TYPE Original Research
PUBLISHED 23 January 2026
DOI 10.3389/fenrg.2025.1746997

OPEN ACCESS

EDITED BY
YongGang Li,
North China Electric Power University (Baoding), China

REVIEWED BY
Peishuai Li,
Nanjing University of Science and Technology, China
Qiang Fu,
The Hong Kong Polytechnic University, Hong Kong SAR, China
Lei Chen,
Yangzhou University, China

*CORRESPONDENCE
Chao Xu,
xuchao.21@163.com

RECEIVED 15 November 2025
REVISED 08 December 2025
ACCEPTED 18 December 2025
PUBLISHED 23 January 2026

CITATION
Ni J, Xu C, Fei K, Hu P, Xu Y, Li J and Wang N (2026) Multi-objective partitioning and cooperative peak shaving for distributed photovoltaic clusters via multi-agent reinforcement learning.
Front. Energy Res. 13:1746997.
doi: 10.3389/fenrg.2025.1746997

Multi-objective partitioning and cooperative peak shaving for distributed photovoltaic clusters via multi-agent reinforcement learning

Jun Ni[1], Chao Xu[1]*, Ke Fei[1], Peng Hu[2], Yan Xu[3], Jun Li[2] and Ning Wang[2]

[1]Wuxi Power Supply Company of State Grid Jiangsu Electric Power Co., Ltd., Wuxi, China, [2]Jiangsu Frontier Electric Technology Co., Ltd., Nanjing, China, [3]State Grid Jiangsu Electric Power Co., Ltd., Nanjing, China

Introduction: To address the urgent need for enhanced grid flexibility in high-penetration distributed photovoltaic (PV) systems, this paper proposes a novel two-stage framework integrating multi-objective cluster partitioning and multi-agent reinforcement learning (MARL) for cooperative peak shaving.
Methods: First, a comprehensive multi-objective optimization model defines PV cluster partitioning using three domain-specific metrics: available peak shaving margin (APSM), peak shaving synergy (PSS), and node correlation degree (NCD). This approach replaces conventional clustering algorithms, ensuring partitions balance electrical compactness, regulation capacity, and operational coordination. Subsequently, each partitioned cluster acts as an autonomous agent in a hierarchical MARL framework. Leveraging the multi-agent deep deterministic policy gradient (MADDPG) algorithm, agents collaboratively optimize active/reactive power allocation among clusters while coordinating unit-level control within clusters.
Results: Validated on the IEEE-33 node system, the strategy achieves a 29.39% reduction in peak-to-valley difference and 54.44% improvement in power balance deviation, significantly smoothing net load curves and enhancing voltage stability.
Discussion: This work demonstrates how MARL-driven coordination unlocks the inherent flexibility of distributed PV clusters, providing a scalable solution for grid-edge resource participation in system-level peak shaving.

KEYWORDS

cluster partitioning, cooperative peak shaving, distributed photovoltaics, MADDPG, multi-agent reinforcement learning

1 Introduction

In the context of an accelerated global energy transition, the changing generation mix and load profile have imposed significant operational challenges to power systems due to structural shifts in generation and load profiles (Sun et al., 2022). The rapid integration of distributed energy resources, along with the growing complexity of user-side load characteristics, has intensified peak-to-valley differences in distribution networks (Shi et al., 2025). In 2024, a total of 451.9 GW of PV capacity was added globally, with China accounting for 278 GW and India following at 24.5 GW, whose inherent intermittency and volatility pose substantial challenges to system power balance (Irena, 2025). While flexibility resources such as pumped hydro storage and electrochemical energy storage are being increasingly deployed, their large-scale adoption remains constrained by long construction cycles, strong geographical dependencies, and high costs (Aleksandrov et al., 2025). Moreover, demand-side response mechanisms are still immature and cannot independently fulfill system peak shaving demands (Song et al., 2025). Therefore, it is imperative to develop a novel peak shaving framework that fully exploits the regulatory potential of distributed resources to enhance system flexibility and robustness (Hu et al., 2022). However, unlocking this potential is hindered by two fundamental research gaps. First, it remains challenging to construct PV clusters that are simultaneously electrically compact and functionally adequate for peak shaving; neglecting the latter often leads to clusters with insufficient regulation margins despite their geographical proximity. Second, achieving effective inter-cluster coordination without heavy communication dependencies is critical but unresolved, as traditional centralized or iterative distributed controls struggle with the latency and privacy issues inherent in grid-edge resources.

In recent years, the clustering and partitioning of PV clusters have attracted growing research interest as a means to address scale complexity. Existing studies primarily employ clustering algorithms, optimization techniques, and deep learning methods using geographical, electrical, and operational data to achieve cluster partitioning (Miraftabzadeh et al., 2023). For instance (Jiao et al., 2024), applied a density-based clustering algorithm to smart meter measurement data for topology identification in low-voltage distribution networks, providing a data-driven foundation for electrically-aware partitioning (Li et al., 2024). Utilized a self-organizing map (SOM) for multi-step clustering and equivalent modeling of distributed PV, validating the rationality of cluster-based modeling (Wen et al., 2025). Designed and implemented an intelligent control device for distributed PV clusters, which demonstrated high-precision power control and significantly enhanced operational stability in practical applications. Despite these advancements, these methodologies typically prioritize topological or electrical distances while overlooking the explicit peak shaving regulation capacity required for active grid support. Consequently, traditional clustering methods often fail to produce partitions that balance regulation capability with electrical compactness, creating a mismatch between cluster formation and operational objectives. Thus, there is a critical need for a high-performance, multi-objective clustering approach capable of processing high-dimensional data to achieve precise and functionally meaningful cluster divisions.

Regarding the participation of distributed PV clusters in grid peak shaving, existing research has largely focused on coordinated strategies involving hybrid resources such as energy storage and conventional generation units (Shen et al., 2019). For example (Guo et al., 2021), demonstrated the effectiveness of energy storage in reducing peak-to-valley differences through timed charging and discharging (Li et al., 2025; Zare Ghaleh Seyyedi et al., 2025). Proposed an "aggregation–peak shaving–decomposition" model for PV-storage systems, showing that coordinated operation can alleviate peak shaving pressure while enhancing economic benefits (Zhou et al., 2025). Integrated PV, energy storage, and electric vehicles to effectively reduce the grid load peak-valley difference through optimal power rationing. Nonetheless, these studies predominantly address coordination between distributed PV and external devices, paying limited attention to the inherent peak shaving capabilities of distributed PV itself. More importantly, existing coordinated control methods usually rely on frequent information exchange, suffering from low efficiency and inadequate global optimization (Yin et al., 2024). This reliance on communication infrastructure contradicts the need for autonomous, resilient control at the grid edge, identifying a clear gap for decentralized strategies that can operate with minimal inter-agent communication while maintaining global system stability (He et al., 2025).

In response to the aforementioned challenges, this paper proposes an integrated two-stage framework that synergizes multi-objective cluster partitioning with MARL to unlock the cooperative potential of distributed PV clusters for systematic peak shaving. The main contributions of this work are threefold. First, moving beyond conventional clustering methods, we formulate a multi-objective optimization model for PV cluster partitioning based on three dedicated metrics-APSM, PSS, and NCD -to ensure the formed clusters are structurally sound and functionally adequate for coordinated control. Second, a novel hierarchical control paradigm is established where each partitioned cluster acts as an autonomous agent. Leveraging the MADDPG algorithm, the framework achieves de-centralized decision-making for cooperative power allocation among clusters and precise regulation within a cluster, all under a centralized-training-decentralized-execution scheme. Finally, the proposed strategy is rigorously validated on a modified IEEE-33 node system, demonstrating significant improvements in flattening the net load profile and enhancing voltage stability. The overall architecture of the proposed methodology, illustrating the seamless flow from cluster partitioning to cooperative control, is depicted in Figure 1.

As illustrated in Figure 1, the framework begins with the multi-objective partitioning of distributed PV systems into coherent clusters. Subsequently, these clusters are modeled as agents in the MARL environment, where they learn collaborative peak shaving policies through interaction with the distribution network. The remainder of this paper is organized as follows: Section 2 details the cluster partitioning methodology, Section 3 presents the MADDPG-based multi-agent peak shaving model, Section 4 provides case studies and analysis, and Section 5 concludes the paper.

2 Multi-objective partitioning model for distributed PV clusters

2.1 Formulation of the comprehensive index system

The goal of cluster division is to achieve internal balanced autonomy and minimize the power exchange between clusters.

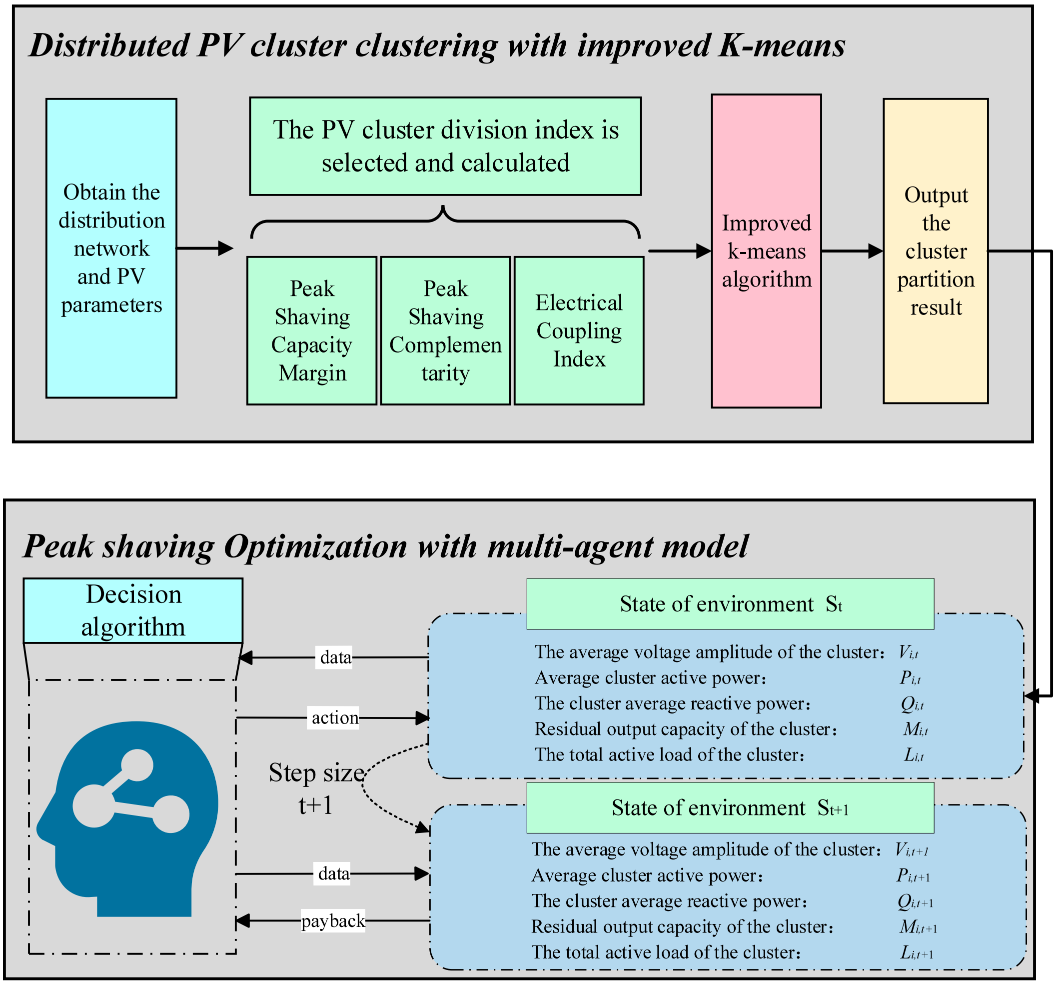

FIGURE 1
Framework of this paper.

Under the framework shown in Figure 1, the comprehensive indicators for PV cluster clustering used in this paper can be classified into three categories based on their characteristics: APSM, PSS, and NCD indicators.

