electronics

MDPI

Article

Feature Extraction Approach for Distributed Wind Power Generation Based on Power System Flexibility Planning Analysis

Sile Hu [1,2], Jiaqiang Yang [1,*], Yuan Wang [3], Chao Chen [1], Jianan Nan [2], Yucan Zhao [1] and Yue Bi [1]

[1] College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China; 12110087@zju.edu.cn (S.H.); 22260119@zju.edu.cn (C.C.); zhaoyucan@zju.edu.cn (Y.Z.); 22360166@zju.edu.cn (Y.B.)
[2] Inner Mongolia Power (Group) Co., Ltd., Hohhot 010020, China; nanjianan@impc.com.cn
[3] Inner Mongolia Electric Power Economic and Technological Research Institute, Hohhot 010090, China; wangyuan9@impc.com.cn
* Correspondence: yangjiaq@zju.edu.cn; Tel.: +86-0571-87951243

Abstract: This study addresses the integral role of typical wind power generation curves in the analysis of power system flexibility planning. A novel method is introduced for extracting these curves, integrating an enhanced *K*-means clustering algorithm with advanced optimization techniques. The process commences with thorough data cleaning, filtering, and smoothing. Subsequently, the refined *K*-means algorithm, augmented by the Pearson correlation coefficient and a greedy algorithm, clusters the wind power curves. The optimal number of clusters is ascertained through the silhouette coefficient. The final stage employs particle swarm and whale optimization algorithms for the extraction of quintessential wind power output curves, essential for flexibility planning in power systems. This methodology is validated through a case study involving wind power output data from a new energy-rich provincial power grid in North China, spanning from 1 January 2019, to 31 December 2022. The resultant curves proficiently mirror wind power fluctuations, thereby laying a foundational framework for power system flexibility planning analysis.

Keywords: power system flexibility; Extended Kalman Filter (EKF) typical wind power curve; clustering; Pearson correlation coefficient; optimization algorithm

check for updates

Citation: Hu, S.; Yang, J.; Wang, Y.; Chen, C.; Nan, J.; Zhao, Y.; Bi, Y. Feature Extraction Approach for Distributed Wind Power Generation Based on Power System Flexibility Planning Analysis. *Electronics* **2024**, *13*, 966. https://doi.org/10.3390/electronics13050966

Academic Editors: Ahmed F. Zobaa and Davide Astolfi

Received: 7 January 2024
Revised: 14 February 2024
Accepted: 29 February 2024
Published: 2 March 2024

1. Introduction

As nations advance toward 'carbon peak and carbon neutrality' objectives, there's a notable shift towards novel power systems, heavily incorporating renewable energy sources like wind and solar. Despite the rapid progress, the inherent unpredictability and intermittency of wind power introduce considerable challenges to the power grid's stability and operational planning. Addressing these issues, system flexibility planning becomes pivotal, ensuring the power grid's resilience, reliability, and economic operation, especially as the share of these volatile energy sources grows. This planning is integral to managing the complexities introduced by fluctuating energy inputs like wind power. [1].

In the context of extensive wind power integration, understanding its influence on system flexibility is paramount. The initial step entails identifying typical daily wind power generation curves through flexibility planning analysis. This approach shifts the focus from the stochastic nature of wind power to a deterministic scenario analysis, catering to the demands of large-scale, time-series data computations and analyses. This transformation is vital for accurate system flexibility assessments in a landscape dominated by renewable energy [2]. In the realm of power system operation, typical curve extraction methods are broadly categorized into traditional and emerging techniques. Traditional methods, like statistical analysis, distill patterns from wind power data to derive typical curves, while manual classification groups curves, selecting exemplars from each. Conversely, emerging methods, like cluster analysis, employ advanced algorithms (e.g., *K*-means, fuzzy C-means) to aggregate and represent large datasets. Another innovative approach, model fitting, constructs mathematical models to mirror wind turbine output, offering a data-driven pathway to identifying typical curves. The authors of [3,4] delved into the characteristic index system of wind power and photovoltaic output. The study focused on categorizing and pinpointing typical curves, a crucial step for understanding and optimizing energy generation patterns in renewable energy systems. The authors of [5] conducted an insightful analysis of power system load curves, pinpointing the day with the highest annual peak-to-valley difference. This day was utilized as a benchmark for electricity balance calculations in wind power systems, emphasizing the significance of understanding daily fluctuations for effective energy management. In [6], the authors advanced traditional approaches by revamping and refining the daily load curve for enhanced clustering accuracy. Recent progress in machine learning has spurred the development and application of sophisticated clustering techniques, significantly enriching the analysis of power system generation curves with improved accuracy and insight. The authors of [7] introduced a sophisticated clustering approach by merging the enhanced fuzzy C-means algorithm (PFCM) with fuzzy linear discriminant analysis (FLDA). This integration aims to optimize the selection of typical load curves, harnessing the strengths of both methods for more accurate and insightful power system analysis.

Furthermore, in [8], the authors introduced a novel approach by utilizing hierarchical clustering algorithms to effectively deduce typical scenarios from regional wind power data. This method aims to comprehensively capture and represent the intrinsic patterns and variabilities in wind power generation across different regions. Reference [9] obtained the optimal clustering scheme by combining the lion optimization algorithm with the *K*-means algorithm, while reference [10] used the grey wolf algorithm combined with the fuzzy C-means clustering algorithm (FCM) to improve the daily load curve clustering effect. The authors of [11] determined the optimal number of clusters based on the silhouette coefficient. Reference [12] introduced the Copula function into the production of typical curves for wind and solar power output scenarios [13]. In [14], the authors used an optimized spectral clustering algorithm for typical scenario generation and analysis.

The proliferation and diversification of new energy sources have led to an expanded distribution of energy stations and a wealth of operational data. This growing data repository significantly enhances the robustness of various clustering methodologies reliant on real operational data, paving the way for more informed and data-driven decisions in the energy sector [15]. The research emphasizes refining clustering algorithms for global analysis of wind and solar power output curves. Notably, it introduces a method combining an improved *K*-means algorithm with intelligent optimization algorithms for extracting typical wind power curves, crucial for power system flexibility planning. The method involves data cleansing, noise filtering, smoothing, and using the Pearson correlation coefficient with a greedy algorithm to enhance the *K*-means clustering, addressing traditional challenges like optimal *K* value selection and clustering center determination [11]. Then, particle swarm algorithms [16] and whale algorithms [17] are used to extract typical wind power curves and perform cross-validation, ultimately generating typical daily wind power output curves for power system flexibility planning analysis.

