector machine	LSTM
RMSE	0.13	0.08
R-square	0.66	0.94

To further analyze the prediction effect of LSTM, the prediction data set is input into each trained model. Figure 5 is the prediction effect diagram of LSTM and each model, in which the predicted value is converted into the true irradiance range through log inverse transformation, and Table 3 is the comparison table of the prediction error between LSTM and each model.

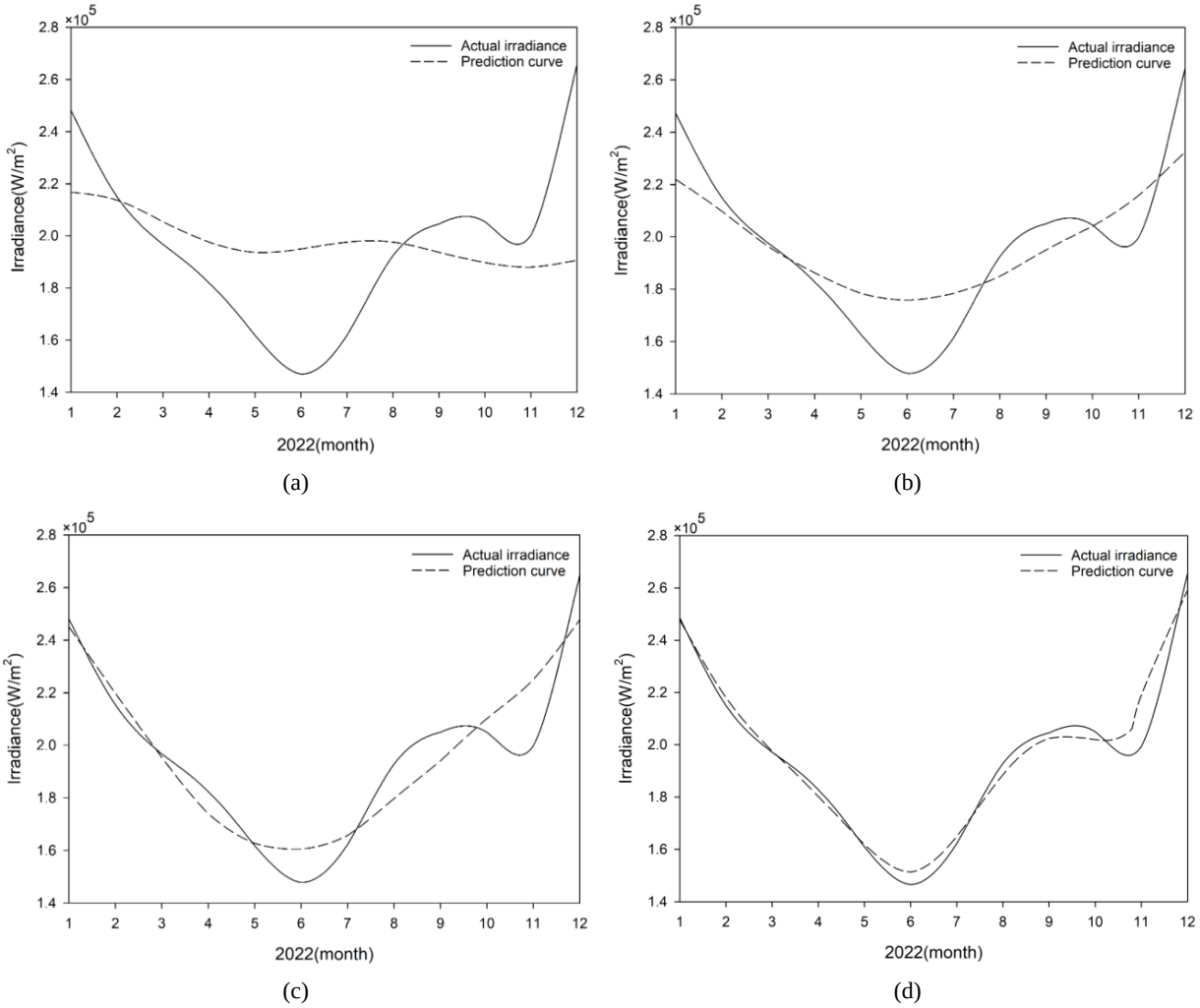


Figure 5. Comparison of prediction results between LSTM and various models

Table 3. Comparison of prediction error results of each model

Model	Multiple regression	Persistent forecasting	Support vector machine	LSTM
MAPE	33.57%	19.99%	12.36%	3.87%
R-square	0.11	0.30	0.73	0.95

It can be seen that, LSTM has better fitting ability and prediction ability. As for the prediction error, since the final irradiance value is in the actual range, the order of magnitude is larger, which is suitable for MAPE as the error measurement standard. For medium - and long-term prediction tasks, the MAPE of the predicted results is within 5%, which is an acceptable range. Therefore, LSTM neural network has better prediction effect in medium and long-term irradiance prediction tasks with limited training sample data, regular change trend of prediction target in time series, and "long time period dependent" condition.

