

Article

Distributed Photovoltaic–Storage Hierarchical Aggregation Method Based on Multi-Source Multi-Scale Data Fusion

Shaobo Yang ^{1,*}, Xuekai Hu ¹, Lei Wang ¹, Guanghui Sun ¹, Min Shi ¹, Zhengji Meng ¹, Zifan Li ¹, Zengze Tu ² and Jiapeng Li ² 

¹ State Grid Hebei Electric Power Company Electric Power Research Institute, Shijiazhuang 050021, China; dyy_huxk@he.sgcc.com.cn (X.H.); 15232102338@126.com (L.W.); 13653383368@139.com (G.S.); 18003218308@139.com (M.S.); mzhengji@126.com (Z.M.); lizifan_hebepri@163.com (Z.L.)

² Electrical Engineering College, Xi'an Jiaotong University, Xi'an 710049, China; tu159690542@stu.xjtu.edu.cn (Z.T.); jiapengli@xjtu.edu.cn (J.L.)

* Correspondence: dyy_yangsb@he.sgcc.com.cn

Abstract

Accurate model aggregation is pivotal for the efficient dispatch and control of massive distributed photovoltaic (PV) and energy storage (ES) resources. However, the lack of unified standards across equipment manufacturers results in inconsistent data formats and resolutions. Furthermore, external disturbances like noise and packet loss exacerbate the problem. The resulting data are massive, multi-source, and heterogeneous, which poses severe challenges to building effective aggregation models. To address these issues, this paper proposes a hierarchical aggregation method based on multi-source multi-scale data fusion. First, a Multi-source Multi-scale Decision Table (Ms-MsDT) model is constructed to establish a unified framework for the flexible storage and representation of heterogeneous PV-ES data. Subsequently, a two-stage fusion framework is developed, combining Information Gain (IG) for global coarse screening and Scale-based Trees (SbT) for local fine-grained selection. This approach achieves adaptive scale optimization, effectively balancing data volume reduction with high-fidelity feature preservation. Finally, a hierarchical aggregation mechanism is introduced, employing the Analytic Hierarchy Process (AHP) and a weight-guided improved K-Means algorithm to perform targeted clustering tailored to the specific control requirements of different voltage levels. Validation on an IEEE-33 node system demonstrates that the proposed method significantly improves data approximation precision and clustering compactness compared to conventional approaches.

Keywords: distributed PV-ES systems; data fusion; multi-scale information; scale-based tree; hierarchical aggregation; improved K-Means



Academic Editor: José Matas

Received: 22 December 2025

Revised: 12 January 2026

Accepted: 20 January 2026

Published: 21 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The advancement of the energy transition has led to widespread applications of PV and ES generation in power grids due to their ease of installation and low cost [1,2]. However, distributed PV and ES exhibit strong randomness and volatility, characterized by numerous and widely dispersed connection points. The high-penetration integration of these resources can easily cause a series of problems, such as reverse power flow, node voltage rise, and reduced grid regulation capability, posing significant challenges to grid operation and control [3,4]. Therefore, aggregation clustering and unified control of large-scale distributed PV and ES systems have gained increasing attention in recent years.

Common aggregation and equivalence methods can be categorized into physical equivalence methods (e.g., capacity weighting, interface fitting, and coherency-based equivalence) and data-driven clustering methods [5]. Physical equivalence methods [6–10] are primarily model-centric, designed to aggregate traditional power plants or wind farms into simplified equivalent models. A fundamental limitation is that they often neglect the electrical network topology and coupling between distributed units, potentially leading to inaccuracies in grid-level dynamic studies. More critically for modern DERs, they lack inherent mechanisms to process heterogeneous operational data streams. Data-driven clustering methods, such as K-Means [11,12], offer flexibility by aggregating units based on statistical similarity in measured data (e.g., output power profiles). This represents a shift towards data-centric analysis. However, their successful application typically rests on a critical assumption: the availability of clean, consistent, and uniformly structured input data. In real-world distributed PV-ES systems, this assumption is violated due to multi-operator management, varied communication protocols, and differing data resolution needs.