2. General Framework for Extracting Typical Daily Wind Power Generation Curves

The framework for extracting typical daily wind power generation curves based on power system flexibility planning analysis is illustrated in Figure 1 and comprises four steps:

1. Eliminate anomalies and missing values from historical daily wind power output data, standardize the output data (output data/daily installed capacity), then use the Extended Kalman Filter (EKF) for data filtering and noise reduction, followed by curve smoothing functions [18]. Afterwards, calculate the daily peak-to-valley difference rate to obtain the wind power daily output dataset for subsequent research.

2. Select the daily generation curves from the processed wind power daily output dataset that fall within the top 10% in terms of peak-to-valley difference rate and conduct statistical analysis on the months when the maximum peak-to-valley differences occur.
3. Perform *K*-means clustering on the selected wind power data to preliminarily analyze data patterns and the clustering *K*-value. The Pearson correlation coefficient plays a pivotal role in the optimization of the *K*-means clustering algorithm, particularly by quantifying linear relationships, enhancing *K*-means clustering, and optimizing cluster representation. Calculate the Pearson correlation coefficient of the selected daily wind power output data, and use the Greedy Algorithm to find the output curve with the lowest correlation as the clustering center [19]. Then, use *K*-means for clustering and repeat several times to find the clustering result corresponding to the maximum silhouette coefficient [12].
4. Extract curves from the clustering result dataset using both particle swarm optimization and whale algorithms. Cross-validate the results obtained from these two methods to finally derive the typical daily wind power generation curves for power system flexibility planning analysis.

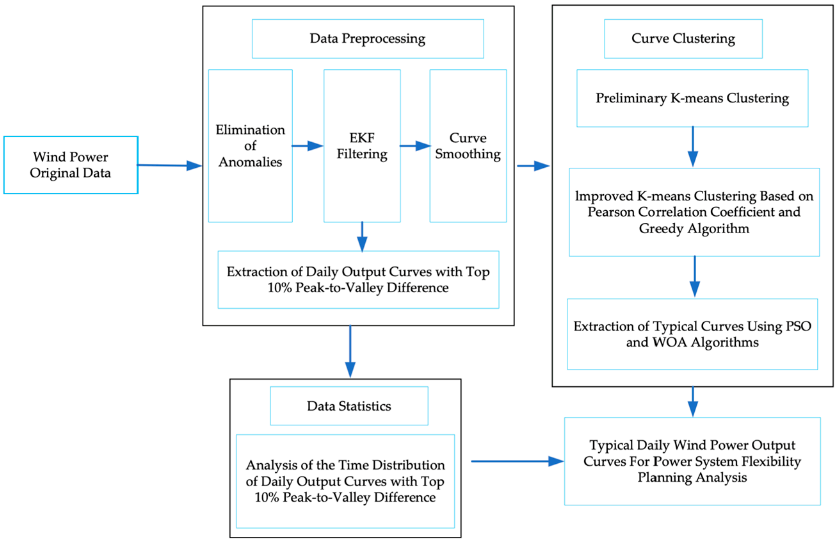
Wind Power Original Data
Data Preprocessing
Elimination of Anomalies
EKF Filtering
Curve Smoothing
Extraction of Daily Output Curves with Top 10% Peak-to-Valley Difference
Data Statistics
Analysis of the Time Distribution of Daily Output Curves with Top 10% Peak-to-Valley Difference
Curve Clustering
Preliminary K-means Clustering
Improved K-means Clustering Based on Pearson Correlation Coefficient and Greedy Algorithm
Extraction of Typical Curves Using PSO and WOA Algorithms
Typical Daily Wind Power Output Curves For Power System Flexibility Planning Analysis

Figure 1. General framework for extracting typical daily wind power generation curves.

3. Data Preprocessing

3.1. Cleaning of Missing and Anomalous Data

Data cleaning is an integral step in our data analysis process, designed to enhance the reliability and accuracy of the dataset by addressing incorrect, incomplete, or inaccurate data entries. This section is of paramount importance as it underpins the credibility of our subsequent analysis and modeling [20].

The process of cleaning the data involved several stages. Initially, we employed an automated script to identify and flag any missing values within the dataset. The criteria for identifying missing data were based on the expected data transmission intervals. Any points that did not adhere to these intervals were considered missing.

Following the identification of missing data, we applied an imputation technique to address these gaps. Specifically, we used a linear interpolation method for time series data where appropriate, which assumes that the change between two data points is linear and can be estimated. This method was chosen for its simplicity and effectiveness in dealing with small gaps in time series data.

In addition to missing data, we also scrutinized the dataset for anomalies that could indicate incorrect or inaccurate data. For this, we implemented a Z-score analysis to detect outliers. Data points with a Z-score greater than 3 were considered extreme and thus were examined manually to determine whether they represented true values or anomalies due to recording errors or data transmission issues.

We further refined the dataset by examining the consistency of the daily output curves. Any curve that showed continuous abnormalities, deviating significantly from the established pattern without justifiable cause, was removed. This decision was based on the standard deviation of the curve's gradient, with a threshold set at two standard deviations from the mean gradient of the dataset.

The enhancements to our methodology not only improve the quality of our data but also provide a transparent and rigorous foundation for our analysis, thus addressing the potential issues inherent in the secondary data collection process from power systems.

3.2. Data Denoising and Smoothing

To enhance the accuracy of the data, the Extended Kalman Filter (EKF) algorithm is used for data filtering and noise reduction, followed by smoothing the filtered data using a smooth function resulting in data for subsequent research [10].

The Extended Kalman Filter (EKF) is a nonlinear state estimation algorithm based on the Kalman Filter, suitable for systems with nonlinear models [18]. EKF linearizes nonlinear problems by Taylor expansion of the nonlinear functions, and then employs the Kalman Filter methodology for state estimation and error correction.

The specific steps are as follows:

1. State equation:

$$x(k) = f(x(k-1), u(k-1)) + \omega(k-1) \tag{1}$$

in which $\omega(k-1)$ represents process noise.

2. Observation equation:

$$y(k) = h(x(k)) + v(k) \tag{2}$$

in which $v(k)$ represents observation noise.