APSM is utilized to quantify the aggregated adjustable power capacity of the PV cluster, providing a fundamental data basis for evaluating its potential contribution to distributed peak shaving. The PSS index assesses the temporal alignment between local PV generation and load demand. A higher PSS indicates that the load profile closely follows the PV output curve, which minimizes the net load deviation and optimizes the cluster's self-balancing capability on a time scale. NCD index evaluates the electrical coupling strength among nodes, ensuring that the formed clusters are electrically compact and structurally suitable for coordinated voltage regulation. The specific indicators are defined as follows:

1. As can be seen from Figure 2, by adjusting the working voltage of the photovoltaic according to the photovoltaic characteristic curve, the output power of the photovoltaic can be changed. In order to effectively control the output of photovoltaic power, it is necessary to ensure that the leading nodes within the photovoltaic cluster have sufficient regulation capacity. In this paper, the MPPT algorithm is adopted, and there are two operating regions on its left and right. Given the reference power P_{ref}, there are two operating points A and B corresponding to it.

If the photovoltaic system operates on the left side of the MPPT, the photovoltaic working voltage needs to be reduced to cut down the photovoltaic output power, that is: $U_{pv,ref} = U_{pv,ref} - U_{step}$.

If the photovoltaic system operates on the right side of the MPPT, the photovoltaic working voltage needs to be increased to reduce the photovoltaic output power, that is: $U_{pv,ref} = U_{pv,ref} - U_{step}$.

The specific formula for peak shaving margin is defined as Equation 1:

$$P_{rest} = P_{\max} - P_{pv} \tag{1}$$

where $P_{\max}$ represents the upper limit of the output of the photovoltaic cluster, P_{pv} is the current target power between $P_{\max}$ and P_{ref}, and P_{rest} indicates the APSM of this photovoltaic cluster.

2. PSS refers to the degree of matching between the renewable energy output power of the photovoltaic cluster as a whole

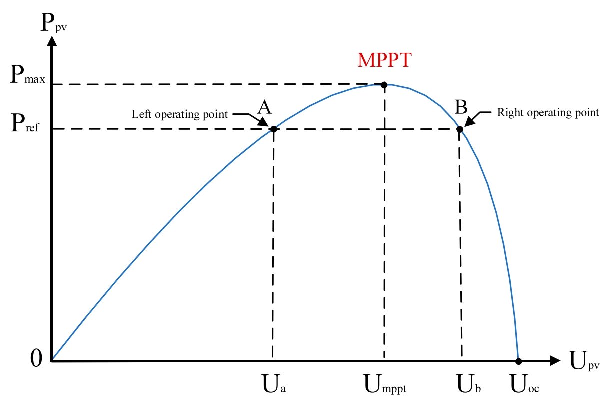

FIGURE 2
Changes in operating points when PV clusters participate in power grid peak regulation.

and the optimized energy demand within a certain time scale, which is demonstrated through the coordinated dispatching among the nodes within the cluster.

As photovoltaic power is an uncontrollable resource, the cluster needs to establish a net load optimization model. The optimization object is all nodes within the cluster. The optimization objective is to make the optimized load curve as close as possible to the PV cluster's power generation curve in terms of time sequence, that is, to minimize the net load within the cluster, as shown in Equations 2, 3.

$$\Delta P_k^t = \sum_{i\in k} P_{load,i}^t - \sum\nolimits_{i\in k} P_{pv,i}^t \tag{2}$$

$$f_\varphi = \min \sum_{t=1}^{T} \left|\Delta P_k^t\right| \tag{3}$$

where ΔP_t^k is the net load value of the *K-th* cluster after optimized scheduling at time t; P_t^k and $P_{re,k}^t$ are the load value within cluster k at time t and the total power value of renewable energy; f_φ is the optimization objective of the net load optimization model; T represents the scheduling period. In order to maximize the utilization of energy within the cluster during the optimal operation stage of the distribution network and reduce cross-regional power interaction, this paper takes the net load size within the cluster after optimized dispatching as the core basis for determining the matching degree of source-load-storage resources within the cluster.

Considering that both photovoltaic output and load output have the characteristic of fluctuating over time, it is necessary to select the average value of the cluster at a specific time scale to quantify the indicators. Based on this, the functional indicators of cluster division and the peak shaving coordination indicators can be expressed as Equations 4, 5:

$$\varphi_{P,k}^t = \begin{cases} \dfrac{P_{pv,k}^t}{P_k^t}, & \Delta P_k^t > 0 \\ 1, & \Delta P_k^t = 0 \\ \dfrac{P_k^t}{P_{pvk}^t}, & \Delta P_k^t < 0 \end{cases} \tag{4}$$

$$\varphi_P = \frac{1}{N_C}\int_1^{N_C}\left(\frac{1}{T}\int_0^T \varphi_{P,k}(t)\,dt\right)dk \tag{5}$$

where $\varphi_{P,k}^t$ is the degree of coordination between PV output and load in the *K-th* cluster at time t, and $\varphi_{P,k}^t \in (0,1]$. φ_P is the peak regulation coordination index of the entire distribution network, N_C is the number of clusters. The more the optimized net load tends to 0, the closer the optimized load curve is to the photovoltaic output curve, indicating that the better the interaction matching degree of internal load resources in the cluster is, and the larger the value will be.

3. This index is based on the modularity concept (Kumari et al., 2025), which is used to measure the structural strength of the community network. Under the given power grid topology, line parameters and reference operating point, the definition of the cluster comprehensive performance index ρ based on electrical distance and cluster comprehensive reserve degree is presented in Equation 6:

$$\rho = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(i,j) \tag{6}$$

where m (calculated as Equation 7) represents the total weight of all the edges of the nodes in the cluster, A_{ij} is the edge weight of nodes i and j, $k_i k_j$ is the product of the total weight of the edges connected to node i and node j, and $\delta(i,j)$ is the identification function, defined as Equation 8.

$$\begin{cases} m = \left(\sum_i \sum_j A_{ij}\right)/2 \\ k_i = \sum_j A_{ij} \\ k_j = \sum_i A_{ij} \end{cases} \tag{7}$$

$$\delta(i,j) = \begin{cases} 1 & \text{Node } i \text{ and Node } j \text{ are in the same cluster} \\ 0 & \text{else} \end{cases} \tag{8}$$

In the distribution network, the electrical distance based on which the network edge weight is determined is generally defined by the sensitivity relationship between power and voltage amplitude (Cai et al., 2020), as the changes of active and reactive power will significantly affect the voltage. The power-voltage sensitivity matrix of active and reactive power can be calculated by the power flow correction equation of Newton-Raphson method in polar coordinate form.

The Jacobian matrix is transformed to obtain the sensitivity matrix, as shown in Equation 9:

$$\begin{bmatrix} \Delta\theta \\ \Delta V \end{bmatrix} = \boldsymbol{J}^{-1}\begin{bmatrix} \Delta P \\ \Delta Q \end{bmatrix} = \begin{bmatrix} \boldsymbol{S}_{P\theta} & \boldsymbol{S}_{Q\theta} \\ \boldsymbol{S}_{PV} & \boldsymbol{S}_{QV} \end{bmatrix}\begin{bmatrix} \Delta P \\ \Delta Q \end{bmatrix} \tag{9}$$

where $\boldsymbol{J}^{-1}$ is the inverse of Jacobian matrix; $\boldsymbol{S}_{P\theta}$ is the active power-voltage phase Angle sensitivity matrix; $\boldsymbol{S}_{Q\theta}$ is reactive power-voltage phase Angle sensitivity matrix; $\boldsymbol{S}_{PV}$ is the active power-voltage amplitude sensitivity matrix; $\boldsymbol{S}_{QV}$ is the reactive power-voltage amplitude sensitivity matrix.

The variation of voltage amplitude at any node and the variation of injected active and reactive power at all nodes are expressed as Equation 10:

$$\begin{bmatrix} \Delta V_1 \\ \Delta V_2 \\ \vdots \\ \Delta V_n \end{bmatrix} = \boldsymbol{S}_{PV} \begin{bmatrix} \Delta P_1 \\ \Delta P_2 \\ \vdots \\ \Delta P_n \end{bmatrix} + \boldsymbol{S}_{QV} \begin{bmatrix} \Delta Q_1 \\ \Delta Q_2 \\ \vdots \\ \Delta Q_n \end{bmatrix} \tag{10}$$

Thus, the electrical distances $A_{VQ}(i,j)$ and $A_{VP}(i,j)$ of nodes i and j based on re-active voltage sensitivity and active voltage sensitivity can be obtained as Equation 11:

$$\begin{cases} A_{QV}(i,j) = S_{QV}(i,i) - S_{QV}(i,j) - S_{QV}(j,i) + S_{QV}(j,j) \\ A_{PV}(i,j) = S_{PV}(i,i) - S_{PV}(i,j) - S_{PV}(j,i) + S_{PV}(j,j) \end{cases} \tag{11}$$

where $S_{VQ}(i,i)$ and $S_{VP}(i,i)$ are the sensitivity coefficients of injected reactive power and active power of node i to its own voltage, respectively. $S_{VQ}(j,j)$ and $S_{VP}(j,j)$ are the sensitivity coefficients of injected reactive and active power of node j to its own voltage, respectively. $S_{VQ}(i,j)$ and $S_{VP}(i,j)$ are the sensitivity coefficients of the injected reactive power and active power of node j to the voltage of node i, respectively. $S_{VQ}(j,i)$ and $S_{VP}(j,i)$ are the sensitivity coefficients of the injected reactive and active power at node i to the voltage at node j, respectively

The following can be obtained as Equation 12:

$$A_{ij} = 1 - \frac{A_{QV}(i,j) + A_{PV}(i,j)}{2 \max A} \tag{12}$$

where A is the electrical distance matrix composed of the electrical distance between any two nodes in the cluster.

Finally, the comprehensive cluster clustering index $d = \{P_{rest}, \varphi_P, A_{ij}\}$ was generated by combining the APSM index, PSS index and NCD indices.

2.2 Multi-objective optimization model and solution method

PV cluster clustering is based on the geographical location, output characteristics, meteorological data and other multi-dimensional characteristics of PV, using clustering algorithm to extract and group its features, photovoltaic power stations with similar output patterns are divided into the same cluster, so as to achieve large-scale management and cooperative peak regulation control.

The clustering index data $d = \{P_{rest}, \varphi_P, \rho\}$ selected above is normalized so that its mean value is 0 and its standard deviation is 1. The processed data can be expressed as Equation 13:

$$d' = \frac{d - d_{\min}}{d_{\max} - d_{\min}} \tag{13}$$

where d' is the normalized value of each clustering index data, d is the actual value of each clustering index data, $d_{\min}$ is the minimum value of each clustering index data, and $d_{\max}$ is the maximum value of each clustering index data.

The improved K-means clustering algorithm that optimizes the initial clustering center is simple and practical. This paper uses the clustering algorithm to divide the PV into different clusters according to the selected clustering index. Suppose that there is n distributed PV systems in the area, forming cluster $Y = \{y1, y2, \ldots\ldots, yn\}$. The optimal number of clusters is determined as K by the elbow method, where $1 < K \le y_n$:

Minimizing the squared error is expressed as Equation 14:

$$Q^{\mathrm{PV}}(J_i, m_i) = \sum_{i=1}^{k} \sum_{m_i \in J_i} \left\| x_i - \frac{1}{|J_i|} \sum_{m_i \in J_i} {m_i}^2 \right\|_2^2 \tag{14}$$

The conventional K-means algorithm is inherently limited by its sensitivity to initial centroid selection and outlier data points, which often leads to suboptimal clustering outcomes. In the context of peak shaving in power systems, such limitations can significantly undermine the effectiveness of cluster-based control strategies. To over-come these challenges and ensure that the clustering results are both structurally sound and functionally aligned with peak shaving requirements, this paper introduces an enhanced K-means methodology. This approach integrates three domain-specific clustering indices-namely, the APSM, PSS, and NCD-to reformulate the clustering process as a multi-objective optimization problem. The objective is to concurrently optimize several competing criteria, thereby achieving a balanced and technically viable partition of PV clusters. This formulation provides the foundation for the multi-objective optimization model, defined as Equation 15:

$$\begin{cases} \min & f_1 = \sum_{k=1}^{K} \sum_{i,j \in C_k} A(i,j) \\ \max & f_2 = \sum_{k=1}^{K} P_{restk} \\ \min & f_3 = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(\varphi_{Pk} - \overline{\varphi_P} \right)^2} \\ \text{s.t.} & \sum_{k=1}^{K} z_{ik} = 1, \forall i \in N \\ & \sum_{i \in N} z_{ik} \ge 1, \forall k = 1, 2, \cdots, K \end{cases} \tag{15}$$

In the formulated model, the objective function f_1 aims to maximize the electrical compactness within each cluster, thereby enhancing structural coherence; f_2 seeks to maximize the total peak-shaving capacity of the system, ensuring sufficient regulation margin across clusters; and f_3 achieves balanced distribution of peak-shaving responsibility among clusters by minimizing the standard deviation of power regulation coefficients. The set of PV nodes is denoted as N, and the predetermined number of clusters is K. The model employs binary decision variables z_{ik} to define node-cluster assignments, constrained such that each node must belong to exactly one cluster, thereby ensuring solution feasibility and practical applicability. The complete work-flow for solving this multi-objective optimization problem is illustrated in Figure 3.