4.2. Medium- and Long-Term Power Consumption Data Prediction of Distributed Photovoltaic Power Stations

Through indirect prediction, that is, using irradiance as a medium, the required power consumption data is predicted. Taking 20#, 35#, 55#, 65# as the object of study, the power consumption data of each power station in 2022 is predicted, and the results are as follows (Figure 6).

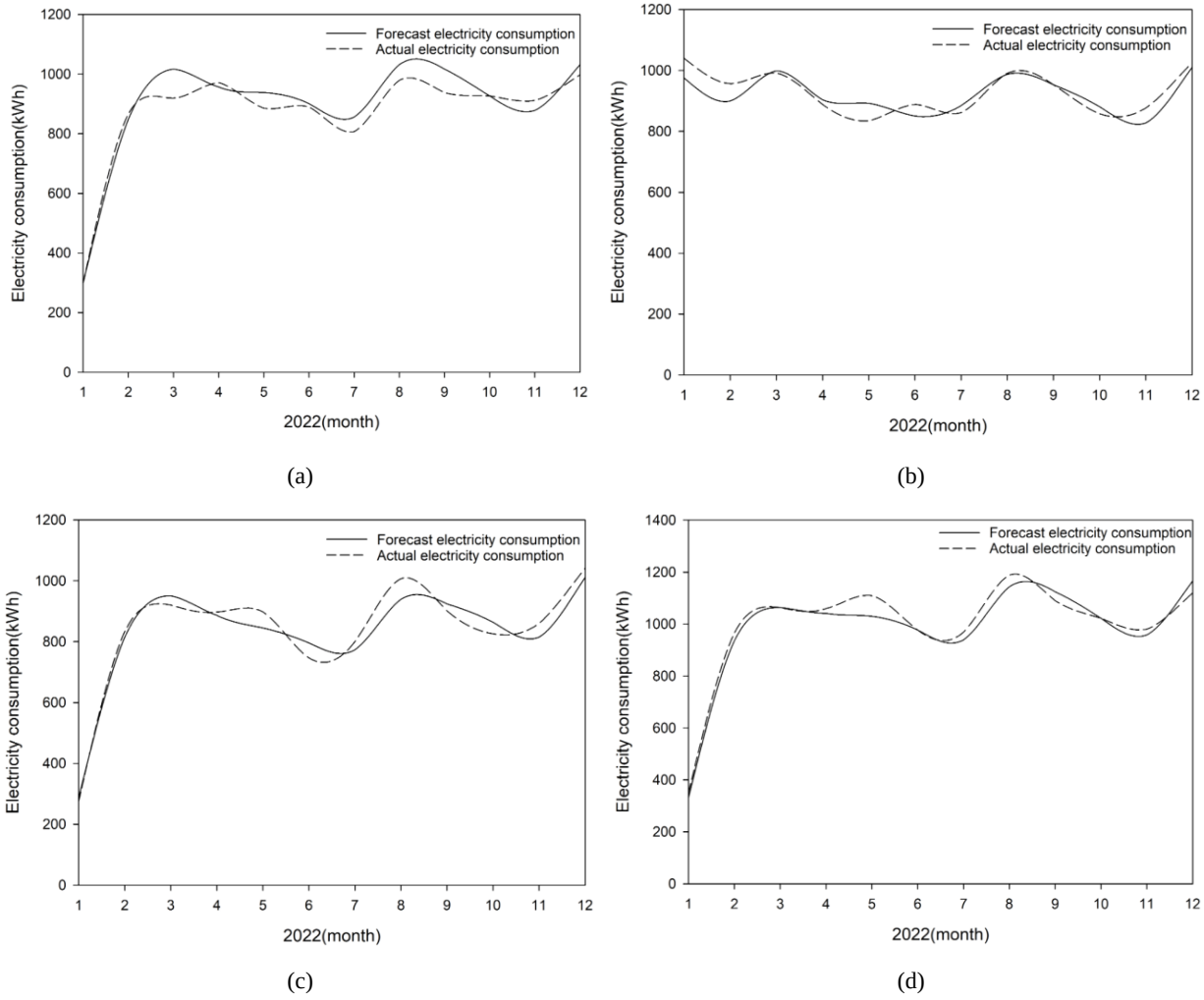


Figure 6. Prediction results of electricity consumption data

In the Figure 6, the MAPE of power station 20# is 3.80%, the MAPE of power station 35# is 3.36%, the MAPE of power station 55# is 3.96%, and the MAPE of power station 62# is 3.09%. The annual MAP errors of the 20, 35, 55, and 65 # power stations are 0.87%, 1.11%, 1.02%, and 1.22%, respectively, by adding up the predicted monthly power generation and comparing them with the actual total power generation in 2022. It can be seen that the MAPE of the monthly predicted value and the actual value of the LSTM-based distributed photovoltaic power generation model fluctuates up and down 3.5%, and the annual predicted value of each power station fluctuates up and down 1.1%, and the prediction effect is good.

5. Conclusion

To sum up, this paper first summarizes the current irradiance prediction model and understands its principle and structure; then, the principle of the LSTM neural network algorithm is introduced in detail, paving the way for the subsequent example analysis and empirical development. Secondly, by comparing the prediction effect of the multiple regression model, persistent prediction model, support vector machine model, and LSTM neural network model, it is found that MAPE is 33.57% and R-square is 0.11 in the prediction of the multi-source regression model. In the persistent prediction model, MAPE is 19.99% and R-square is 0.30. In SVM model prediction, MAPE was 12.36% and R-square was 0.73. In the prediction of LSTM neural networks, MAPE is 3.87% and R-square is 0.95, both of which are superior to other models, confirming the advantages of LSTM neural networks in medium- and long-term irradiance prediction. Finally, in order to further test the prediction effect of the LSTM neural network model, distributed PV power stations 20#, 35#, 55#, and 65# are taken as the research objects to forecast the electricity consumption data of each power station in 2022. Finally, it is found that the MAPE of the monthly forecast is 3.5%, and the annual forecast is 1.1%, which has high robustness. It can be put into practical applications to predict the electricity consumption data of distributed photovoltaic power stations in real time.