3. Observation steps:

State vector prediction:

$$x_{pred} = [x_1 + x_2, x_2]^T \tag{3}$$

Transition matrix:

$$F = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \tag{4}$$

Covariance matrix predicted value:

$$P_{pred} = FP_0F^T + Q \tag{5}$$

4. Update steps:

Observed value and predicted value:

$$y_{yred} = x_{pred}(1) \tag{6}$$

Observation Matrix:

$$H = [1, 0] \tag{7}$$

Kalman Gain:

$$K = P_{pred}H^T\left(HP_{pred}H^T + R\right)^{-1} \tag{8}$$

Estimated state vector:

$$x_{est} = x_{pred} + K\left(y(k) - y_{yred}\right) \tag{9}$$

Estimated covariance matrix:

$$P_{est} = (I - KH)P_{pred} \tag{10}$$

In this context, x_1 and x_2 represent the two components of the state vector, specifically the wind power output and its rate of change, respectively. Q and R represent the covariance matrices of process noise and observation noise, respectively. In the program, the initial values of the covariance matrices are as follows:

$$P_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{11}$$

After filtering, the data is further smoothed using the moving average method. The moving average method is a commonly used smoothing technique that can smooth a one-dimensional vector, removing noise and jitter. Its basic concept is as follows: for a vector of length N, a window of length m is selected. The data within this window is averaged to obtain a new data point. Then, the window is shifted one unit to the right, and the above operation is repeated until all data points have been processed.

As shown in equation:

$$\mathrm{y}_i = \frac{1}{\mathrm{m}} \sum_{j=1}^{m} x_{i-j+\lfloor \frac{m}{2} \rfloor} \tag{12}$$

where x_i represents the i data point of the original data vector, y_i represents the i data point of the smoothed data vector, m represents the window size, and $\lfloor \cdot \rfloor$ represents the floor function.

3.3. Data Normalization

Considering that the wind power output curve is significantly influenced by the installed capacity, to find a general rule, this paper adopts the method of normalizing the data by dividing the daily wind power output by the daily wind power installation capacity.

$$x' = \frac{x}{p} \tag{13}$$

In the equation, x' is the normalized wind power output value, and p is the daily installed wind power capacity.

4. Clustering of Wind Power Output Curves

The influence of environmental factors on wind power output introduces significant variability and distinct typologies in output curves. K-means clustering, an eminent unsupervised learning algorithm, adeptly categorizes these curves by their shape and size using Euclidean distance. This algorithm is praised for its simplicity, straightforward implementation, and computational efficiency, which enables it to process large datasets with commendable scalability. Its ability to cluster substantial volumes of data efficiently makes it an indispensable tool for analyzing and understanding the complex dynamics of wind power output curves, providing valuable insights into their classification and underlying patterns. However, K-means requires pre-specifying the number of clusters K, which depends on experience or other algorithms, lacking automation. Additionally, it is sensitive to the choice of initial cluster centers. Since cluster centers are randomly selected, this can lead to convergence to local optima instead of global optima and inconsistency in clustering results [21].

To address these issues, this paper uses the silhouette coefficient method to preliminarily find the K value with a larger silhouette coefficient. It calculates the Pearson correlation coefficient of the selected data, and uses the greedy algorithm to find the K data points with the lowest correlation under the K value to serve as cluster centers. Afterwards, K-means clustering is performed.

The silhouette coefficient is a clustering evaluation metric used to assess the reasonableness and quality of clustering results [11]. It is based on the comparison of intra-cluster similarity and inter-cluster dissimilarity, with a value range of $[-1, 1]$. A higher value indicates better clustering effect.

Specifically, for each sample, the silhouette coefficient is calculated as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{14}$$

The provided text is a description of the silhouette coefficient and the Pearson correlation coefficient, key concepts used in statistical analysis and clustering.

The silhouette coefficient involves two distances: $a(i)$, which represents the average distance of sample i to the other samples in the same cluster (intra-cluster similarity), and $b(i)$, which represents the average distance of sample i to all the samples in the nearest cluster that i is not a part of (inter-cluster dissimilarity). For the silhouette coefficient of the entire dataset, the average value of the silhouette coefficients for all samples can be taken.

The Pearson Correlation Coefficient is a statistical measure used to assess the strength of a linear relationship between two variables. Its value ranges between -1 and 1, where 0 indicates no linear relationship, and values closer to 1 or -1 indicate a stronger linear relationship between the two variables [22]. The formula for calculating the Pearson correlation coefficient is as follows:

$$r_{x_i,x_j} = \frac{Cov\left(x_i, x_j\right)}{S(x_i)\cdot S\left(x_j\right)} \tag{15}$$

where $Cov\left(x_i, x_j\right)$ represents the covariance between x_i and x_j, and $S(x_i)$ and $S\left(x_j\right)$ are the standard deviations of x_i and x_j, respectively.

The greedy algorithm is a common heuristic algorithm that achieves a global optimum by selecting a local optimum at each step. Specifically, the greedy algorithm chooses the best solution at each step in decision making until a certain goal is reached or no further optimization can be made. After obtaining the correlation coefficient for each curve, the greedy algorithm is used to calculate the selection of k points in the dataset such that their sum is minimized.

The specific steps are as follows:

1. Choose a starting point r_i and add it to the set of selected points;
2. Calculate the distance d_{ij} from all unselected points to the selected points, i.e., the distance between the ith point and the jth point, and find the point r_j with the minimum distance to the selected points, then add it to the set of selected points;
3. Repeat step 2 until the number of selected points reaches the preset value k.

In this problem, the weight of each point is its value r, so the sum of the selected points is calculated as:

$$\sum_{i=1}^{k} r_i \tag{16}$$

In the program, the sum of the points can be calculated by iterating over the selected points, as follows:

$$\sum_{i=1}^{k} r_i = \sum_{i=1}^{k} r_i + \sum_{i<j,i,j\in sselected\ points} d_{ij} \tag{17}$$

where d_{ij} represents the distance between the ith point and the jth point, and r_i represents the value of the ith point. Finally, the value of the above formula is the minimum sum of the selected k points.

Using the above method to select K cluster centers, clustering is completed using the K-means algorithm.