Based on the enhanced K-means clustering algorithm and integrated with a comprehensive peak shaving evaluation index-comprising the APSM, PSS, and NCD indices-the optimal partitioning scheme for each cluster is determined. Finally, the load clustering results are generated and visualized. The detailed procedure is as follows:

1. Data Preprocessing and Similarity Matrix Construction: First, the three key characteristic indicators for each photovoltaic

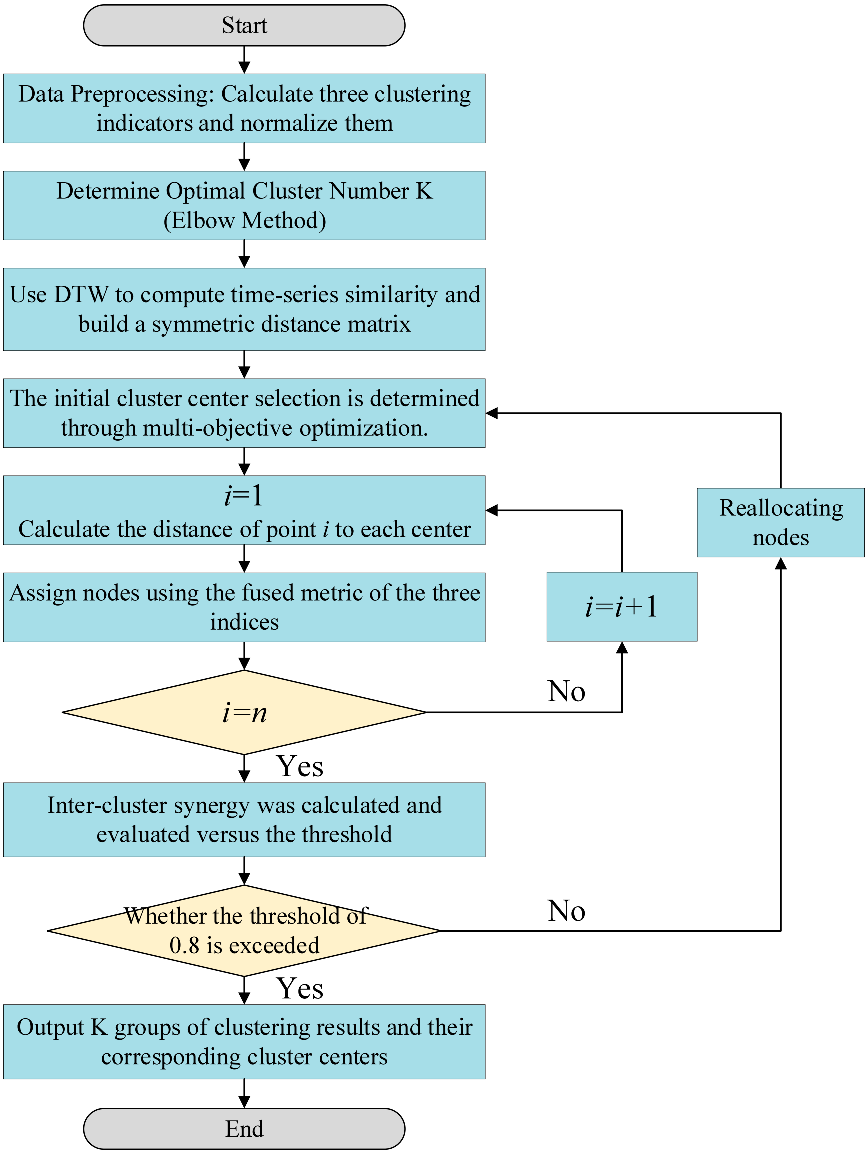

FIGURE 3
Flowchart of PV cluster clustering based on improved *K*-means.

(PV) node in the region are computed using the selected clustering metrics, and a dynamic time warping (DTW) distance matrix is constructed.

2. Determination of Optimal Cluster Number: The optimal number of clusters, K, is determined using the Elbow Method or other appropriate criteria. This ensures a practical balance between clustering compactness and the operational feasibility of the peak shaving task.
3. Initial Centroid Generation: Multi-objective Optimization Guidance. The initialization process is guided by two stages: The first initial centroid is selected as the midpoint of the node pair exhibiting the minimum distance in the DTW matrix. Subsequently, an improved k-mean sampling strategy guided by multi-objective optimization is adopted. The minimum weighted distance from each sample to the already selected centroids is calculated by fusing the three indicators' weights, and the next centroid is chosen with a probability proportional to the squared distance.
4. Iterative Assignment and Centroid Update: For each node, its comprehensive distance to the K centroids is calculated by integrating the three characteristic indicators, and the node is assigned to the nearest cluster. The cluster centroids are then updated by the mean or weighted mean of all intra-cluster samples. This assignment-update process repeats until the centroids stabilize.
5. Synergy Constraint Verification: For each pair of clusters, the Inter-Cluster Contribution Correlation is calculated. It is then verified whether the synergy coefficient for all cluster pairs exceeds the predefined threshold, which is set at 0.8 in this study. For each cluster c and time,

define the cluster's direct contribution to balance deviation improvement as Equation 16:

$$S_c(t) = \frac{|\Delta P_{\text{rest}}(t)| - |\Delta P(t)|}{|\Delta P(t)| + \epsilon} \tag{16}$$

where $|\Delta P_{\text{rest}}(t)|$ denotes the improvement under PV regulation, $|\Delta P(t)|$ denotes the baseline deviation without PV regulation, and $\epsilon > 0$ is a small constant to avoid division by zero.

Over the evaluation window $(T = 1, \ldots, T)$, the mean contribution of cluster c is calculated as Equation 17:

$$\overline{S}c = \frac{1}{|T|} \sum t = 1^T S_c(t) \tag{17}$$

The peak-shaving synergy coefficient between clusters m and n is defined as the Pearson correlation of their contribution time series, as shown in Equation 18:

$$\gamma_{mn} = \frac{\sum_{t=1}^{T} \left(S_m(t) - \overline{S}_m\right)\left(S_n(t) - \overline{S}_n\right)}{\sqrt{\sum_{t=1}^{T} \left(S_m(t) - \overline{S}_m\right)^2} \sqrt{\sum_{t=1}^{T} \left(S_n(t) - \overline{S}_n\right)^2}} \tag{18}$$

where γ_{mn} denotes the Inter-Cluster Contribution Correlation between cluster m and cluster n, which is a scalar value ranging from −1 to 1.

6. Synergy-based Node Reallocation and Local Optimization:For any cluster pair that violates the synergy constraint, critical nodes identified as having the largest positive marginal impact on synergy improvement are reallocated. After each reallocation, the synergy coefficients are recalculated, and the process returns to Step (4) to continue the iteration until all cluster pairs satisfy the threshold or a maximum iteration/reallocation limit is reached.
7. Termination and Output: The algorithm terminates upon satisfying the synergy threshold or reaching the maximum iteration/reallocation limit. The final output includes cluster labels, centroid positions, and evaluation metrics, which are then utilized for subsequent peak shaving task allocation and visualization.

In summary, this section has established a comprehensive multi-objective partitioning model for distributed PV clusters, integrating three domain-specific metrics-APSM, PSS, and NCD-to guide the clustering process. By reformulating the partitioning task as a multi-objective optimization problem and employing an improved K-means algorithm, the proposed model ensures that the resulting clusters are not only electrically compact and structurally coherent but also functionally equipped for coordinated peak shaving.

Building upon this well-structured cluster foundation, the next section will present a hierarchical multi-agent reinforcement learning framework designed to leverage these clusters as autonomous decision-making agents. The goal is to enable cooperative peak shaving through decentralized control, thereby achieving system-wide optimization in both active and reactive power dispatch while maintaining voltage stability and operational efficiency.

3 A hierarchical multi-agent reinforcement learning framework for cooperative peak shaving

3.1 Formulation of multi-agent peak shaving environment

In machine learning, an agent collects observations from its environment and employs a policy to make decisions based on these observations. Upon executing an action, the agent influences the environment, leading to a state transition. Subsequently, the agent receives a reward determined by the new environmental state, after which it initiates the next cycle of decision-making. The definitions of the agent, observation space, action space, and reward function are detailed in references (Cao et al., 2020; Cao et al., 2022).

3.1.1 Agent definition

The distributed PV system is divided into K clusters by the modified K-means clustering of Section 2. In this section, each PV cluster is defined as an agent, and the multi-agent system is constructed. The set of agents is set as Equation 17:

$$Agent = \{Ag_1, Ag_2, \ldots, Ag_K\} \tag{19}$$

where Ag_K is the corresponding K-*th* cluster, and each cluster acts as a decision-making body responsible for the power regulation of its internal PV unit. The co-operative relationship between clusters is directly determined by the electrical coupling strength characterized by the NCD.

3.1.2 Observation space design

Since this paper considers the PV output under cluster control, and each agent makes decisions only based on the electrical information of the distributed PV access nodes under its own control, the observation space of the designed agent is shown in Equation 20.

$$O_t^i = \{\bar{V}_{i,T}, \bar{P}_{i,T}, \bar{P}_{i,T}, M_{i,T}, L_{i,T}\} \tag{20}$$

where $\bar{V}_{i,T}$ is the average voltage amplitude of cluster i at time period T, reflecting the overall voltage level of the cluster. $\bar{P}_{i,T}$ is the average active power of the average voltage amplitude of cluster i at time period T. $\bar{Q}_{i,T}$ is the average active power load of cluster i at time period T. $M_{i,T}$ is the remaining output capacity of cluster i in time period T. $L_{i,T}$ is the sum of the active power consumed by all users in cluster i in time period T.

3.1.3 Action space design

The peak shaving problem studied in this paper needs to address two key objectives simultaneously: power balance and voltage quality. To achieve this goal, the action space is designed to coordinate the active and reactive power adjustment instructions of the PV inverter, as shown in Equation 21:

$$A_{i,T} = \left[\Delta P_{i1,T}, \Delta Q_{i1,T}, \Delta P_{i2,T}, \Delta Q_{i2,T}, \ldots, \Delta P_{in_i,T}, \Delta Q_{in_i,T}\right] \tag{21}$$

where n_i is the number of PV units in cluster i, $\Delta P_{i1,T}$ is the active power regulation of the j-*th* PV unit in cluster i at time period T, and $\Delta Q_{ij,T}$ is the reactive power regulation of the j-*th* PV unit in cluster i at time period T.

In order to meet the operation constraints of the PV inverter, this paper uses tanh function as the activation function of the output layer of the neural network, limits the original action output in the range of [-1,1], and then obtains the actual power regulation amount through linear scaling, as Equations 22, 23.

$$\begin{cases} \Delta P_{ij,T} = \alpha_{ij,T} \cdot \tanh\left(a_{ij,P}^{\text{raw}}\right) \\ \Delta Q_{ij,T} = \beta_{ij,T} \cdot \tanh\left(a_{ij,Q}^{\text{raw}}\right) \end{cases} \tag{22}$$

$$\begin{cases} \alpha_{ij,T} = P_{ij,T} - P_{\min} \\ \beta_{ij,T} = \sqrt{S_{ij}^2 - \left(P_{ij,T} + \Delta P_{ij,T}\right)^2} \end{cases} \tag{23}$$

where $a_{ij,P}^{\text{raw}}$ and $a_{ij,Q}^{\text{raw}}$ are the original active and reactive power action values output by the neural network, respectively. $\alpha_{ij,T}$ is the maximum active power reduction of the PV unit at the current time, where $P_{ij,T}$ is the current active power output and $P_{\min}$ is the minimum allowed active power output. $\beta_{ij,T}$ is the maximum adjustable reactive power of the PV unit at the current time.