Existing research on distributed PV and ES aggregation typically assumes that information from all units has the same data structure and is fully available. Reference [13] proposed an aggregation feasible region modeling method based on Minkowski summation, achieving a standardized representation of the feasible region for distributed PV and ES clusters and simplifying calculations through k-order approximation models. Reference [14], focusing on heterogeneous resources like data center clusters, distributed PV, and ES, constructed a distributionally robust optimization model based on Wasserstein distance and designed a flexibility aggregation boundary identification method to characterize power regulation boundaries and enhance market response capability. Reference [15], starting from uncertainty set projection and mapping relationships, proposed a coordinated optimization framework to achieve synergistic operation between cluster resources and system control objectives, strengthening the adjustability and controllability of distributed PV and ES in complex market environments. In reality, due to diverse operators and communication methods, the multi-source heterogeneous data generated during the operation of distributed PV and ES (e.g., PV output, ES status, environmental monitoring, and grid dispatch) features a wide spatiotemporal distribution, significant structural differences, and strong dynamics. Raw data is multi-source, multi-scale, noisy, and often incomplete. Directly applying these clustering algorithms to such raw data leads to suboptimal or misleading clusters [16]. Furthermore, neither traditional equivalence nor standard data-driven clustering provides a structured way to align aggregation results with the hierarchical control objectives of different grid levels (e.g., prioritizing voltage support at the distribution level versus energy balancing at the transmission level) [17]. Therefore, it is necessary to hierarchically construct aggregation models for distributed PV and ES based on multi-source heterogeneous data fusion, tailored to different voltage levels and scenario requirements, to meet the diverse needs of distributed PV and ES systems for participating in user-side reliable power supply and grid-side regulation and control.

In summary, the aggregation of distributed PV and ES is of great practical importance for the unified dispatch and control of power systems. However, existing research rarely addresses the multi-source and multi-scale heterogeneous nature of operational data or the diverse control requirements across different network hierarchies. To overcome these limitations, this paper proposes a principled, domain-aware integration of established techniques to form a coherent pipeline. Specifically, we develop a two-stage multi-scale data fusion framework (IG + SbT) designed to handle data heterogeneity without requiring massive training data or creating black-box models, ensuring transparency essential for grid operations. Subsequently, we introduce a hierarchical aggregation method (AHP + K-Means++)

that explicitly incorporates domain knowledge through structured priority weighting, rather than relying on opaque, data-driven similarity kernels. This combination is specifically chosen to balance computational practicality, interpretability, and alignment with power system operational constraints—criteria often overlooked by purely data-centric machine learning approaches. The effectiveness and superiority of the proposed approach are validated through comprehensive case studies based on the IEEE-33 node test system, demonstrating not only algorithmic performance but also tangible utility in system-level control scenarios.

The overall flowchart of the proposed hierarchical aggregation method based on multi-source multi-scale data fusion is shown in Figure 1. The proposed methodology consists of two coherent stages. Stage I transforms raw, heterogeneous PV-ES data from multiple sources and scales into a high-quality, fused feature dataset through a standardized pipeline including data cleaning, a two-stage fusion strategy (IG + SbT), and a specialized mechanism for handling missing data. Stage II utilizes this dataset to perform hierarchical aggregation tailored to different voltage levels: it extracts three operational metrics, determines level-specific weighting via AHP, and executes targeted clustering using a weight-guided improved K-Means++ algorithm. The final output comprises distinct aggregated clusters for the 400 V station area, 10 kV distribution network, and 110 kV transmission network, each aligned with the specific control objectives of its respective level.