5. Extraction of Typical Wind Power Curves

Upon clustering the wind power output curves, we categorize them into K distinct daily output scenarios. This necessitates the derivation of a representative curve for each category. While the prevailing literature commonly employs averaging to determine representative curves, this method is prone to significant inaccuracies due to the impact of outliers within clusters. To address this issue, our study employs the correlation coefficient method. This method selects the curve that demonstrates the highest cumulative correlation with all other curves in its category as the representative curve. Initially, we compute the Pearson correlation coefficients for the scenario curves within each category. Subsequently, we utilize both the particle swarm optimization (PSO) algorithm and the whale optimization algorithm (WOA) to identify the curve that maximizes the sum of correlations with other curves. The PSO algorithm is preferred for its straightforward implementation, minimal parameter requirements, and efficacy in addressing nonlinear problems. However, its vulnerability to local optima is a notable drawback. In contrast, the WOA, characterized by its unique escape mechanisms, versatility, and balanced approach to exploration and exploitation, may necessitate more iterations but adeptly circumvents the issue of local optima. By integrating these two algorithms and conducting mutual verification during the process of extracting typical curves, we can enhance the accuracy of the results. The curve identified through this rigorous process is then designated as the most representative curve for its category.

5.1. Particle Swarm Optimization (PSO)

The particle swarm optimization (PSO) algorithm is inspired by social behavior patterns of organisms such as birds and fish. It is a stochastic optimization technique that guides a population of particles through the search space by updating generations based on the individual and collective experience of the swarm. Each particle updates its trajectory towards its own best-known position and towards the best-known global position in the search space [16].

Specifically, the basic process of the particle swarm algorithm is as follows:

1. Initialize the population:

Randomly generate a certain number of particles, which are individuals, each with an initial position and velocity.

2. Calculate the fitness function:

Evaluate the fitness value for each particle based on its position in the fitness function.

3. Search for the optimal solution:

Identify the particle with the highest fitness value in the current population, which represents the current optimal solution.

4. Update position and velocity:

Based on the information about the current optimal solution, and the positions and velocities of other particles, update the position and velocity of each particle.

5. Repeat steps 2–4 until the stopping criteria are met.

The position and velocity are updated according to the following formulas:

$$\begin{gathered} v_{id}(t+1) = \omega \cdot v_{id}(t) + c_1 \cdot rand()(p_{id} - x_{id}(t)) + c_2 \cdot rand() \cdot (g_d - x_{id}(t)) \\ x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \end{gathered} \tag{18}$$

The first equation updates the velocity based on the particle's previous velocity, its best-known position, and the global best-known position. The second equation updates the particle's position based on its new velocity. In this context, $v_{id}(t)$ represents the velocity of the ith particle in the dth dimension at iteration t + 1, $x_{id}(t)$ represents the position of the ith particle in the dth dimension at iteration t + 1, p_{id} represents the best-known position of particle i in the dth dimension, g_d represents the best-known position among all particles in the dth dimension, ω represents the inertia weight, balancing exploration and exploitation, c_1 and c_2, respectively, represent cognitive and social factors, and $rand()$ represents a random number function.

5.2. Whale Optimization Algorithm (WOA)

The whale algorithm is a heuristic optimization algorithm based on the group behavior of whales in nature, simulating the behavior of whale pods when foraging for food. This algorithm iteratively updates the position and velocity of each whale in the search for an optimal solution [23].

Specifically, the basic process of the whale algorithm is as follows:

1. Initialize the population:

   Randomly generate a certain number of whales, which are individuals, each with an initial position and velocity.

2. Calculate the fitness function:

   Evaluate the fitness value for each whale based on its position in the fitness function.

3. Search for the optimal solution:

   Identify the whale with the highest fitness value in the current population, which represents the current optimal solution.

4. Update position and velocity:

   Based on the information about the current optimal solution, and the positions and velocities of other whales, update the position and velocity of each whale.

5. Repeat steps 2–4 until the stopping criteria are met.

The position and velocity are updated according to the following formulas:

$$\begin{gathered} v_{id}(t+1) = v_{id}(t) + a \cdot A \cdot (p_{id} - x_{id}(t)) + a \cdot C \cdot (g_d - x_{id}(t)) \\ x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \end{gathered} \tag{19}$$

In this context, $v_{id}(t)$ represents the velocity of the ith whale in the dth dimension, x_{id} represents the position of the ith whale in the dth dimension, p_{id} represents the position of the current best solution in the dth dimension, g_d represents the average position of all whales in the dth dimension, a represents the learning factor, and A and C, respectively, represent the cognitive and social factors.

6. Case Study Analysis

The computational analysis conducted in this study utilized MATLAB 2022b on a Windows 11-based computer system equipped with an i7 processor and 16GB of RAM. The dataset comprises wind power generation data from an expansive new energy grid in North China, covering the period from 1 January 2019, to 31 December 2022. This region, which includes Northeast China, North China, and Northwest China, harbors the nation's richest onshore wind resources. With its installed wind power capacity surpassing 25 GW, the area presents a valuable opportunity for investigating the characteristics of wind power production. Statistical analysis of wind power variability in the region revealed that fluctuations occurring on a 15 min timescale across the grid have attained a level necessitating attention. In contrast, fluctuations shorter than 15 min exert a minimal impact on grid operations. Furthermore, in the interest of computational efficiency, data analysis was conducted using a 15 min timescale.

6.1. Data Preprocessing and Statistics

Anomalies and missing values are removed from the historical daily wind power output data, and the output data is normalized (output data/daily installed capacity).

Subsequently, the data is filtered for noise reduction using the Extended Kalman Filter (EKF), followed by smoothing with the smooth function, and the daily peak-to-valley difference rate is calculated. The daily output curves with a peak-to-valley difference rate in the top 10% (about 150 curves) are extracted for further research. The comparison of the data before and after preprocessing is shown in Figure 2. The preprocessed curves are smoother and retain the characteristics of the curves.

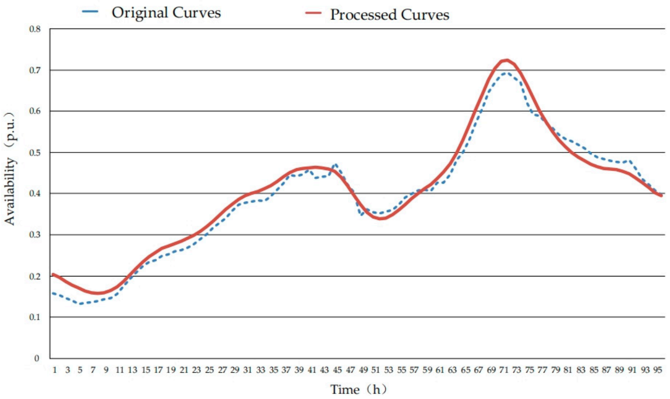

Figure 2. Comparison of wind power output data before and after processing.