This design method naturally satisfies the following inverter capacity constraints and active power regulation priority constraints, as shown in Equation 24:

$$\begin{cases} \left(P_{ij} + \Delta P_{ij}\right)^2 + \left(Q_{ij} + \Delta Q_{ij}\right)^2 \le S_{ij}^2 \\ \left|\Delta P_{ij,T}\right| \le \eta \cdot \alpha_{ij,T}^{\max} \end{cases} \tag{24}$$

where $\eta \in [0,1]$ is the active power regulation coefficient, which reflects the system's acceptable level of generation loss. $\eta = 0$ denotes that active power reduction is forbidden, and $\eta = 1$ denotes that maximum active power reduction is allowed.

3.1.4 Reward function design

The reward function guides the learning direction by quantifying how good or bad the agent's behavior is. In this paper, multiple objectives such as voltage quality, network loss, peak shaving effect and cluster coordination are comprehensively considered, and a multi-objective reward function is designed as follows

a. Peak shaving effect reward, R_{peak} is the core goal of this paper, aiming to achieve power balance through active power regulation, and its calculation formula is presented in Equation 25.

$$\begin{cases} R_{\text{peak}} = -k_1 \cdot \left|\dfrac{P_{\text{net},i}}{P_{\text{base},i}}\right| - k_2 \cdot \max\left(0, P_{\text{net},i} - P_{\text{limit}}\right) \\ P_{\text{net},i} = \sum\limits_{j \in Ci} \left(PPV, j + \Delta P_{ij,T}\right) - P_{\text{base},i} \\ P_{\text{base},i} = \dfrac{1}{T} \sum\limits_{t=1}^{T} \sum\limits_{j \in C_i} P_{L,j}(t) \end{cases} \tag{25}$$

where $P_{\text{net},i}$ is the net power of the cluster, $P_{\text{base},i}$ is the benchmark value that the cluster conforms to, P_{limit} is the upper limit of the net power allowed, k_1, k_2 is the weight coefficient, and $k_2 \gg k_1$.

b. Voltage quality reward R_{voltage} ensures system voltage safety in the peak shaving process, as defined in Equation 26:

$$R_{\text{voltage}} = -\sum_{j \in C_i} \left(V_{j,T} - V_{\text{ref}}\right)^2 - \lambda \cdot N_{\text{violation}} - \mu \cdot \sigma_{V,i} \tag{26}$$

Among them: the first term penalizes voltage deviation and encourages voltage stability at the rated value; The second term imposes a penalty on the number of nodes that exceed the voltage limit. The third term penalizes intra-cluster voltage inconsistencies.

c. Economic operation reward $R_{economic}$ Balance peak shaving effect and economic cost, as calculated in Equation 27:

$$R_{economic} = -c_p \cdot \sum_{j=1}^{n_i} \max\left(0, -\Delta P_{ij,T}\right) - c_q \cdot \sum_{j=1}^{n_i} \left|\Delta Q_{ij,T}\right| - c_l \cdot P_{loss,i} \tag{27}$$

where the first term only penalizes active power reduction $\Delta P_{ij,T}$, and does not penalize active power increase; The second penalty reactive power adjustment range, avoid excessive movement of equipment; Network loss cost: The third term penalizes the system active power loss.

In the framework of clustering agent peak-shaving, the system state is transferred deterministic by power flow calculation, and the discount factor is set to coordinate the current peak-shaving effect and long-term operation performance. This design enables each cluster agent to make distributed decisions based on local information, while ensuring global performance through multi-period optimization. Figure 4 shows the distributed peak shaving control framework of the cluster agent.

3.2 MADDPG-based control architecture

MADDPG is a deep reinforcement learning algorithm for multi-agents in partially observable environments, which is configured with an Actor network and a Critic network for each agent. At training time, the Critic network can use the global state and the actions of all agents; However, the Actor network can only use local observation states. When executed, only the Actor network is needed and decisions are made based on local observations.

In this paper, the PV cooperative peak shaving problem is modeled as a cooperative MARL problem. Based on the clustered multi-agent environment constructed in Section 3.1, the basic components of the MADDPG algorithm are as follows. In this section, the environment state space, agent observation space, action space and reward function in MADDPG framework are formally defined, aiming to guide multiple cluster agents to coordinately control the active and reactive power output of their internal PV units according to local observation data, and finally realize the comprehensive optimization goal of global peak shaving and voltage stability in distribution network. The flowchart of the MADDPG algorithm is shown in Figure 5.

In MADDPG, The Actor network functions as a decentralized controller, mapping the local observation to a deterministic action. Following the standard notation in (Cao et al., 2020), we simplify the deep neural network representation as a parameterized policy function. For the $i-th$ agent, the action a_i is generated as Equation 28:

$$a_t^i = \mu_i\left(o_t^i | \theta_i^\pi\right) \tag{28}$$

10.3389/fenrg.2025.1746997

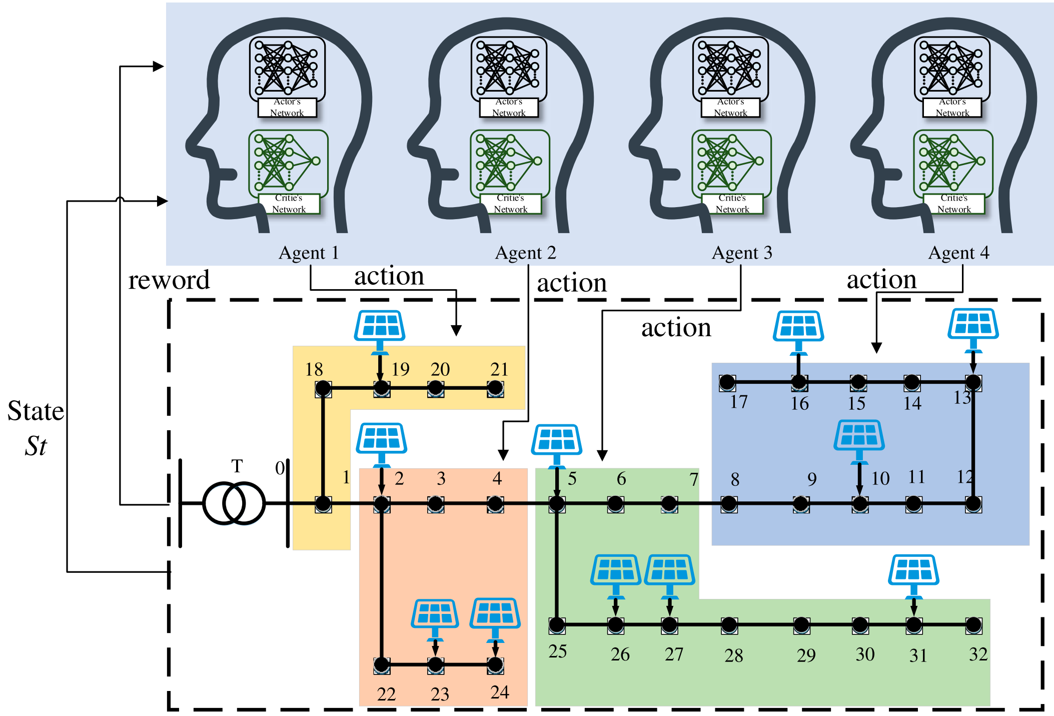

FIGURE 4
Distributed peak shaving control framework based on clustering agents.

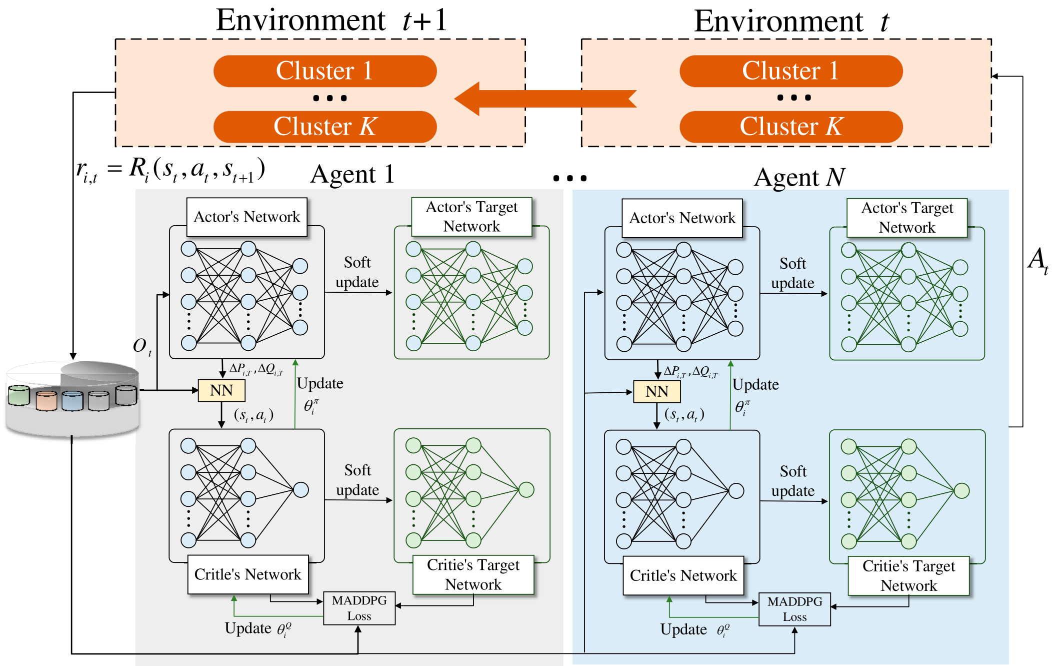

FIGURE 5
Flow chart of MADDPG algorithm.

where μ_i denotes the policy function parameterized by weights θ_i^{π}, and o_i is the local observation vector defined in Section 3.1. The output layer employs a hyperbolic tangent (tanh) activation function to ensure the action is bounded within the feasible regulation range $[-1, 1]$.

The Critic network serves as a centralized evaluator during the training phase to address the non-stationarity of the multi-agent environment. It estimates the joint action-value function by taking the global state information and joint actions as inputs. The Q-value for agent i is calculated as Equation 29:

$$Q_i = Q_i\left(\mathbf{o}, \mathbf{a}; \theta_i^Q\right) \tag{29}$$

where $\boldsymbol{o} = \{o_1, \ldots, o_N\}$ and $\boldsymbol{a} = \{a_1, \ldots, a_N\}$ represent the set of observations and actions from all N clusters, respectively, and θ_i^Q represents the parameters of the Critic network. This centralized critic structure allows the agent to implicitly learn the cooperative strategies embedded in the complex power grid interactions.

Critic Network Optimization: Based on temporal difference learning, the Critic network is updated by minimizing the Bellman error, defined in Equation 30.

$$\begin{cases} \mathcal{L}\left(\theta_i^Q\right) = \mathbb{E}_{(\mathbf{o},\mathbf{a},r,\mathbf{o}')\sim\mathcal{D}}\left[\left(y_i - Q_i\left(\mathbf{o}, \mathbf{a} | \theta_i^Q\right)\right)^2\right] \\ y_i = r_i + \gamma Q_i^{\text{target}}\left(\mathbf{o}', \mathbf{a}' | \theta_i^{Q'}\right) \\ \mathbf{a}' = \left[\pi_1^{\text{target}}(o_1'), \ldots, \pi_K^{\text{target}}(o_K')\right] \end{cases} \tag{30}$$

Update the gradient parameters as per Equation 31:

$$\nabla_{\theta_i^Q}\mathcal{L} = -2\mathbb{E}\left[\left(y_i - Q_i\right) \cdot \nabla_{\theta_i^Q} Q_i(\mathbf{o}, \mathbf{a})\right] \tag{31}$$

Actor Network Optimization: Based on the policy gradient theorem, the Actor network is updated in the direction of boosting Q-values, as shown in Equation 32:

$$\nabla_{\theta_i^\mu} J \approx \mathbb{E}_{\mathbf{x},a}\left[\nabla_{\theta_i^\mu}\mu_i(o_i) \cdot \nabla_{a_i} Q_i(\mathbf{x}, a_1, \ldots, a_N)|_{a_i=\mu_i(o_i)}\right] \tag{32}$$

To stabilize the training, the soft update mechanism is used to synchronize the target network, as described in Equation 33:

$$\begin{cases} \theta_i^{Q'} \leftarrow \tau\theta_i^Q + (1-\tau)\theta_i^{Q'} \\ \theta_i^{\pi'} \leftarrow \tau\theta_i^{\pi} + (1-\tau)\theta_i^{\pi'} \end{cases} \tag{33}$$

where $\tau \ll 1$ is the soft update coefficient.