6. Declarations

6.1. Author Contributions

Conceptualization, J.Y.; methodology, J.Y.; software, C.X.; validation, J.Y. and J.H.; formal analysis, C.X.; investigation, J.H.; resources, J.Y.; data curation, X.Y.; writing—original draft preparation, J.Y.; writing—review and editing, C.X.; visualization, X.Y.; supervision, J.Y.; project administration, J.H.; funding acquisition, J.Y. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

This work was supported by science and technology project of STATE GRID Shanxi Electric Power Company "Research on the Analysis of Regional Characteristics and State Assessment of Distributed Photovoltaic Power Generation" (52051L230101).

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7. References

[1] Jovijari, F., & Mehrpooya, M. (2024). Development of crude oil desalination unit by using solar flat plate collectors. Applied Thermal Engineering, 239, 122110. doi:10.1016/j.applthermaleng.2023.122110.

[2] Awogbemi, O., & Von Kallon, D. V. (2023). Towards the development of underutilized renewable energy resources in achieving carbon neutrality. Fuel Communications, 100099.doi:10.1016/j.jfueco.2023.100099.

[3] Rafiq, M., Mahr, M. S., Imran, R., Shaban, M., Al-Saeedi, S. I., Hasanin, T. H. A., Salim, M., & Ibrahim, M. A. A. (2023). Towards Development of High-Performance Perovskite Solar Cells Based on Pyrrole Materials for Hole Transport Layer by Using Computational Approach. Journal of Computational Biophysics and Chemistry, 22(8), 1097–1113. doi:10.1142/S2737416523420127.

[4] Charbonnier, F., Morstyn, T., & McCulloch, M. (2024). Home electricity data generator (HEDGE): An open-access tool for the generation of electric vehicle, residential demand, and PV generation profiles. MethodsX, 12, 102618. doi:10.1016/j.mex.2024.102618.

[5] Liang, Y., Li, P., Su, W., Li, W., & Xu, W. (2024). Development of green data center by configuring photovoltaic power generation and compressed air energy storage systems. Energy, 292. doi:10.1016/j.energy.2024.130516.