Statistical analysis of the processed wind power data on a monthly basis reveals that the days with the top 10% peak-to-valley differences in wind power output are distributed throughout the year, as shown in Figure 3. The occurrence probability is higher in spring, autumn, and winter, corresponding to January–May and September–December, while the frequency is lower in summer, corresponding to June–August. The seasonal distribution probability of days with large peak-to-valley differences is significant, but even during the low wind season in summer, there are days with large wind power daily output peak-to-valley differences. There is no apparent regularity in the distribution of various types of curves, so it is necessary in power system flexibility planning analysis to comprehensively consider various curve shapes and timings of the typical wind power output curves with large peak-to-valley differences. Furthermore, with the large-scale integration of wind power into the grid, the existing wind power data can represent the wind resources and wind power distribution in all regions of the studied provincial power grid. After standardizing the output data, it can provide a reliable research foundation for future power system flexibility planning analysis.

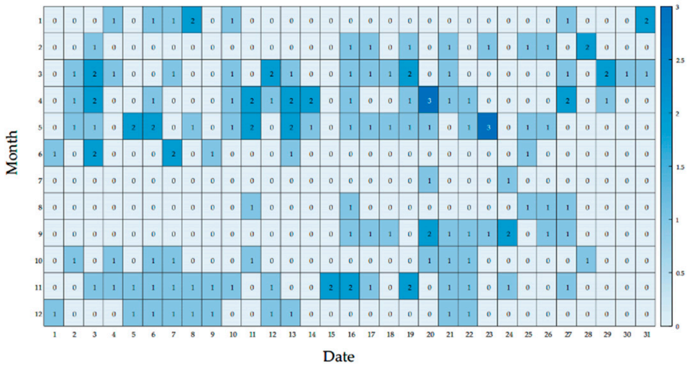

Figure 3. Monthly distribution of wind power daily output curve for the top 10% of peak valley difference.

6.2. Clustering of Wind Power Output Curves

6.2.1. *K*-Means Clustering

Firstly, direct *K*-means clustering is applied to the selected daily wind power output data to find the range of cluster numbers *K* with the maximum silhouette coefficient *SC*.

Taking K values from 2 to 10, clustering is performed 10 times for each K value, and the clustering statistics are shown in Table 1.

Table 1. The results of using K-means algorithm for 10 rounds of clustering.

***K*-Value**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
SC1	0.439	0.421	0.452	0.411	0.435	0.375	0.388	0.38
SC2	0.451	0.473	0.445	0.41	0.432	0.405	0.373	0.383
SC3	0.452	0.443	0.443	0.397	0.433	0.378	0.43	0.382
SC4	0.389	0.425	0.445	0.42	0.427	0.426	0.355	0.411
SC5	0.451	0.398	0.443	0.432	0.414	0.361	0.385	0.39
SC6	0.448	0.413	0.439	0.418	0.432	0.403	0.378	0.372
SC7	0.379	0.42	0.407	0.451	0.434	0.362	0.394	0.377
SC8	0.439	0.409	0.446	0.42	0.442	0.403	0.358	0.388
SC9	0.452	0.428	0.452	0.415	0.421	0.385	0.408	0.39
SC10	0.378	0.409	0.453	0.435	0.431	0.406	0.402	0.413
$SC_{(ave)}$	0.428	0.424	0.443	0.421	0.43	0.39	0.397	0.389

As shown in the above table, using direct K-means for clustering leads to unstable results due to the random selection of cluster centers. The average silhouette coefficient SC_{ave} is larger when K is between 3 and 7 and smaller for values of K from 8 to 10.

6.2.2. Clustering Using the Improved *K*-Means Algorithm

Calculate the Pearson correlation coefficient of the selected wind power daily output data, use greedy algorithm to find K curves with the lowest correlation corresponding to different K values, and use them as clustering centers for clustering. The cluster center curve numbers and contour coefficients corresponding to different K values are shown in Table 2.

Table 2. Improved K-means algorithm clustering results.

***K*-Value**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
SC	0.462	0.469	0.471	0.483	0.484	0.471	0.457	0.436

As shown in the above table, the improved K-means clustering algorithm has a significant improvement in clustering performance compared to traditional K-means clustering algorithms. The SC value reaches its maximum value of 0.484 when the K value is 7. Therefore, selecting 7 as the K value, the clustering effect is shown in Figure 4.

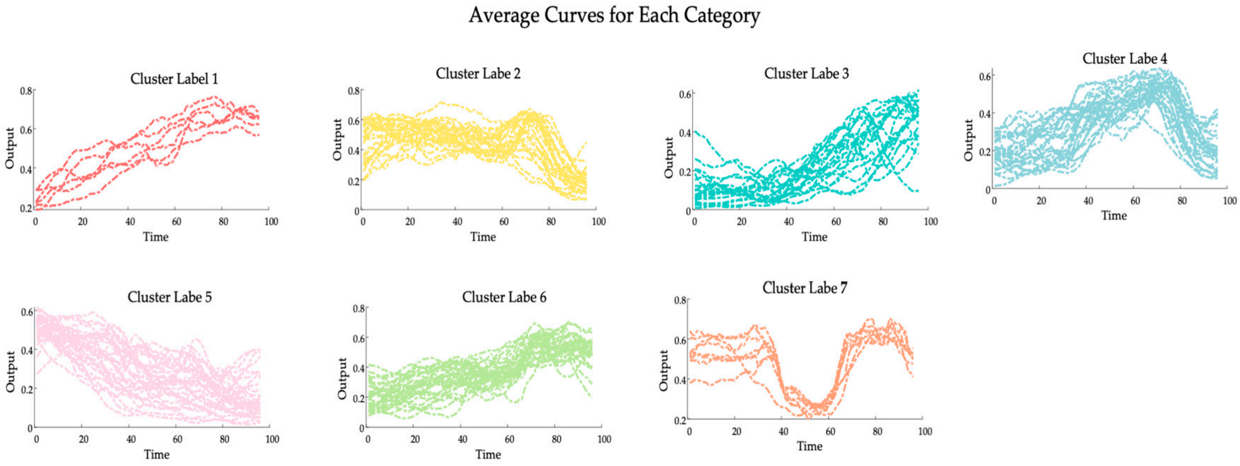

Figure 4. Clustering of the top 10% daily wind power output curves with peak-to-valley difference.