Upon completion of the training process, the network parameters are fixed, and only the Actor components are retained for deployment. Each agent's actuators can then make real-time decisions based on local observations. Since the augmented information is utilized exclusively by the Critic during training, agents are capable of exhibiting cooperative behavior and generating decisions that remain robust with respect to the actions of other agents, relying solely on local information. The flow of the real-time decentralized control algorithm is presented in Algorithm 1.

Initialize replay buffer $\mathcal{D}$, actor networks θ_i^{π}, critic networks θ_i^Q, and target networks $\theta_i^{\pi'}\theta_i^{Q'}$
For episode = 1: E do
  Initialize the random operation state of the distribution network O_0
  For t = 1: T do
    The action of each agent is determined by its prevailing observation: Equation 20
    Performing joint actions: Equation 28, the new state is obtained by power flow calculation O_{t+1}
    Calculate each agent's reward r_t: Equation 25-27
    Store experiences: (s_t, a_t, r_t, s_{t+1}) to $\mathcal{D}$
    IF experience sample is sufficient then
      Mini-batch samples are randomly sampled from $\mathcal{D}$
      For each agent i do
        Calculate the target Q-value: $y_i = r_i + \gamma Q_i^{\pi'}(o_1', \ldots, o_K', \pi_1'(o_1'), \ldots, \pi_K'(o_K'))$
        Update the critic network: Equation 31
        Update the actor network: Equation 32
        Soft update target network: Equation 33
      End for
    End if
  End for
End for

Algorithm 1. Multi-agent peak shaving training algorithm based on MADDPG.

3.3 Implementation of the hierarchical cooperative control strategy

This section presents the specific implementation mechanism of the collaborative peak shaving strategy for PV clusters, based on the multi-agent environment and the MADDPG algorithm framework established in the previous section. By leveraging a hierarchical decision-making architecture and a real-time feedback mechanism, the strategy achieves a closed-loop control system that integrates global optimization with local execution. The detailed implementation steps of the proposed approach are illustrated in Figure 6:

Step 1: Power allocation among clusters.

At the system level, the power allocation among clusters is implicitly learned by the MADDPG algorithm in offline training and solidified into the policy of each agent. During the training process, the Critic network with global information guides each Actor to understand the impact of its own actions on the system as a whole and the potential reactions of other clusters. This enables each cluster to autonomously determine its share of power regulation when a system load disturbance is observed, based on its electrical distance from the disturbance node, its own peak shaving capacity, and the operating status of other clusters. The clusters with close electrical distance and strong adjustment ability will take the initiative to undertake more peak shaving tasks, otherwise, it will provide auxiliary support, so as to achieve global near-optimal power allocation without online communication and negotiation.

Step 2: Coordination of units within a cluster.

Inside the cluster, the refinement and execution of power commands are also made by the trained Actor network to achieve

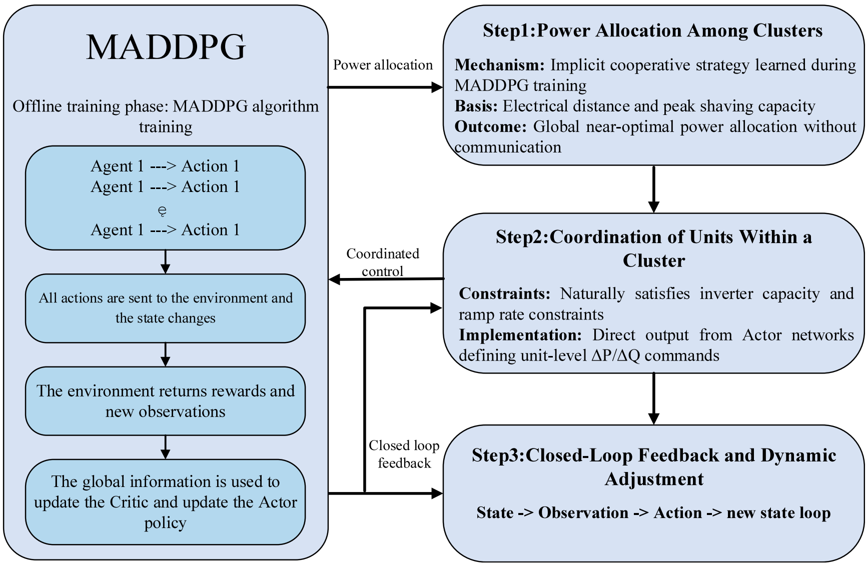

FIGURE 6
Flowchart of layered coordination control for distributed PV clusters based on multi-objective partitioning.

TABLE 1 Parameters of PV systems.

PV number	Access node	Standard unitary maximum output	Actual maximum capacity
PV1	2	0.7634	763.4 kW
PV2	5	0.7278	727.8 kW
PV3	10	0.4816	481.6 kW
PV4	13	0.5524	552.4 kW
PV5	16	0.4613	461.3 kW
PV6	19	0.5554	555.4 kW
PV7	23	0.7146	714.6 kW
PV8	24	0.7200	720.0 kW
PV9	26	0.5352	535.2 kW
PV10	27	0.7557	755.7 kW
PV11	31	0.7807	780.7 kW

distributed autonomous decision-making. After the cluster agent determines the total power adjustment amount of its cluster according to its observation space, the action space output by its Actor network directly defines the active and reactive power adjustment instructions of each PV unit in the cluster. The network has learned in training how to decompose the total task into each unit according to the real-time operation point of each unit, proportion-ally or according to the principle of optimal efficiency, under the premise of meeting the inverter capacity constraints and hill climbing rate constraints.

Step 3: Closed-loop feedback and dynamic adjustment.

The execution results of each cluster will change the operation state of the distribution network, and these changes will be immediately reflected in the local observation of each agent at the next time. The agent then makes a new round of decisions, resulting in a continuous rolling optimization process. The dynamic adjustment mechanism based on real-time feedback makes the strategy automatically adapt to the uncertain factors such as light fluctuation and load change, and continuously track the optimal operating point of the system, thus ensuring the accuracy and robustness of the peak shaving effect.

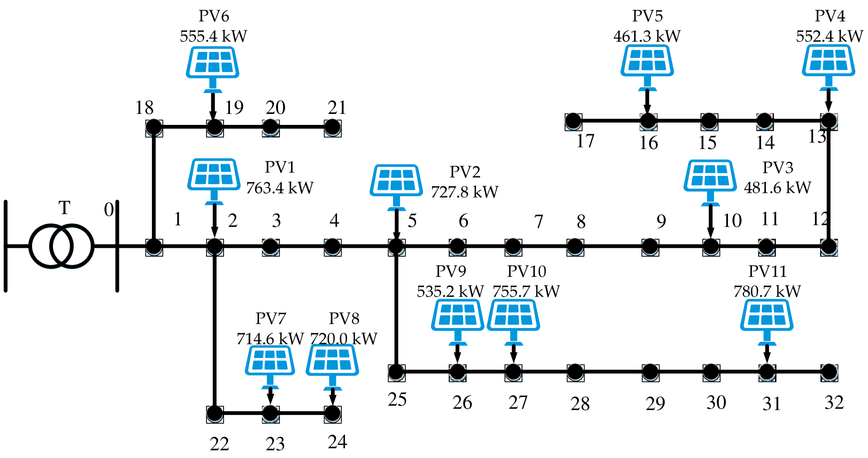

FIGURE 7
Topological diagram of IEEE 33-node system.

TABLE 2 Clustering results of PV systems.

Cluster number	Node number
1	1, 18, 19, 20, 21
2	5, 6, 7, 25, 26, 27, 28, 29, 30, 31, 32
3	2, 3, 4, 22, 23, 24
4	8, 9, 10, 11, 12, 13, 14, 15, 16, 17

4 Simulation validation and case study analysis

4.1 Experimental setup and test system configuration

To comprehensively evaluate the effectiveness and robustness of the proposed multi-objective partitioning and cooperative peak shaving strategy for PV clusters, simulations were conducted on the IEEE-33 node system (Gao et al., 2017). This system is a well-established benchmark in power distribution network studies, characterized by a radial topology consisting of 33 nodes and 32 branches. The IEEE-33 node system is particularly suited for this study due to its moderate number of nodes, clear structure, and flexibility in simulating various operational scenarios, including load fluctuations and high penetration of distributed PV systems.

In this study, a high penetration scenario was constructed by integrating 11 distributed PV systems into the IEEE-33 node system. The specific access points for these PV systems are detailed in Table 1, and the topological structure of the system is shown in Figure 7. The selection of these nodes was based on practical considerations, such as ensuring a diverse distribution of PV systems across the network to simulate real-world conditions effectively. The installed capacities of these PV systems range from 46.13 kW to 78.07 kW, reflecting a realistic distribution of PV installations in modern distribution networks.

The simulations were performed on a hardware platform equipped with an NVIDIA GTX 4050Ti GPU and an Intel Core i9-13900 KF CPU, ensuring sufficient computational power for the complex MARL algorithms employed in this study. The software environment consisted of the Windows 10 operating system, with all algorithms implemented using the PyTorch neural network framework and Python 3.10.0.

To simulate real-world load disturbances, a random disturbance factor was introduced into the IEEE-33 node system. The specific disturbance values for each hour of the day are detailed in Table 4. These disturbances were designed to mimic the random fluctuations in load that are commonly observed in power systems, providing a realistic testbed for evaluating the proposed peak shaving strategy.

By carefully configuring the test system and simulation environment, this study ensures that the results are both scientifically rigorous and practically relevant, providing valuable insights into the application of multi-objective partitioning and cooperative peak shaving strategies in modern distribution networks with high penetration of distributed PV systems.

4.2 PV cluster partitioning results and analysis

Following the determination of the optimal cluster number ($K = 4$) via the elbow method, the improved K-means algorithm, guided by the comprehensive multi-objective optimization model, was leveraged to partition the distributed PV system. This advanced clustering approach moves beyond conventional methods by explicitly optimizing the defined metrics-APSM, PSS, and NCD-to ensure the resulting clusters are not only geographically and electrically coherent but also functionally prepared for cooperative control. The final partitioning scheme, which effectively delineates the network into four distinct clusters with balanced characteristics, is quantitatively summarized in Table 2. More intuitively, the spatial distribution of these clusters within the network topology is

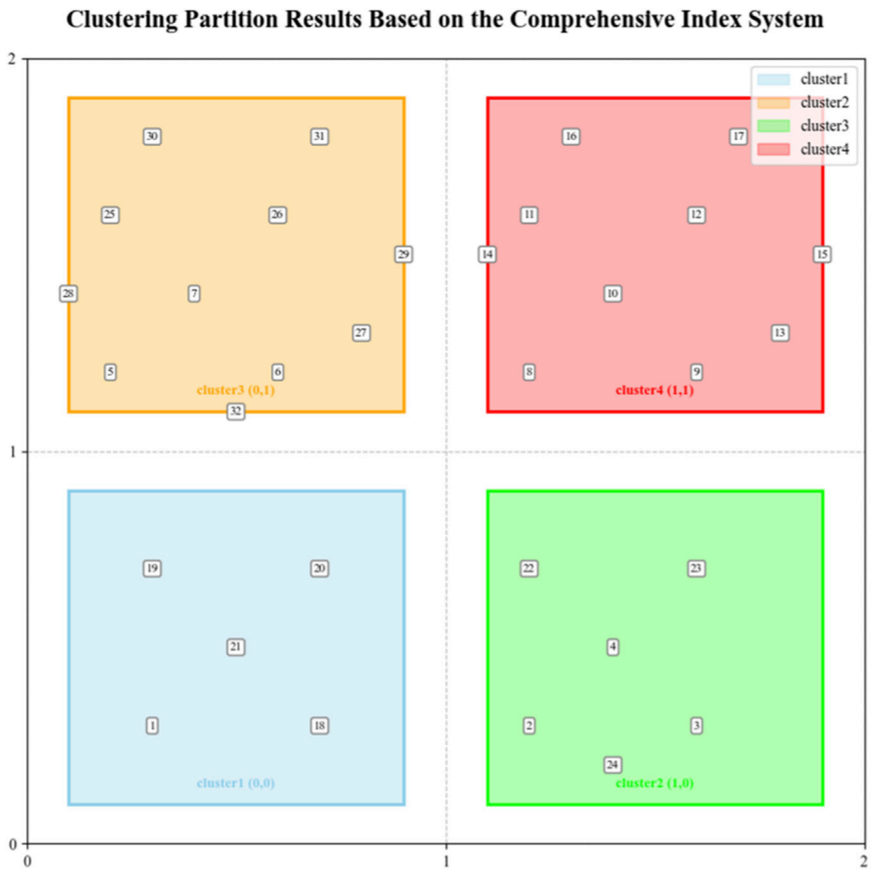

FIGURE 8
The clustering partition results are carried out according to the selected clustering index.

visually presented in Figure 8, clearly illustrating the physical grouping of nodes.