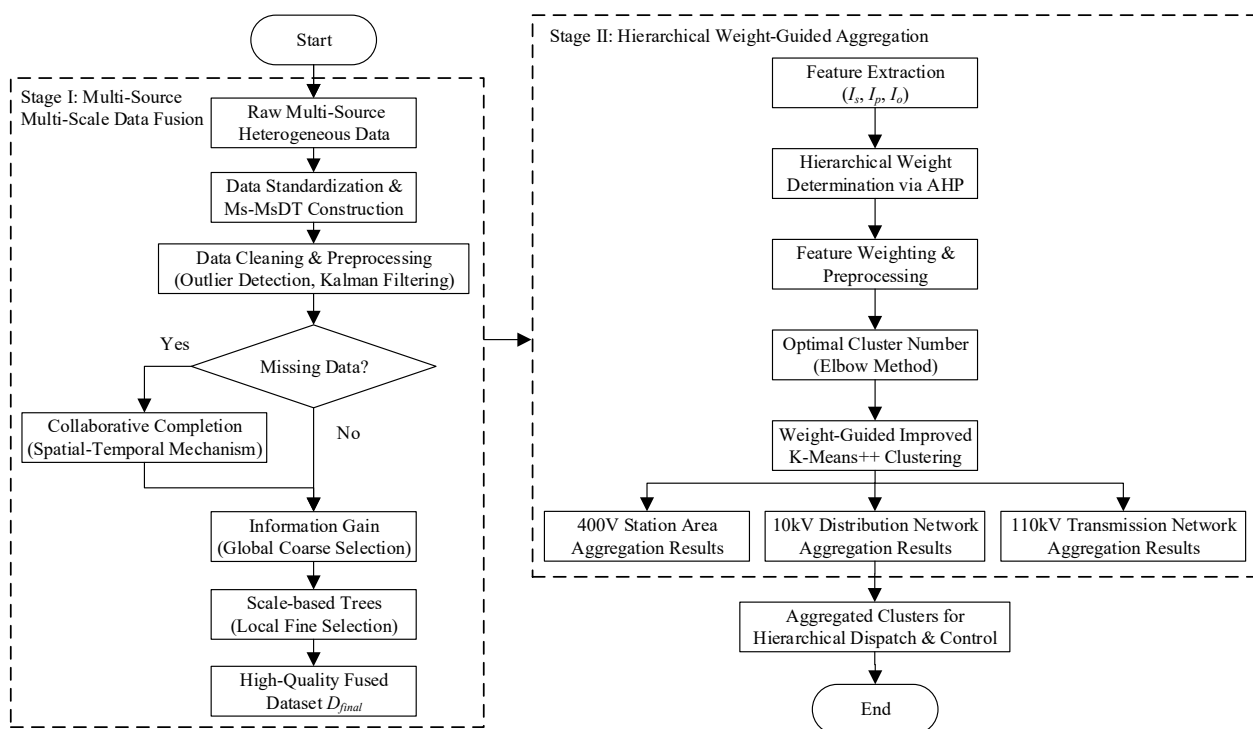


Figure 1. Overall flowchart of the proposed hierarchical aggregation method based on multi-source multi-scale data fusion.

2. Multi-Source Multi-Scale Heterogeneous Data Fusion

Operational data from distributed PV-ES systems exhibit typical multi-source, multi-scale, heterogeneous characteristics. Addressing this challenge requires a fusion approach that is both effective and suited to the practical constraints of power system applications, including limited labeled data, the need for interpretability, and the avoidance of computationally intensive training. Therefore, instead of employing complex, data-hungry models like deep learning, we construct a lightweight, explainable pipeline that transforms raw, chaotic data into high-quality, consistent, and information-rich feature datasets. This

pipeline lays a transparent foundation for subsequent aggregation, where every fusion decision (scale selection) can be traced and justified—a crucial feature for operational trust and regulatory compliance.

2.1. Data Standardization and Multi-Scale Data Structure Construction

In practical operation, distributed PV-ES data exhibit significant format heterogeneity due to differences in operators, equipment manufacturers, communication protocols, etc. As shown in Tables 1 and 2, first unify various raw data formats into timestamped structured time-series tables. More critically, recognize that each physical or operational attribute (e.g., active power, SOC) typically exists at multiple observational or statistical scales. For instance, output data from the same PV station may simultaneously include (1) second-level raw sampling values (reflecting instantaneous fluctuations); (2) 5 min average values (for short-term forecasting); (3) 15 min scheduled values (for market bidding); and (4) hourly statistical values (for performance evaluation). This multi-scale characteristic forms the basis for subsequent fusion optimization.

Table 1. Standardized PV data format (with scale marking).

Timestamp	PV Station ID	Active Power (MW)	Reactive Power (Mvar)	Voltage (kV)	Capacity (MW)	Base Voltage (kV)	Scale Mark
20 August 2023 12:00:00	PV001	1.5	0.1	230	2	220	L2 (5 min average)
...	...	...	...	...	...	...	...

Table 2. Standardized ES data format (with scale marking).

Timestamp	ES Station ID	Rated Power (kW)	Charge/Discharge Rate (C)	SOC (%)	Voltage (kV)	Capacity (kWh)	Base Voltage (kV)	Scale Mark
20 August 2023 12:00:00	SE001	500	0.5	65	225	1000	110	L1 (second-level)
...	...	...	...	...	...	...	...	...

Definition 1. *PV-ES Multi-scale Information Table (MsIT).* An MsIT can be formally represented as a triple:

$$\text{MsIT} = (O, A_S, F) \quad (1)$$

where

- $O = \{o_1, o_2, \dots, o_m\}$ is a finite non-empty set of objects, representing individual PV-ES units or monitoring points.
- $A_S = \{a_i^j | i = 1, 2, \dots, n; j = 1, 2, \dots, t_i\}$ is a multi-scale attribute set. a_i denotes the i -th attribute (e.g., “active power”, “SOC”), with superscript j indicating the j -th scale of that attribute. Scale indices are typically arranged from fine to coarse, i.e., $j = 1$ represents the finest granularity (e.g., second-level), $j = t_i$ indicates the coarsest granularity (e.g., daily-level).
- $F = \{f : O \times A_S \rightarrow V\}$ is an information function assigning an attribute value $f(o, a_i^j) \in V$ to each object $o \in O$ at each scale, where V is the value domain.

Definition 2. *PV-ES Multi-Source Multi-Scale Decision Table (Ms-MsDT) [18].* In practical systems, data typically originates from s different data sources (e.g., cloud platforms of different

operators, data collectors of different models). Each source may observe different attribute subsets and scale sets. We define the multi-scale data in this multi-source environment as

$$Ms\text{-}MsDT = \{MsDT_k | k = 1, 2, \dots, s\} \quad (2)$$

where data from each source can be represented as a quadruple:

$$MsDT_k = (O, A_{S_k} \cup \{d\}, F_k, G) \quad (3)$$

Here, d is a single, predefined categorical decision attribute that captures a key structural grouping in the system, such as ‘Operator ID’, $G = \{g : O \rightarrow V_d\}$ is the decision information function, and V_d is the decision value domain.