To further discover patterns, the mean of each cluster after clustering is calculated, as shown in Figure 5.

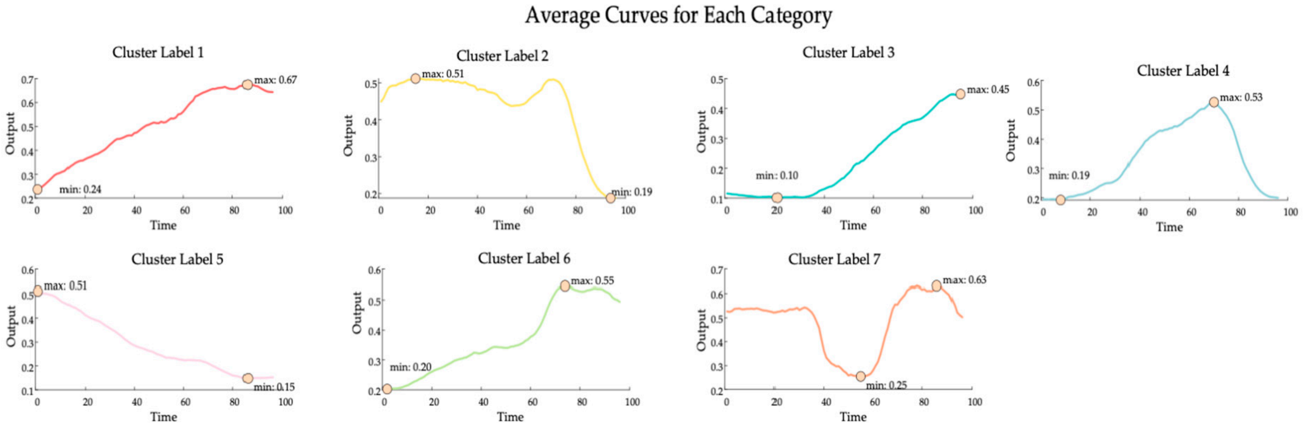

Figure 5. Mean clustering of the top 10% daily wind power output curves with peak-to-valley difference.

Based on the clustering results, the seven types of daily wind power output curves with significant peak-to-valley differences are as follows: rapid rise type, rapid decline type, slight decline followed by rapid rise type, large amplitude rise then fall type, large amplitude fall then rise type, stable decline followed by rapid decline type, and stable rise followed by rapid rise type.

6.3. Extraction of Typical Wind Power Curves

This paper employs two swarm intelligence algorithms, PSO and WOA, using the curve extraction method proposed in Section 4 for solution. The main parameters are shown in Table 3, the extraction results are illustrated in Figure 6, and specific numerical values are provided in Table 4.

Table 3. Key parameters of the wind power typical output curve extraction model.

Algorithm	Parameter	Value
PSO	Population Size	50
	Maximum Iterations	100
	Inertia Weight ω	0.7
	Acceleration Factor C1	1.5
	Acceleration Factor C2	1.5
WOA	Population Size	50
	Maximum Iterations	100
	Spiral Update Constant	1

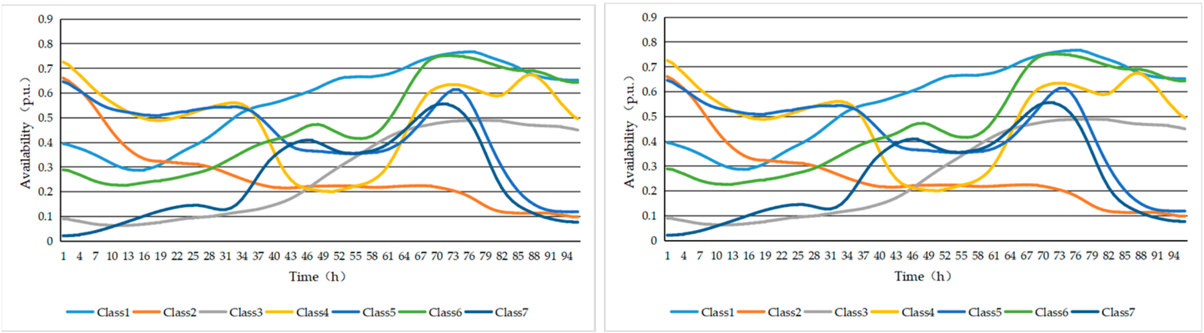

Figure 6. Typical curves extracted by PSO and WOA swarm intelligence algorithms.

Table 4. Data of typical daily wind power generation curves based on power system flexibility planning analysis.