This rational partition provides a critical structural foundation for the subsequent MARL framework, effectively defining the autonomous decision-making agents for the cooperative peak-shaving task. The following analysis delves into the specific composition and inherent properties of each cluster.

The results show that the distributed PV nodes in the IEEE-33 node system are clustered into four clusters with significant physical significance by the improved K-means clustering algorithm. Finally, four types of clusters dominated by core indicators such as peak shaving capacity and response speed were formed. Finally, the system clustering diagram is obtained according to the final layer division results, as shown in Figure 9.

Based on the improved K-means clustering algorithm, combined with three core indicators of APSM, PSS, and NCD, the optimal number of clusters is determined to be 4, and the cluster division of PV nodes in IEEE-33 node system is completed. The results of PV cluster division will provide a clear control unit for MADDPG algorithm to construct a multi-agent environment, and will also provide key prior information for the Critic network to learn the interaction mode between clusters and the Actor network to generate the global optimal collaborative peak shaving strategy, thereby ensuring the effectiveness of algorithm training and peak shaving decisions.

To validate the physical significance of the partitioning results, a quantitative analysis of the three proposed metrics was conducted as shown in Figure 10. Regarding regulation capacity, a 40% safety margin was applied to the APSM to ensure robustness against PV output uncertainties. Under this condition, Cluster 2 retains the highest effective capacity of 1.12 MW, establishing it as a capacity-oriented unit for bulk peak shaving. In terms of PSS, Cluster 3 achieves the highest coefficient of 0.68, indicating a strong temporal alignment between source and load. Consequently, Cluster 3 should be prioritized for active power regulation due to its high self-balancing potential. Conversely, Cluster 4 exhibits a negative synergy of −0.04, reflecting a typical residential load pattern that is asynchronous with PV generation. Finally, regarding structural topology, Figure 10c shows that Cluster 4 possesses a significantly elevated NCD of 3.83, compared to the average of 1.6 in other clusters. This high NCD characterizes Cluster 4 as a voltage-sensitive cluster located at the feeder end-zone where electrical coupling is strong. Therefore, its control agent naturally focuses on voltage stability maintenance rather than large-scale power flow regulation.

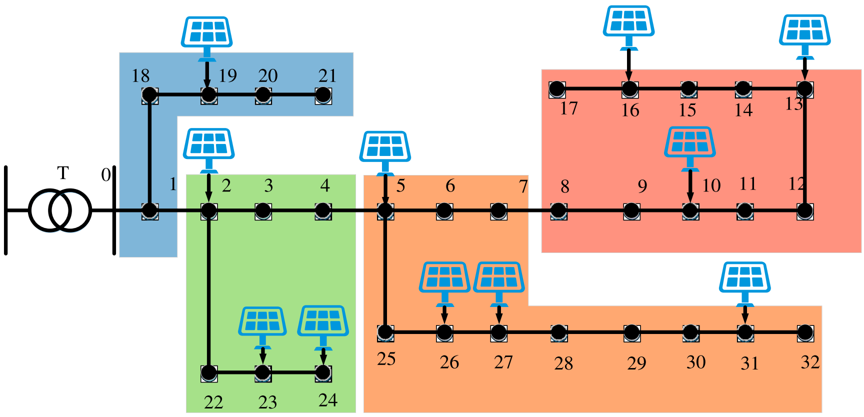

FIGURE 9
Topological mapping of multi-objective clustering results on the IEEE-33 system.

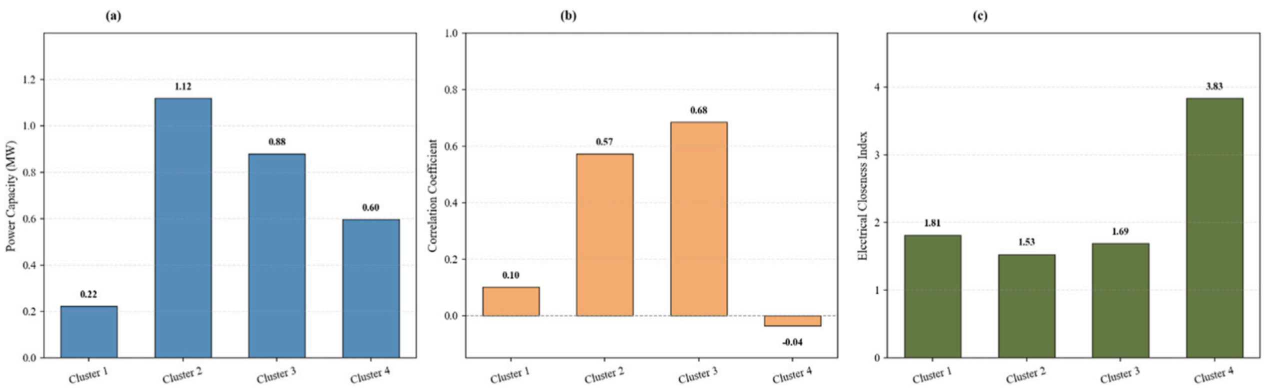

FIGURE 10
Cluster characteristics profiling based on proposed metrics. **(a)** Available Peak Shaving Margin(APSM). **(b)** Peak Shaving Synergy (PSS). **(c)** Node Correlation Degree (NCD).

It is important to note that the clustering stage serves primarily as a structural decomposition strategy to reduce the dimensionality of the control problem. The system's overall robustness is fundamentally guaranteed by the MADDPG agents. Trained under stochastic environments, these agents act as an adaptive control layer capable of compensating for minor topological variations resulting from clustering parameter sensitivities. This hierarchical design ensures that the peak-shaving performance is not overly sensitive to the precise tuning of clustering weights, provided the partitions maintain basic electrical coherence.

4.3 Experimental setup and test system configuration

To validate the effectiveness of the proposed cooperative peak-shaving strategy for PV clusters, the MADDPG algorithm was employed to train autonomous agents based on the clustering results obtained in Section 4.2. Within this framework, each cluster defined by the multi-objective partitioning process was modeled as an independent agent. Guided by centralized training, these agents performed decentralized decision-making to optimize both inter-cluster power allocation and intra-cluster unit coordination. Table 3 presents the hyperparameter settings of the MADDPG algorithm, which were carefully tuned to ensure robust performance. Specifically, the discount factor γ = 0.98 was selected to balance immediate and future rewards, allowing decision impacts to reasonably extend across the 24-h cycle while avoiding excessive emphasis on distant states that could hinder convergence. The hidden layer dimension h = 256 was chosen to provide sufficient network capacity while preventing overfitting caused by overly large dimensions, thereby achieving a balance between representational power and training efficiency.

In order to verify the effectiveness of the distributed cooperative peak shaving control method based on multi-agent proposed in this paper and reflect the random fluctuation characteristics of load in the power system, this paper introduces a random disturbance factor into the IEEE-33

TABLE 3 Parameter settings of the proposed method.

Parameters	Parameter values
Actor network learning rate	0.005
Critic network learning rate	0.01
Soft update coefficient	0.005
Initial exploration noise	0.5
Minimum exploration noise	0.05
Explore the noise attenuation coefficient	0.998
Experience pool capacity	50,000
Discount factor	0.98
Hidden layer dimension	256
Total number of training rounds	200
Early stop patience value	30

node system connection to simulate the peak shaving scene of the actual power grid, and the specific disturbance values are shown in Table 4.

Figure 11 shows the reward value of the agent during the training process, the blue curve in the upper left subfigure is the total reward of each training round, and the red curve is the average reward of the "10 rounds moving average". It can be seen that with the increase of the number of training rounds, the total reward shows an overall upward trend, while the "10-round moving average" curve gradually flattens out at 120 training rounds, indicating that each node has learned the optimal strategy to reduce the network loss.

The top right subfigure shows the contribution of each reward component such as balance reward, regulation reward, and voltage reward in the last round of training, indicating that the agent achieved the collaborative optimization of multi-objective in the final policy. The left subgraph quantifies the actual performance of the last round strategy from the dimensions of balance deviation, total regulation, and voltage deviation, which corroborates the metric optimization logic behind the average return improvement. The right subfigure "Exploration noise attenuation curve" presents the attenuation process of exploration noise with training rounds, reflecting the strategy transition of the algorithm from "exploration oriented" to "exploitation oriented".

Figures 12, 13 show the 24-h output of the 11 connected photovoltaics in the given example. The output of distributed PV included in the system is 0 at 0–5 and 20–24, and increases with the increase of light intensity at 5–20.

Figures 14, 15 illustrate the adjusted power output profiles for individual PV units and aggregated clusters, respectively, following the MADDPG optimization under simulated load fluctuations. As shown in Figure 14, the output of each specific PV unit is dynamically regulated based on the decision-making of its corresponding cluster agent. Furthermore, Figure 15 aggregates these responses, demonstrating how the collective adjustment of PV clusters aligns with the system's peak-shaving requirements.

To further illustrate the collaborative behaviors of PV clusters, a typical operating day is selected from the simulation results for detailed analysis. Figure 16 shows the Controlled Power Output Ratio of each PV cluster during the peak shaving period, while Figure 17 presents the Overall Adjustment Ratio of each cluster. Together, these two figures demonstrate the coordinated decision-making mechanism and differentiated functional roles within the multi-agent system, achieved without direct communication.

Figure 16 depicts the temporal evolution of the collaboration degree among the four PV clusters over a typical operating day. During non-generation periods (01:00–05:00 and 21:00–24:00), the PV systems are inactive and do not participate in peak shaving. From 06:00 to 11:00, Cluster 3 consistently maintains the highest synergy level at approximately 0.44, followed by Clusters 2 and 1, indicating that Cluster 3 assumes the dominant regulatory role during this period. As the system transitions into the high-load period (12:00–20:00), the collaboration degrees exhibit significant fluctuations, reflecting active but unevenly distributed cooperative regulation. A notable shift occurs between 19:00 and 20:00, when Cluster 1's synergy degree surges to 0.74 while Cluster 3's drops sharply to 0.03, marking Cluster 1 as the primary regulator during this critical interval. Complementing the trends in Figures 16, 17 also provides the power adjustment ratios of each cluster relative to their baseline output. During the peak shaving phase, Clusters 3, 2, and 1 predominantly contribute by implementing power curtailment to maintain grid balance. In contrast, during the valley filling phase, distinct complementary behaviors emerge: Cluster 4 takes the leading role through downward regulation, while the other three clusters perform positive adjustments by increasing power injection.

The red curve in Figure 18 is the "before peak shunted" curve, showing significant fluctuation characteristics, which intuitively reflects the impact effect of load disturbance on the balance node power without PV regulation. The fluctuation of the blue curve is the "after peak adjustment" curve, which is significantly narrower, reflecting the regulation effect of the algorithm on the power balance.

The simulation results quantitatively demonstrate the performance of the pro-posed MADDPG-based control strategy in mitigating power fluctuations. The application of the strategy led to a reduction in the peak-to-valley difference from 3.3174 MW to 2.3423 MW, corresponding to an improvement of 29.39%. Concurrently, an average enhancement of 59.68% in power balance deviation was achieved. These quantitative outcomes confirm the strategy's capability to flatten the net load curve and maintain active power balance within the distribution network, effectively alleviating the impact of load disturbances under high PV penetration.

The robustness of the proposed framework is fundamentally underpinned by the decentralized coordination mechanism established through the MARL framework. Relying on the centralized-training-decentralized-execution paradigm, MADDPG agents are trained in an environment characterized by stochastic

TABLE 4 24-h values of load disturbances.