Ms-MsDT completely characterizes the complex structure of multi-source multi-scale PV-ES data and serves as the starting point for method processing. To realize both computational efficiency and interoperability, the Ms-MsDT is implemented as a hierarchical, indexed in-memory data structure.

The Ms-MsDT acts as a canonical data model, decoupling data ingestion from processing logic. The fusion and aggregation algorithms are programmed solely against the unified Ms-MsDT interface. To support a new data source (e.g., a new operator’s API format), only the specific data ingestion adapter needs to be implemented to populate the Ms-MsDT structure; the core analytical pipeline remains unchanged. This design ensures long-term interoperability and extensibility as new data sources and formats emerge in evolving power systems.

The Ms-MsDT model and the subsequent fusion framework inherently account for potential conflicts among heterogeneous data sources. Conflicts are addressed through a two-stage process: initial mitigation via information-gain-based scale selection within each source, followed by consistency-driven reconciliation across sources using the scale-tree search mechanism. Persistent inconsistencies are flagged for further analysis.

2.2. Data Cleaning and Preprocessing

This step is performed independently for each data source $MsDT_k$, aiming to eliminate obvious errors and noise, providing clean data for subsequent scale fusion.

2.2.1. Outlier Detection

Outlier identification based on physical constraints:

PV output, P_{pv} , must satisfy

$$0 \leq P_{pv} \leq P_{pv}^{max} \quad (4)$$

where P_{pv}^{max} is the maximum theoretical output of PV modules (considering irradiance, temperature, etc.).

The energy storage state of charge, SOC, must satisfy

$$SOC^{min} \leq SOC \leq SOC^{max} \quad (5)$$

where SOC^{min} and SOC^{max} are the safety lower and upper limits, specified by the manufacturer.

Data points violating these constraints are identified as outliers and marked as NaN.

2.2.2. Kalman Filter-Based Denoising

Considering sampling and transmission noise, Kalman filtering is applied for time-series data smoothing. We treat the finest scale sequence of each attribute as the underlying true state sequence. The state-space model is as follows:

State equation:

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{w}_k, \quad \mathbf{w}_k \sim \mathcal{N}(0, \mathbf{Q}) \quad (6)$$

where $\mathbf{x}_k$ is the state vector at time k (may contain multiple correlated attributes), $\mathbf{A}$ is the state transition matrix, $\mathbf{w}_k$ is process noise, and the covariance matrix $\mathbf{Q}$ reflects the intensity of system state changes.

Observation equation:

$$\mathbf{z}_k = \mathbf{H}\mathbf{x}_k + \mathbf{v}_k, \quad \mathbf{v}_k \sim \mathcal{N}(0, \mathbf{R}) \quad (7)$$

where $\mathbf{z}_k$ is the observation vector at time k (possibly from different scales), $\mathbf{H}$ is the observation matrix, and $\mathbf{v}_k$ is observation noise with a covariance matrix $\mathbf{R}$ determined by sensor accuracy.

Furthermore, in the multi-scale context, we model cross-scale change correlations by introducing cross-scale correlation parameters ρ_{ij}^{scale} into the process noise covariance matrix $\mathbf{Q}$. For example, if second-level data fluctuates sharply at a certain moment, the corresponding minute-level average should also reflect some change trend. This modeling enables the filtering process to collaboratively utilize multi-scale observation information, yielding more robust state estimates than single-scale filtering.

In this study, the state vector $\mathbf{x}_k$ represents the actual (denoised) values of key operational variables (e.g., PV output, SOC) at the finest scale. Given the relatively high sampling rate (e.g., second-level) and the physical inertia of power/energy states, we adopt a random walk model for short-term prediction. Therefore, $\mathbf{A}$ is constructed as an identity matrix:

$$\mathbf{A} = \mathbf{I} \quad (8)$$

The matrix $\mathbf{Q}$ models the uncertainty in the state transition. In our multi-scale fusion context, $\mathbf{Q}$ is designed not only to capture the inherent variability of each signal but also to encode cross-scale correlations.

1 Diagonal Elements (Variance of Each State):

The i -th diagonal element Q_{ii} is estimated from historical data as the variance of the innovation sequence (the difference between consecutive measurements) for that state variable over a moving window:

$$Q_{ii} = \sigma_{\Delta x_i}^2 = \text{Var}(x_i(t) - x_i(t-1)) \quad (9)$$

2 Off-Diagonal Elements (Multi-Scale Correlation):

The off-diagonal element Q_{ij} between state i and state j is calculated by introducing a cross-scale correlation parameter ρ_{ij}^{scale} :

$$Q_{ij} = \rho_{ij}^{scale} \cdot \sqrt{Q_{ii} \cdot Q_{jj}} \quad (10)$$

Here, $\rho_{ij}^{scale} \in [-1, 1]$ quantifies the expected correlation between changes at different scales of the same (or related) physical quantity. It can be derived from historical data or set based on domain knowledge (e.g., finer and coarser scales of the same PV plant are highly positively correlated).