Data Point	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
1	0.395	0.661	0.090	0.726	0.646	0.289	0.022
2	0.391	0.649	0.088	0.715	0.638	0.286	0.023
3	0.383	0.632	0.084	0.694	0.624	0.278	0.025
4	0.376	0.613	0.081	0.675	0.611	0.269	0.027
5	0.368	0.592	0.077	0.653	0.596	0.260	0.030
6	0.359	0.566	0.072	0.631	0.579	0.251	0.034
7	0.349	0.536	0.069	0.610	0.565	0.243	0.039
8	0.338	0.505	0.066	0.592	0.553	0.236	0.045
9	0.327	0.475	0.065	0.577	0.543	0.232	0.052
10	0.315	0.445	0.064	0.562	0.535	0.228	0.058
11	0.305	0.418	0.064	0.548	0.530	0.227	0.065
12	0.297	0.395	0.064	0.535	0.526	0.227	0.073
13	0.291	0.374	0.065	0.523	0.522	0.228	0.080
14	0.288	0.357	0.066	0.513	0.519	0.231	0.087
15	0.287	0.344	0.068	0.504	0.515	0.234	0.095
16	0.290	0.334	0.071	0.497	0.511	0.237	0.102
17	0.295	0.327	0.073	0.491	0.509	0.241	0.110
18	0.302	0.324	0.075	0.489	0.509	0.244	0.117
19	0.311	0.323	0.078	0.489	0.511	0.246	0.123
20	0.324	0.321	0.080	0.492	0.514	0.249	0.129
21	0.337	0.319	0.083	0.496	0.518	0.253	0.134
22	0.350	0.317	0.087	0.502	0.521	0.257	0.138
23	0.363	0.315	0.090	0.508	0.524	0.262	0.142
24	0.375	0.313	0.093	0.514	0.527	0.268	0.145
25	0.385	0.312	0.096	0.521	0.531	0.273	0.146
26	0.396	0.310	0.098	0.528	0.534	0.278	0.145
27	0.409	0.306	0.099	0.535	0.538	0.284	0.142
28	0.423	0.301	0.101	0.541	0.540	0.292	0.137
29	0.439	0.294	0.103	0.548	0.543	0.301	0.132
30	0.456	0.287	0.106	0.553	0.543	0.311	0.128
31	0.474	0.278	0.110	0.557	0.543	0.323	0.128
32	0.491	0.270	0.113	0.561	0.545	0.335	0.135
33	0.507	0.261	0.117	0.560	0.545	0.347	0.148
34	0.521	0.253	0.120	0.555	0.541	0.359	0.170
35	0.532	0.245	0.123	0.543	0.535	0.370	0.199
36	0.540	0.238	0.126	0.522	0.525	0.381	0.231
37	0.546	0.232	0.130	0.490	0.507	0.391	0.265
38	0.551	0.226	0.135	0.451	0.488	0.399	0.297
39	0.555	0.221	0.141	0.406	0.466	0.405	0.325
40	0.561	0.218	0.147	0.359	0.444	0.411	0.346
41	0.567	0.216	0.154	0.314	0.422	0.417	0.364
42	0.574	0.215	0.163	0.278	0.402	0.424	0.379
43	0.582	0.216	0.172	0.250	0.386	0.432	0.391
44	0.589	0.218	0.184	0.231	0.376	0.443	0.400
45	0.596	0.220	0.197	0.220	0.370	0.455	0.407
46	0.604	0.222	0.212	0.213	0.367	0.466	0.409
47	0.612	0.223	0.227	0.209	0.365	0.472	0.407
48	0.621	0.223	0.242	0.206	0.364	0.474	0.400
49	0.632	0.224	0.258	0.203	0.363	0.468	0.391
50	0.642	0.224	0.273	0.201	0.360	0.458	0.380
51	0.651	0.224	0.288	0.202	0.358	0.447	0.370
52	0.659	0.224	0.302	0.207	0.356	0.437	0.362
53	0.664	0.224	0.315	0.211	0.355	0.426	0.357
54	0.666	0.223	0.328	0.218	0.354	0.420	0.356
55	0.667	0.222	0.342	0.223	0.355	0.417	0.357
56	0.666	0.220	0.356	0.228	0.356	0.416	0.359

Table 4. *Cont.*

Data Point	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
57	0.666	0.219	0.370	0.235	0.358	0.419	0.362
58	0.667	0.218	0.383	0.245	0.359	0.426	0.366
59	0.669	0.218	0.397	0.259	0.361	0.440	0.373
60	0.673	0.219	0.410	0.279	0.365	0.460	0.382
61	0.677	0.220	0.421	0.307	0.372	0.487	0.395
62	0.684	0.222	0.432	0.342	0.382	0.520	0.411
63	0.692	0.223	0.441	0.383	0.395	0.555	0.430
64	0.701	0.224	0.449	0.430	0.411	0.593	0.449
65	0.712	0.225	0.455	0.478	0.432	0.632	0.470
66	0.722	0.225	0.461	0.524	0.455	0.667	0.491
67	0.731	0.225	0.465	0.561	0.479	0.697	0.511
68	0.738	0.224	0.470	0.591	0.503	0.722	0.530
69	0.745	0.222	0.474	0.610	0.530	0.738	0.544
70	0.751	0.219	0.478	0.622	0.556	0.747	0.554
71	0.755	0.215	0.481	0.629	0.581	0.751	0.556
72	0.759	0.210	0.484	0.634	0.601	0.752	0.554
73	0.762	0.205	0.487	0.634	0.614	0.751	0.547
74	0.765	0.199	0.488	0.633	0.615	0.749	0.534
75	0.767	0.191	0.489	0.630	0.599	0.747	0.514
76	0.767	0.181	0.489	0.623	0.570	0.744	0.487
77	0.766	0.168	0.489	0.615	0.530	0.739	0.450
78	0.761	0.155	0.490	0.607	0.484	0.733	0.405
79	0.752	0.143	0.489	0.599	0.435	0.726	0.355
80	0.744	0.132	0.489	0.592	0.387	0.719	0.303
81	0.737	0.124	0.488	0.588	0.341	0.712	0.255
82	0.730	0.119	0.487	0.592	0.300	0.705	0.215
83	0.722	0.116	0.484	0.605	0.263	0.699	0.185
84	0.714	0.114	0.480	0.626	0.231	0.695	0.162
85	0.704	0.113	0.476	0.647	0.203	0.692	0.145
86	0.692	0.113	0.473	0.666	0.181	0.691	0.132
87	0.681	0.114	0.470	0.675	0.163	0.691	0.121
88	0.673	0.114	0.469	0.673	0.148	0.689	0.112
89	0.667	0.114	0.468	0.658	0.137	0.685	0.103
90	0.662	0.114	0.467	0.638	0.129	0.678	0.096
91	0.659	0.112	0.466	0.612	0.124	0.670	0.091
92	0.656	0.110	0.465	0.584	0.121	0.660	0.086
93	0.654	0.107	0.462	0.557	0.119	0.653	0.082
94	0.653	0.103	0.459	0.533	0.119	0.647	0.079
95	0.653	0.099	0.455	0.510	0.119	0.644	0.077
96	0.654	0.097	0.451	0.496	0.120	0.642	0.076

The table illustrates that the particle swarm optimization (PSO) and whale optimization algorithm (WOA) yield identical representative curves for wind power output, corroborating the efficacy of both algorithms. PSO is celebrated for its straightforwardness and efficiency in nonlinear optimization challenges, albeit prone to entrapment in local optima. Conversely, WOA employs distinct strategies to circumvent local optima, enhancing its search robustness, albeit at the expense of increased iterations. The congruence in outcomes may stem from the characteristics of wind power output data and the algorithms' proficiency in converging upon solutions that faithfully encapsulate the data's inherent patterns. This efficiency likely results from the algorithms' adeptness in navigating and capitalizing on the search space, culminating in comparable optimal or suboptimal solutions.