Time/hour	Load disturbance value (MW)	Time/Hour	Load disturbance value (MW)
1	0.575	13	0.950
2	0.550	14	0.850
3	0.508	15	0.850
4	0.499	16	0.870
5	0.508	17	0.930
6	0.570	18	0.990
7	0.680	19	1.010
8	0.950	20	0.990
9	1.010	21	0.840
10	1.040	22	0.800
11	1.020	23	0.800
12	1.000	24	0.700

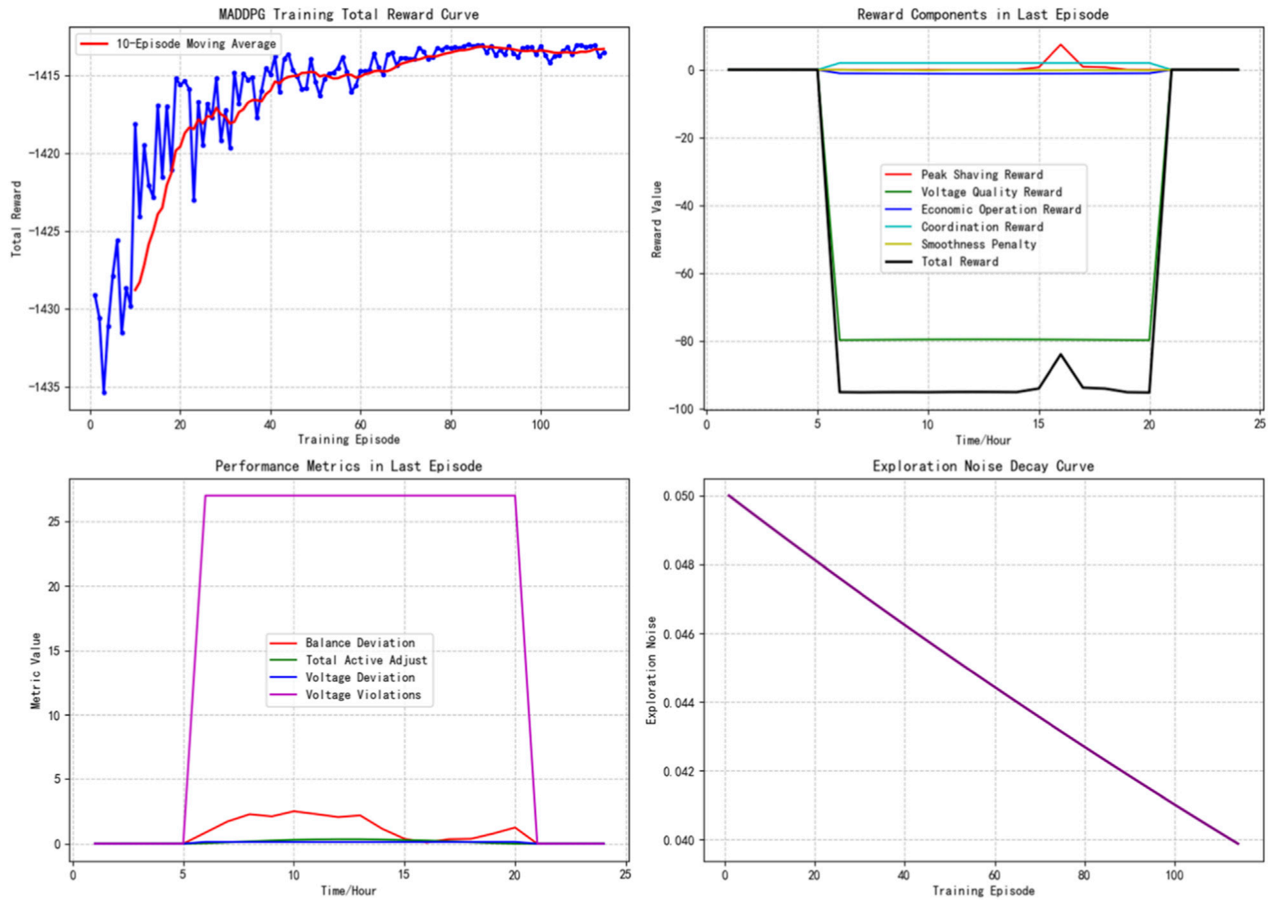

FIGURE 11
Average reward curve of agent training.

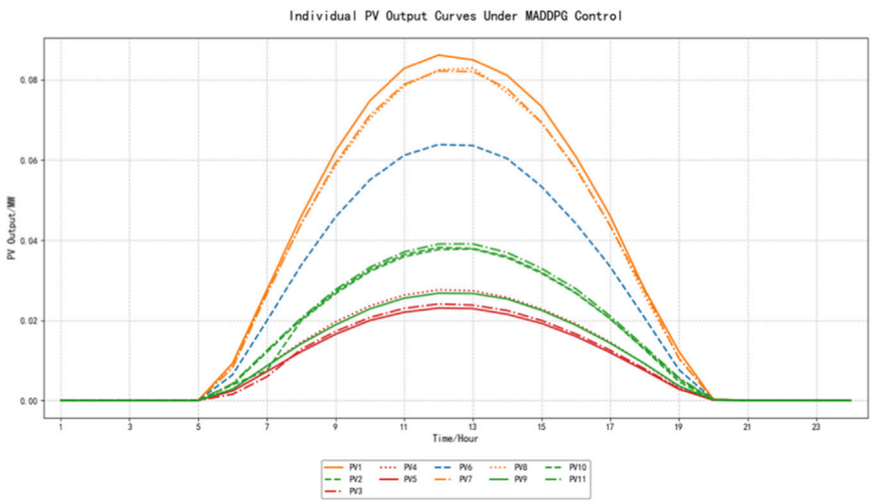

FIGURE 12
Active power output curves of individual PV units.

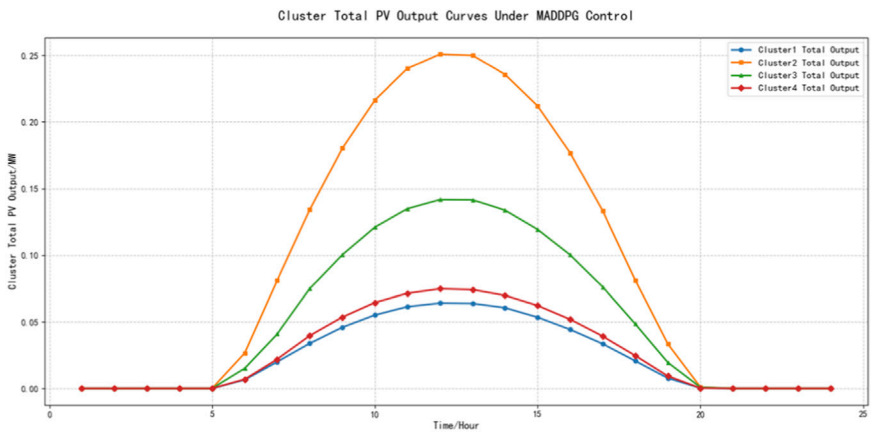

FIGURE 13
Aggregated active power output curves for each PV cluster.

load fluctuations and PV output uncertainties. This learning capability enables agents to develop cooperative strategies that are resilient to structural variations. Consequently, even if clustering parameters vary-leading to slight shifts in cluster boundaries-the agents can adaptively learn to adjust their policies to compensate for these changes during the training process. This mechanism allows the system to execute decisions based solely on local observations without real-time inter-cluster communication, ensuring that peak shaving performance remains robust and applicable to complex real-world scenarios regardless of precise parameter tuning.

5 Conclusion

This study has successfully addressed the critical challenge of integrating high-penetration PV systems into the grid by introducing a novel, integrated framework that synergizes multi-objective cluster partitioning with hierarchical MARL for cooperative peak shaving. The proposed methodology demonstrates a significant advancement in leveraging the inherent flexibility of distributed PV clusters to enhance grid stability and power quality. The key conclusions derived from this work are summarized as follows.

1. The multi-objective partitioning model, built upon the comprehensive indices of APSM, PSS, and NCD e, proves to be a robust foundation for the entire framework. By optimizing these metrics, the improved clustering algorithm effectively transitions the system from a collection of individual, disparate PV units into a set of functionally coherent and structurally sound clusters. This initial step is crucial, as it ensures that the formed clusters possess the inherent capability and synergistic potential necessary for effective coordinated control, thereby

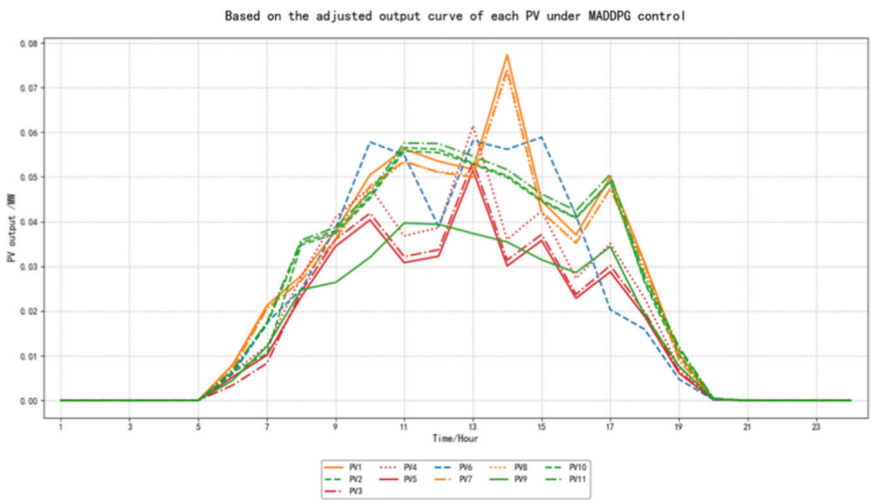

FIGURE 14
The adjusted output curve of each PV.

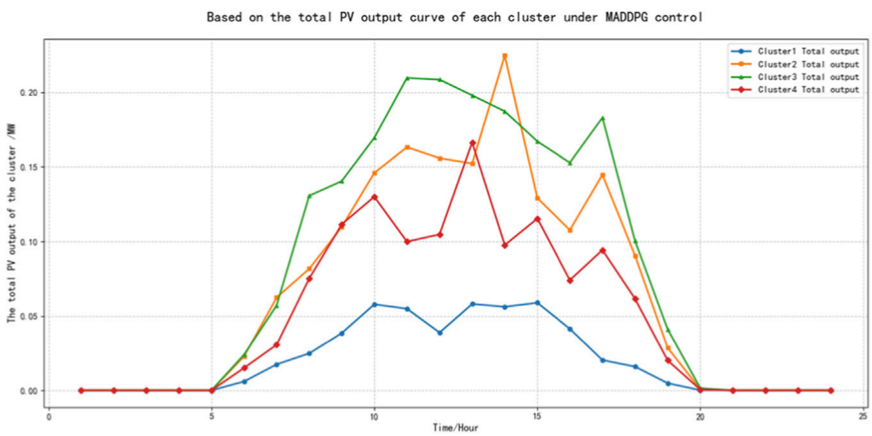

FIGURE 15
The adjusted output curves of each PV cluster.

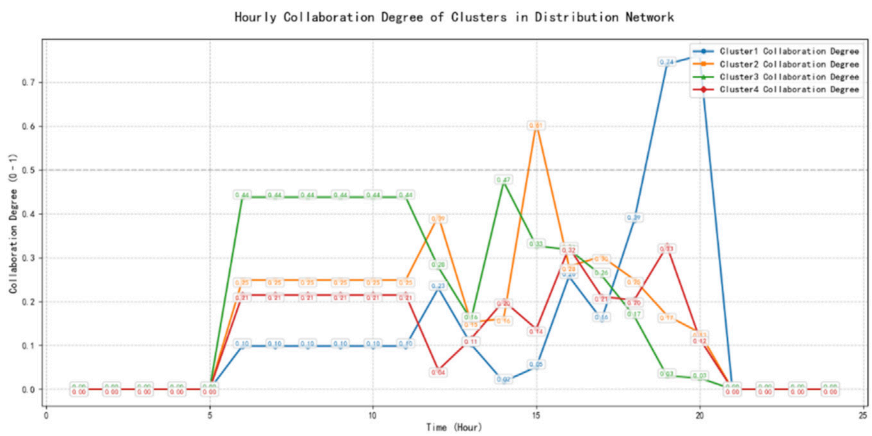

FIGURE 16
Cluster capacity contribution ratio.