This formulation allows the filter to coherently smooth data across scales, improving robustness when measurements from different scales are available.

Standard Kalman filter prediction and update steps are shown in Figure 2.

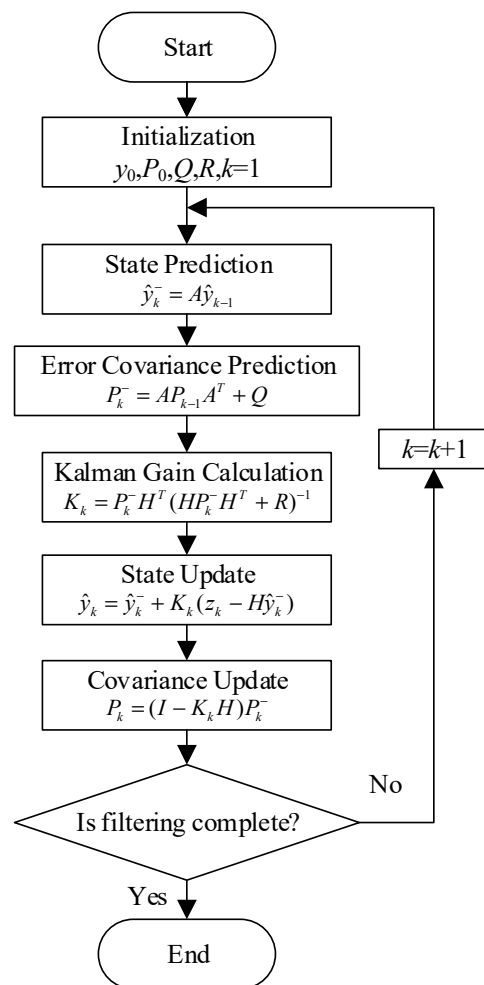


Figure 2. Flowchart of multi-scale collaborative Kalman filtering.

2.3. Scale Coarse Selection Based on Information Gain (IG)

For each preprocessed $MsDT_k$, we need to select the most representative scale from the multiple scales $\{a_i^1, a_i^2, \dots, a_i^{t_i}\}$ of each attribute a_i to achieve preliminary data source fusion and dimensionality reduction. We employ Information Gain (IG) as the scale selection criterion. IG originates from information theory and measures the reduction in uncertainty of the decision attribute d , given knowledge of an attribute (scale) value.

Definition 3. Information Gain. Let decision attribute d partition object set O into r equivalence classes: $O/d = \{D_1, D_2, \dots, D_r\}$. The entropy of decision d is defined as

$$H(d) = -\sum_{c=1}^r p(D_c) \log_2 p(D_c) \quad (11)$$

where $p(D_c) = |D_c|/|O|$.

Given the partition $O/a_i^j = \{X_1, X_2, \dots, X_m\}$ under attribute a_i at scale j , the conditional entropy is

$$H(d|a_i^j) = \sum_{t=1}^m p(X_t) H(d|X_t) = - \sum_{t=1}^m \frac{|X_t|}{|O|} \sum_{c=1}^r \frac{|X_t \cap D_c|}{|X_t|} \log_2 \frac{|X_t \cap D_c|}{|X_t|} \quad (12)$$

The information gain of scale a_i^j is then

$$IG(a_i^j) = H(d) - H(d|a_i^j) \quad (13)$$

A larger $IG(a_i^j)$ indicates stronger discriminative power of that scale for decision d , containing richer information.

This stage achieves fast, information-preserving data dimensionality reduction and cross-source alignment, establishing a unified data foundation for subsequent fine processing.

2.4. Fine Selection Based on Scale-Based Tree

After the coarse selection via Information Gain (Algorithm 1), each attribute in $MsDT_{fused}$ is represented by a single, globally optimal scale. While this achieves efficient data reduction and source alignment, it overlooks the potential heterogeneity in scale requirements among individual objects. In distributed PV-ES systems, the optimal representation scale for the same attribute (e.g., active power) can vary significantly across units due to factors like local meteorological volatility, control strategies, or grid connection characteristics. For instance, a PV unit in a stable region may be adequately characterized by hourly averages, whereas a unit experiencing rapid irradiance fluctuations requires minute- or second-level data to preserve its dynamic signature. Relying solely on a global scale may over-smooth critical variations for some objects or retain unnecessary detail for others, ultimately limiting the fidelity of the aggregated model. To address this, we reconstruct a full-scale data view for each attribute and introduce a fine-grained, object-level adaptive scale selection mechanism based on Scale-based Trees.

Algorithm 1: Scale Coarse Selection Algorithm

Input: Preprocessed $Ms\text{-}MsDT = \{MsDT_1, MsDT_2, \dots, MsDT_s\}$.