[6] Qiu, Z., Tian, Y., Luo, Y., Gu, T., & Liu, H. (2024). Wind and Photovoltaic Power Generation Forecasting for Virtual Power Plants Based on the Fusion of Improved K-Means Cluster Analysis and Deep Learning. Sustainability, 16(23), 10740. doi:10.3390/su162310740.

[7] Yu, X. P., Li, P., Zhang, Y., Li, H., Yang, M., Zheng, Y., & Xue, M. (2022, November). Research on New Energy Generation Market Transaction Based on Sales Risk Control Strategy. 2022 IEEE 6th Conference on Energy Internet and Energy System Integration (EI2), 2008-2014. doi:10.1109/EI256261.2022.10117401.

[8] Lu, P., Ye, L., Zhong, W., Qu, Y., Zhai, B., Tang, Y., & Zhao, Y. (2020). A novel spatio-temporal wind power forecasting framework based on multi-output support vector machine and optimization strategy. Journal of Cleaner Production, 254, 119993. doi:10.1016/j.jclepro.2020.119993.

[9] Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de-Pison, F. J., & Antonanzas-Torres, F. (2016). Review of photovoltaic power forecasting. Solar energy, 136, 78-111. doi:10.1016/j.solener.2016.06.069.

[10] Zhang, W., Li, Q., & He, Q. (2022). Application of machine learning methods in photovoltaic output power prediction: A review. Journal of Renewable and Sustainable Energy, 14(2), 022701. doi:10.1063/5.0082629.

[11] Al-Dahidi, S., Madhiarasan, M., Al-Ghussain, L., Abubaker, A. M., Ahmad, A. D., Alrbai, M., ... & Zio, E. (2024). Forecasting solar photovoltaic power production: a comprehensive review and innovative data-driven modeling framework. Energies, 17(16), 4145. doi:10.3390/en17164145.

[12] Herraiz, Á. H., Marugán, A. P., & Márquez, F. P. G. (2020). Photovoltaic plant condition monitoring using thermal images analysis by convolutional neural network-based structure. Renewable Energy, 153, 334-348. doi:10.1016/j.renene.2020.01.148

[13] Bo, G., Chao, M., Chongbiao, Z., Weijie, Q., Chao, F., & Chao, Z. (2023). Output Forecast of Distributed Photovoltaic Power Generation based on Spatial-Temporal Graph Neural Network. Journal of Electric Power Systems and Automation, 35, 125–133.

[14] Wang, S., Yan, S., Li, H., Zhang, T., Jiang, W., Yang, B., ... & Wang, J. (2024). Short-term prediction of photovoltaic power based on quadratic decomposition and residual correction. Electric Power Systems Research, 236, 110968. doi:10.1016/j.epsr.2024.110968.

[15] Hou, L., Ding, H., Liu, Y., & Wang, S. (2022). Evaluation and suggestion on the subsidy policies for rural clean heating in winter in the Beijing-Tianjin-Hebei region. Energy and Buildings, 274, 112456. doi:10.1016/j.enbuild.2022.112456.

[16] Wang, Y., Chen, L., & Shi, X. (2023). Prediction of scrap volume and recyclable resource potential of distributed photovoltaic power generation equipment in the Beijing-Tianjin-Hebei region. Resources Science, 45(10), 2076–2088. doi:10.18402/resci.2023.10.12.

[17] Molina, M. G., & Espejo, E. J. (2014). Modeling and simulation of grid-connected photovoltaic energy conversion systems. International Journal of Hydrogen Energy, 39(16), 8702-8707. doi:10.1016/j.ijhydene.2013.12.048.

[18] Zhang, C., Yan, X., & Nie, J. (2023). Economic analysis of whole-county PV projects in China considering environmental benefits. Sustainable Production and Consumption, 40, 516-531. doi:10.1016/j.spc.2023.07.020.

[19] Li, J., Wang, P., Dong, H., & Shen, J. (2022). Multi/many-objective evolutionary algorithm assisted by radial basis function models for expensive optimization. Applied Soft Computing, 122, 108798. doi:10.1016/j.asoc.2022.108798.

[20] de Souza, L. P., Sanches-Neto, F. O., Junior, G. M. Y., Ramos, B., Lastre-Acosta, A. M., Carvalho-Silva, V. H., & Teixeira, A. C. S. C. (2022). Photochemical environmental persistence of venlafaxine in an urban water reservoir: A combined experimental and computational investigation. Process Safety and Environmental Protection, 166, 478-490. doi:10.1016/j.psep.2022.08.049.