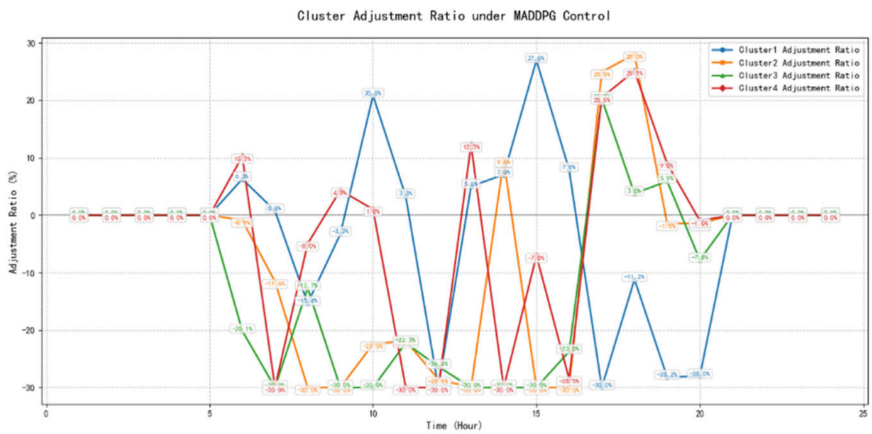

FIGURE 17
The cluster adjusted the proportion as a whole after MADDPG optimization.

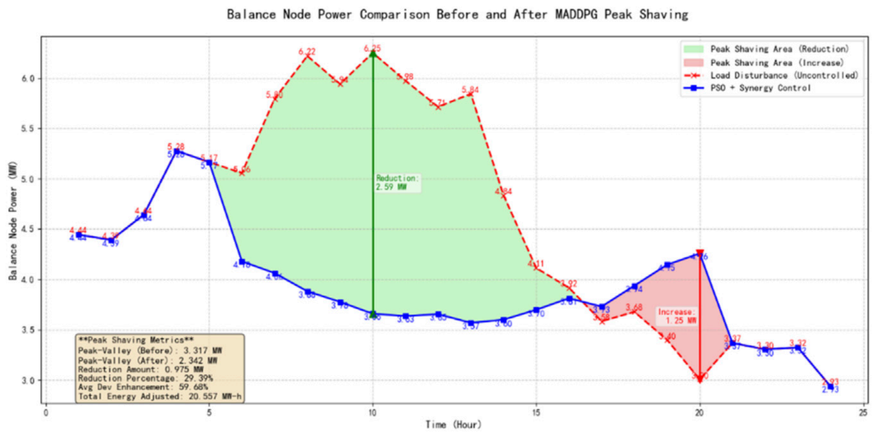

FIGURE 18
Peak shaving performance at the balance node under MADDPG control.

directly addressing the limitations of traditional clustering methods that lack explicit peak-shaving objectives.

2. The hierarchical MARL framework, centered on the MADDPG algorithm, effectively unlocks the cooperative potential of the partitioned PV clusters. The "centralized-training-with-decentralized-execution" paradigm enables each cluster agent to learn sophisticated policies that achieve near-optimal power allocation among clusters and precise power regulation within a cluster, relying solely on local observations. This capability was vividly demonstrated by the highly coordinated adjustment patterns of the different clusters, showcasing how the multi-agent system autonomously balances peak-shaving responsibilities based on real-time grid conditions without the need for continuous communication.
3. The simulation results on the IEEE-33 node system provide compelling evidence of the strategy's effectiveness. The 29.39% improvement in the peak-to-valley difference and the 54.44% enhancement in the average power balance deviation confirm that the proposed strategy significantly smooths the net load curve and mitigates power fluctuations at the balance node. These quantitative improvements, along with the maintained voltage stability, validate that the framework successfully meets its dual objectives of ensuring power balance and enhancing voltage quality.

In conclusion, this work presents a scalable and effective solution for managing the complexities of future distribution networks with high-penetration renewable energy. The integrated "clustering-partitioning and multi-agent control" framework reduces the computational and control complexity of distributed PV systems while demonstrating strong adaptability and robustness. However, for the transition from simulation to engineering application, practical deployment constraints must be carefully considered. Specifically, the stochastic nature of communication networks, particularly data transmission latency, may impact the synchronization stability of multi-agent coordination. Furthermore, the economic feasibility regarding the hardware costs of deploying edge computing nodes capable of running reinforcement learning algorithms needs to be balanced against operational gains. Future research will focus on extending the framework to incorporate heterogeneous resources like energy storage and electric vehicles, testing its performance under more extreme weather conditions, and investigating its practical deployment in real-world grid operations.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JN: Writing – original draft, Conceptualization, Investigation, Formal Analysis. CX: Funding acquisition, Methodology, Supervision, Project administration, Writing – review and editing. KF: Software, Data curation, Writing – review and editing, Validation, Investigation. PH: Writing – review and editing, Resources, Formal Analysis, Visualization. YX: Project administration, Resources, Writing – review and editing, Funding acquisition. JL: Validation, Investigation, Data curation, Writing – review and editing. NW: Methodology, Writing – review and editing, Software.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work is supported by the Science and Technology Project of State Grid Jiangsu Electric Power Company (J2024180).

Conflict of interest

Authors JN, CX, and KF were employed by Wuxi Power Supply Company of State Grid Jiangsu Electric Power Co., Ltd. Authors PH, JL, and NW were employed by Jiangsu Frontier Electric Technology Co., Ltd. Author YX was employed by State Grid Jiangsu Electric Power Co., Ltd.

The authors acknowledge that this work was supported by State Grid Jiangsu Electric Power Company. The funder participated in the overall design of the research. The funder did not participate in data collection, analysis, or interpretation, nor did they contribute to the writing of this paper or the decision to submit it for publication.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

References

Aleksandrov, A., Chekichev, A., Parushev, A., Shehova, D., and Popov, R. (2025). "Analysis of energy storage technologies," in Proceedings of the 2025 19th Conference on Electrical Machines, Drives and Power Systems (ELMA) (Sofia, Bulgaria: Energy/Applied Energy/Renewable Energy/Solar Energy), 1–5. doi:10.1109/ELMA65795.2025.11083439

Cai, X., Chen, H., and Ding, T. (2020). A network partition-based method for voltage control in distribution networks with distributed PV. Proceedings of the 2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2). Wuhan, China: IEEE. 2156–2161. doi:10.1109/EI250167.2020.9346833

Cao, D., Hu, W., Zhao, J., Huang, Q., Chen, Z., and Blaabjerg, F. (2020). A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters. *IEEE Trans. Power Syst.* 35, 4120–4123. doi:10.1109/TPWRS.2020.3000652

Cao, D., Zhao, J., Hu, W., Yu, N., Ding, F., Huang, Q., et al. (2022). Deep reinforcement learning enabled physical-model-free two-timescale voltage control method for active distribution systems. *IEEE Trans. Smart Grid* 13, 149–165. doi:10.1109/TSG.2021.3113085

Gao, Y., Hu, X., Yang, W., Liang, H., and Li, P. (2017). Multi-objective bilevel coordinated planning of distributed generation and distribution network frame based on mult scenario technique considering timing characteristics. *IEEE Trans. Sustain. Energy* 8, 1415–1429. doi:10.1109/TSTE.2017.2713053

Guo, Y., Zhang, Q., and Wang, Z. (2021). Cooperative peak shaving and voltage regulation in unbalanced distribution feeders. *IEEE Trans. Power Syst.* 36, 5235–5244. doi:10.1109/TPWRS.2021.3069781

He, A., Pan, H., Dai, Y., Si, X., Yuen, C., and Zhang, Y. (2025). ADMM for Mobile edge intelligence: a survey. *IEEE Commun. Surv. Tutor.* 27, 3020–3057. doi:10.1109/COMST.2024.3521024

Hu, D., Ye, Z., Gao, Y., Ye, Z., Peng, Y., and Yu, N. (2022). Multi-agent deep reinforcement learning for voltage control with coordinated active and reactive power optimization. *IEEE Trans. Smart Grid* 13, 4873–4886. doi:10.1109/TSG.2022.3185975

IRENA (2025). *Renewable capacity statistics 2025*. Available online at: https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2025/Mar/IRENA_DAT_RE_Capacity_Statistics_2025.pdf (Accessed November 12, 2025).

Jiao, F., Li, Z., Ai, J., Yang, H., Deng, Y., Li, D., et al. (2024). Topology identification method for low-voltage distribution node networks based on density clustering using smart meter real-time measurement data. *IEEE Access* 12, 83600–83610. doi:10.1109/ACCESS.2024.3407827

Kumari, A., Kumar, A., and Kumar, P. (2025). A new modularity metric for disjoint and overlapping community structure evaluation in weighted complex networks. *Int. J. Data Sci. Anal.* 20, 5653–5677. doi:10.1007/s41060-025-00791-2.2

Li, S., Liu, F., Jiao, L., Chen, P., and Li, L. (2024). Self-supervised self-organizing clustering network: a novel unsupervised representation learning method. *IEEE Trans. Neural Netw. Learn. Syst.* 35, 1857–1871. doi:10.1109/TNNLS.2022.3185638

Li, X., Du, Y., Tan, Z., Xing, D., and Tan, C. (2025). Optimal dispatch and cost allocation model for combined peak shaving of source-load-storage under high percentage penetration of renewable energy. *Renew. Energy* 255, 123845. doi:10.1016/j.renene.2025.123845

Miraftabzadeh, S. M., Colombo, C. G., Longo, M., and Foiadelli, F. (2023). *K*-Means and alternative clustering methods in modern power systems. *IEEE Access* 11, 119596–119633. doi:10.1109/ACCESS.2023.3327640

Shen, J.-J., Shen, Q.-Q., Wang, S., Lu, J.-Y., and Meng, Q.-X. (2019). Generation scheduling of a hydrothermal system considering multiple provincial peak-shaving demands. *IEEE Access* 7, 46225–46239. doi:10.1109/ACCESS.2019.2906831

Shi, Y., Zhang, L., Yang, Y., Li, Q., and Zhang, H. (2025). A review on the short-term strategy for reducing the peak-valley difference and the long-term energy structure optimization strategy in cities based on the integration of "power generation - power grid - power load - energy storage". *Front. Energy Res.* 13, 1538811. doi:10.3389/fenrg.2025.1538811

Song, B., Qin, K., Wen, M., Zhu, M., Zou, K., and He, G. (2025). Coordinated multi-level scheduling method considering uncertainty of renewable energy and load. *Front. Energy Res.* 13, 1612065. doi:10.3389/fenrg.2025.1612065

Sun, X., Qiu, J., Yi, Y., and Tao, Y. (2022). Cost-effective coordinated voltage control in active distribution networks with photovoltaics and Mobile energy storage systems. *IEEE Trans. Sustain. Energy* 13, 501–513. doi:10.1109/TSTE.2021.3118404

Wen, X., He, M., Zhou, K., Fan, Q., Ye, H., and Li, B. (2025). "Design and implementation of intelligent control device for distributed photovoltaic clusters," in Proceedings of the 2025 International Conference on Smart Grid and Sustainable Energy (SGSE) (China: Zhangye), 267–270. doi:10.1109/SGSE66281.2025.11200380

Yin, Z., Zhou, Z., Yu, F., Gao, P., Ni, S., and Li, H. (2024). A cloud–edge collaborative multi-timescale scheduling strategy for peak regulation and renewable energy integration in distributed multi-energy systems. *Energies* 17, 3764. doi:10.3390/en17153764

Zare Ghaleh Seyyedi, A., Gitizadeh, M., Fakhrooeian, M., and Jabareh Nasero, M. (2025). Achieving the goals of energy arbitrage, peak-shaving, and PV self-consumption using PV-BTM BESS microgrids coupled with a distribution network. *J. Energy Storage* 112, 115479. doi:10.1016/j.est.2025.115479

Zhou, L., Wang, Y., Liu, N., Li, X., and Zhang, K. (2025). Energy collaborative optimization strategy for virtual power plants participating in green electricity trading and peak regulation ancillary services. *Int. J. Electr. Power Energy Syst.* 172, 111284. doi:10.1016/j.ijepes.2025.111284