Output: Preliminary fused multi-scale decision table $MsDT_{fused}$.

Steps:

1. For each data source $k = 1$ to s , do the following:
 - a. For each attribute a_i in this data source, do the following:
 - i. Calculate information gain $IG(a_i^j)$ for all available scales $\{a_i^j | j = 1, \dots, t_i\}$ of this attribute.
 - ii. Select scale j^* that maximizes $IG(a_i^{j^*})$. If multiple scales have equal IG values, prioritize the coarser scale (larger j) to simplify data representation while retaining key information.
 - iii. Record data at selected scale $a_i^{j^*}$.
 - b. Combine the optimal scale data of all attributes from this source to form the source's concise information table IT_k .
 2. Merge concise information tables $\{IT_1, IT_2, \dots, IT_s\}$ from all s data sources based on the common object set O and unified decision attribute d to form the preliminary fused multi-scale decision table $MsDT_{fused}$. Note that each attribute in $MsDT_{fused}$ now contains only one scale (the optimal scale selected from each source).
-

Definition 4. Scale-based Tree (SbT). For attribute a_i , given its complete data across all scales, its scale tree SbT_i is defined as follows:

- The root node is labeled with the attribute name a_i .
- Except for the root, each node is a two-tuple $\langle j, V \rangle$, where j is the scale identifier ($j = 1$ represents the finest scale, larger values indicate coarser scales), and V is the set of attribute values (or equivalence class identifiers) corresponding to that node at scale j .
- Parent–child relationships represent scale refinement from coarse to fine. If the parent node is $\langle j + 1, V_{parent} \rangle$, its child node set is $\{ \langle j, V_{child} \rangle \mid V_{child} \subseteq V_{parent}, \text{ and } V_{child} \text{ is an equivalence class at scale } j \}$.

The algorithm requires building a tree and performing DFS for each attribute. Let the number of attributes be n , objects m , and average tree depth h , complexity is approximately $O(n \cdot m \cdot h)$. Its core advantage lies in achieving object-level, attribute-level optimal scale adaptive selection based on decision consistency, ensuring discriminative capability while favoring the coarsest sufficient scale. This rule-based, transparent approach stands in contrast to potential deep learning alternatives (e.g., attention-based multi-scale fusion). While such models could theoretically learn complex fusion patterns, they would require extensive labeled training data that is rarely available in practical grid scenarios and would result in opaque fusion logic unsuitable for operational auditing. Our SbT method provides an interpretable, data-efficient mechanism that directly embeds the domain principle: “Use the simplest representation that preserves relevant distinctions.”

It is worth noting that the decision attribute d used in the Ms-MsDT model and the SbT algorithm serves as a guiding proxy for the intrinsic heterogeneity within the PV-ES population (e.g., equipment type, operator affiliation, or regional label). Its primary function is not to define the final operational clusters, but to steer the multi-scale fusion process toward preserving the fundamental structural patterns in the data. This ensures that the output fused dataset retains a high-fidelity representation of the underlying system diversity. Subsequent feature engineering for specific control objectives is then performed on this high-quality, structurally rich dataset. This design separates generic data quality enhancement from application-specific modeling, improving the robustness and generalizability of the framework.

2.5. Missing Data Collaborative Completion Mechanism

In practical operation, complete data loss for certain units during specific periods may occur due to communication interruptions. For such large-scale continuous missing scenarios, we propose a collaborative completion method combining spatial correlation (geographical/electrical proximity) and multi-scale temporal patterns.

Step 1. Spatial Correlation Initial Estimation:

For the missing unit i during the missing period T_{miss} , find the K spatially or electrically most similar neighboring units set $N(i)$. Based on these neighboring units’ multi-scale fused data at the same period $\{ \mathbf{x}_j(t) \mid j \in N(i), t \in T_{miss} \}$, obtain initial missing data estimates via weighted average:

$$\hat{\mathbf{x}}_i^{(0)}(t) = \sum_{j \in N(i)} w_{ij} \cdot \mathbf{x}_j(t), \quad \sum_j w_{ij} = 1 \quad (14)$$

Weights w_{ij} can be determined based on electrical distance, historical output correlation, etc.

Step 2. Construct Temporary Multi-Scale Sequence:

Concatenate the initial estimate sequence $\hat{\mathbf{x}}_i^{(0)}(t)$ with actual valid data from periods before and after the missing interval (e.g., 2 h each), forming a complete temporary time series. For this sequence, compute statistics (mean, variance, quantiles, etc.) at different

time granularities via sliding windows, artificially constructing a temporary multi-scale data view.

Step 3. Multi-Scale Fusion Completion:

Treat this temporary multi-scale data as a “virtual data source” and apply the aforementioned Algorithm 1 (IG coarse selection) and Algorithm 2 (SbT fine selection). The decision attribute d can be temporarily set as a derived ‘data pattern category’ (obtained by clustering the neighboring units’ typical profiles), which serves as a proxy for the local behavioral pattern during the imputation process. Through two-stage fusion, obtain multi-scale optimal estimates $\hat{x}_i^{final}(t)$ for completed data.