7. Conclusions

This paper proposes a method for extracting typical daily wind power generation curves based on power system flexibility planning analysis and concludes the following:

1. The data preprocessing method, which combines Extended Kalman Filter with curve smoothing, can effectively improve data quality and retain the morphological characteristics of the wind power output curves.
2. The use of the Pearson correlation coefficient combined with the greedy algorithm to optimize *K*-means clustering can effectively solve the problems of selecting the *K* value and the random selection of cluster centers in the traditional *K*-means algorithm. The clustering results are more accurate and effective.
3. The use of the Pearson correlation coefficient combined with swarm intelligence algorithms such as PSO and WOA can effectively extract typical wind power output curves, with consistent calculation results.

This paper introduces a method for extracting typical daily wind power generation curves tailored to power system flexibility planning analysis. The seven distinct types of wind power output curves identified offer a robust data foundation for planning and analyzing power system flexibility. While primarily focusing on intra-day wind power output, the proposed algorithm is also applicable to longer-term analyses, including weekly, monthly, or seasonal time scales. However, due to the constraints of the data collected, future research will further explore the characteristics of wind power output over extended periods.

Author Contributions: Conceptualization, S.H. and J.Y.; Methodology, S.H.; Software, Y.W. and Y.Z.; Validation, C.C. and J.N.; Data curation, J.N.; Writing—original draft, J.Y.; Writing—review & editing, S.H. and Y.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Inner Mongolia Electric Power (Group) Co., Ltd Science and Technology Project, grant number 2023-4-6.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from Inner Mongolia Electric Power (Group) Co., Ltd and are available from the corresponding author with the permission of Inner Mongolia Electric Power (Group) Co., Ltd.

Conflicts of Interest: Author Sile Hu and Jianan Nan was employed by the company Inner Mongolia Power (Group) Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Lu, Z.; Li, H.; Qiao, Y. Challenges and planning for power systems with a high proportion of renewable energy resources. *Autom. Electr. Power Syst.* **2016**, *40*, 12.
2. Huang, F. Method for Generating Typical Renewable Energy Scenarios for Power System Analysis. Master's Thesis, Hefei University of Technology, Hefei, China, 2019.
3. Wang, J.; Zhang, Y.; Wan, X. Study on the characteristic index system and classification of typical curves for photovoltaic output. *Demand Side Manag. Electr. Power* **2017**, *19*, 5.
4. Wang, J.; Zhang, Y.; Wan, X.; Zhang, X. Study on the index system of output characteristics of new energy for grid operation—Wind power output characteristic index system. *Power Grid Clean Energy* **2016**, *32*, 11.
5. Liu, C.; Cao, Y.; Huang, Y.; Li, P.; Sun, Y.; Yuan, Y. Method for annual planning of wind power based on time series simulation. *Autom. Electr. Power Syst.* **2014**. [CrossRef]
6. Al-Otaibi, R.; Jin, N.; Wilcox, T.; Flach, P. Feature construction and calibration for clustering daily load curves from smart-meter data. *IEEE Trans. Ind. Inform.* **2017**, *12*, 645–654. [CrossRef]
7. Wu, H.; Zhu, C.; Zhang, Y.; Long, Y. A method for selecting typical days based on an improved fuzzy clustering algorithm. *Smart Power* **2022**, *50*, 60–67.
8. Lin, L.; Fei, H.; Liu, R.; Pan, X. Method for selecting typical wind power output scenarios based on hierarchical clustering algorithm. *Power Syst. Prot. Control* **2018**, *46*, 1–6.
9. Hu, X.; Wang, L.; Zhang, H.; Chang, Y.; Wang, Y. Research on an improved K-Means clustering algorithm based on lion group optimization. *Control Eng.* **2022**, *29*, 1996–2002. [CrossRef]
10. Wu, Y.; Gao, C.; Cao, H.; Chen, L.; Tang, J.; Li, H. Daily load curve clustering analysis based on grey wolf optimization clustering algorithm. *Power Syst. Prot. Control* **2020**, *48*, 68–76. [CrossRef]
11. Zhou, H.B.; Gao, J.T. Automatic method for determining cluster number based on silhouette coefficient. *Adv. Mater. Res.* **2014**, *951*, 227–230. [CrossRef]

12. Zhou, S.; Xu, Z.; Liu, F. Method for determining the optimal number of clusters based on agglomerative hierarchical clustering. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *28*, 3007–3017. [CrossRef]
13. Song, Y.; Li, H. Generation of wind and solar power output scenarios based on kernel density estimation and Copula function. *Electr. Technol.* **2022**, *23*, 56–63.
14. Li, J.; Wen, J.; Cheng, S.; Wei, H. Scenario generation method considering Copula correlation of multiple wind farm outputs. *Proc. CSEE* **2013**, *33*, 30–36.
15. Bai, B.; Han, M.; Lin, J.; Sun, W. Scenario reduction method for renewable energy with wind power and photovoltaics. *Power Syst. Prot. Control* **2021**, *49*, 141–149. [CrossRef]
16. Imran, M.; Hashim, R.; Khalid, N. An overview of particle swarm optimization variants. *Procedia Eng.* **2013**, *53*, 491–496. [CrossRef]
17. Gharehchopogh, S. A comprehensive survey: Whale Optimization Algorithm and its applications. *Swarm Evol. Comput.* **2019**, *48*, 1–24. [CrossRef]
18. Leung, H.; Zhu, Z. An aperiodic phenomenon of the extended Kalman filter in filtering noisy chaotic signals. *IEEE Trans. Signal Process.* **2000**, *48*, 1807–1810. [CrossRef]
19. Wang, C.; Ge, P.; Sun, L.; Wang, F. Research on user-side flexible load scheduling method based on greedy algorithm. *Energy Rep.* **2022**, *8* (Suppl. S14), 192–201. [CrossRef]
20. Wu, T.; Zhang, B.; Wang, Y.; Zhang, C. A comprehensive review of data cleaning. *Mod. Libr. Inf. Technol.* **2007**, 2. [CrossRef]
21. Ai, X.; Yang, Z.; Hu, H.; Wang, Z.; Peng, D.; Zhao, L. Load curve clustering method and application for VPP based on improved k-means algorithm. *Electr. Power Constr.* **2020**, *41*, 9.
22. Zhao, Y.; Lin, W. Study on typical new energy output scenarios based on Pearson correlation coefficient combined with density peak and entropy weight method. *Electr. Power* **2023**, *56*, 193–202.
23. Qi, S.; Yongjin, Y.; Yubin, W.; Haishu, G. Comprehensive optimization of distribution network using improved whale algorithm. *J. Power Syst. Autom.* **2021**, *33*, 8.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.