Algorithm 2: Reverse Scale Selection Algorithm Based on SbT (Fine Fusion)

Input: $MsDT_{fused}$ and corresponding full-scale data for each attribute.

Output: Fine-fused final dataset D_{final} (can be represented in tuple or interval value form).

Steps (taking attribute a_i as an example):

1. Data Annotation and Preprocessing: To ensure clear inclusion relationships between adjacent scales in the scale tree, raw data must be preprocessed. For any object o_x , if the equivalence class X corresponding to its value $f(o_x, a_i^j)$ at fine scale j does not satisfy $X \subseteq Y$ with equivalence class Y corresponding to its value $f(o_x, a_i^{j+1})$ at coarse scale $j + 1$, then annotate the fine-scale value as $f'(o_x, a_i^j) = f(o_x, a_i^j) \star \text{label}(Y)$ to establish a clear hierarchical relationship (symbol $\star$ denotes annotation operation).

2. Build Scale Tree: Construct scale tree SbT_i for attribute a_i using annotated full-scale data.

3. Depth-First Consistency Search:

a. Initialize an empty list(tuple_list) to store selected nodes.

b. Starting from the coarsest scale nodes (direct children of the root), perform depth-first traversal (DFS) on SbT_i .

c. For currently visited node $node = \langle j, V \rangle$:

i. Find all objects whose attribute values belong to $V : X = \{o \in O | f(o, a_i^j) \in V\}$.

ii. Consistency Judgment: Check decision attribute d values of all objects in set X . If all objects in X have the same decision value (i.e., $\exists D_c \in O/d, s.t. X \subseteq D_c$), then node $node$ is decision-consistent.

iii. If $node$ is decision-consistent, record this node in tuple_list and do not further search its descendants (finer scales). This means that for these objects, the current scale j is sufficient to clearly distinguish their decision classes, requiring no finer granularity information.

iiii. If $node$ is not decision-consistent, continue deep search of its child nodes (finer scales) until a consistent node is found or the finest scale leaf node is reached. When reaching a leaf node, record it regardless of consistency.

4. Result Extraction: After traversal, tuple_list contains all selected nodes for the attribute a_i . Each object o is assigned a unique selected node $\langle j^*(o), V \rangle$ based on its attribute value, where $j^*(o)$ is the optimal scale for that object on the attribute a_i .

Tuple form output: Final data is represented as

$$D_{final} = \{(o, a_i, j^*(o), f(o, a_i^{j^*(o)})) | \forall o \in O, \forall a_i\}.$$

Interval value form output: Transform selected scale values back to the value range of the finest scale, forming interval-valued data, e.g., $f(o, a_i) \in [v_{min}, v_{max}]$.

Step 4. Iterative Optimization:

Substitute completed data into Step 1, recalculate spatial correlation weights, and iterate the above process until convergence (e.g., estimate change less than threshold ϵ).

This mechanism is designed for scenarios of prolonged contiguous data loss, where capturing both spatial correlation and the unit's intrinsic multi-scale temporal patterns is crucial. While Step 1 provides a spatially informed initial estimate $\hat{\mathbf{x}}_i^{(0)}(t)$, it may not reflect the unit's specific dynamic pattern. Steps 2–3 address this by constructing a temporary multi-scale view of this estimate and applying a lightweight, pattern-guided refinement using the core logic of our fusion framework. Here, the 'virtual data source' is the single unit's sequence, and the decision attribute d is a 'pattern category' (e.g., 'highly fluctuating', 'smooth', 'evening discharge') obtained by clustering the neighboring units' profiles. The IG and SbT steps are thus applied only to this very small dataset to adjust the estimate so its multi-scale statistics align with the identified pattern. The iterative step (Step 4) allows this pattern refinement to inform the spatial weights, seeking a consistent solution. While more complex than simple interpolation, this approach is justified for critical units or long missing periods where imputation accuracy significantly impacts aggregation model fidelity.

3. Hierarchical Weight-Guided Distributed PV-ES Aggregation

After Chapter 2 processing, we obtain a high-quality, multi-scale fused PV-ES unit feature dataset D_{final} . This chapter aims to aggregate these units into several clusters whose interpretability and alignment with grid control priorities are as important as their statistical compactness. While advanced manifold learning or spectral clustering with custom kernels could identify complex data structures, they would struggle to incorporate explicit, hierarchical domain knowledge (e.g., "voltage sensitivity matters most at 10 kV") in a transparent manner. Therefore, we employ a structured knowledge-integration approach using AHP to translate operational priorities into weights, followed by weighted K-Means++. This combination ensures clusters are not only internally coherent but also semantically meaningful and actionable for system operators, with a clear rationale traceable to the AHP judgments.

3.1. Hierarchical Control-Oriented Aggregation Metric System Construction

Extract three key feature categories from the fused data D_{final} to construct aggregation metrics:

3.1.1. Active Power Sensitivity Metric (I_S)

This metric characterizes how effectively a unit can influence and regulate node voltage by injecting active power. Based on linearized power flow equations,

$$\begin{bmatrix} \Delta \mathbf{P} \\ \Delta \mathbf{Q} \end{bmatrix} = \begin{bmatrix} \mathbf{J}_{P\theta} & \mathbf{J}_{PV} \\ \mathbf{J}_{Q\theta} & \mathbf{J}_{QV} \end{bmatrix} \begin{bmatrix} \Delta \boldsymbol{\theta} \\ \Delta \mathbf{V} \end{bmatrix} \quad (15)$$

Under steady-state operation with minimal PV-ES reactive output, approximate $\Delta \mathbf{Q} \approx 0$, yielding an active-voltage sensitivity relationship:

$$\Delta \mathbf{V} \approx \mathbf{S}_{PV} \cdot \Delta \mathbf{P}, \quad \mathbf{S}_{PV} = \mathbf{J}_{QV}^{-1} \cdot (-\mathbf{J}_{Q\theta} \cdot \mathbf{J}_{P\theta}^{-1} \cdot \mathbf{J}_{PV} + \mathbf{J}_{QV}) \quad (16)$$

Matrix $\mathbf{S}_{PV}$ is the active-voltage sensitivity matrix. For unit i connected at node n , take the absolute value (or log-transformed value) of the corresponding diagonal element S_{PV}^{nn} as sensitivity scalar s_i . To enhance discriminability, construct a feature vector:

$$\mathbf{v}_S^i = [s_i, s_i^2, \text{sign}(S_{PV}^{nn})]^T \quad (17)$$

where $\text{sign}(S_{PV}^{nn})$ denotes the sign of the self-sensitivity, indicating the direction of local voltage change (+1 for voltage rise, −1 for voltage drop) in response to active power injection at node n .

This vector simultaneously contains sensitivity magnitude, nonlinear degree, and influence direction information.

3.1.2. PV Output Characteristic Metric (I_P)

Characterizes the daily output curve shape of units. Discretize unit i 's fused typical daily output curve $P_i(t)$ into T periods (e.g., $T = 96$, 15 min intervals). Calculate average output for each period k :

$$\bar{P}_i^k = \frac{1}{N_{day}} \sum_{d=1}^{N_{day}} P_i^d(t_k), \quad k = 1, \dots, T \quad (18)$$

To capture curve shape, further compute first-order differences (reflecting fluctuations) and normalization factor:

$$\Delta \bar{P}_i^k = \bar{P}_i^k - \bar{P}_i^{k-1}, \quad P_i^{norm} = \frac{\max(\bar{P}_i) - \min(\bar{P}_i)}{\max(\bar{P}_i)} \quad (19)$$

Construct a comprehensive feature vector:

$$\mathbf{v}_P^i = [\bar{P}_i^1, \dots, \bar{P}_i^T, \Delta \bar{P}_i^1, \dots, \Delta \bar{P}_i^T, P_i^{norm}]^T \quad (20)$$

This vector completely describes the output average level, temporal variation pattern, and fluctuation rate.

3.1.3. Operator Difference Metric (I_O)

Identifies unit management affiliation, directly affecting local coordination and control costs. For M operators, use one-hot encoding:

$$\mathbf{v}_O^i = [0, \dots, 1, \dots, 0]^T \in \mathbb{R}^M \quad (21)$$

Alternatively, if cooperative or competitive relationships exist among operators, assign differentiation coefficients $z_a, z_b, \dots$, then $\mathbf{v}_O^i = [z_{op(i)}]$.

For each unit i , concatenate all features to obtain the original high-dimensional feature vector:

$$\mathbf{v}_i^{raw} = [(\mathbf{v}_S^i)^T, (\mathbf{v}_P^i)^T, (\mathbf{v}_O^i)^T]^T \quad (22)$$

3.2. Hierarchical Differentiated Weight Determination Based on AHP

The objectives and constraints of power grid operation differ fundamentally at different voltage levels. Consequently, the emphasis placed on each of the three metrics also varies.

400 V Station Area Level: Operation focuses on ensuring low-voltage user power quality and equipment safety, achieving autonomous station area balance. Due to the small station area scope, units under the same operator are easier to coordinate and control (e.g.,

receiving unified dispatch commands), so the operator difference metric (I_O) should have maximum weight.

10 kV Distribution Network Level: Main challenges involve voltage limit violations and line overloads caused by high distributed generation penetration. Core control means at this level involve adjusting reactive and/or active power to modify node voltages, making units with high grid voltage sensitivity key regulation resources. The active power sensitivity metric (I_S) should have maximum weight at this level.

110 kV Transmission Network Level: It focuses on system-wide power balance, frequency stability, and congestion management. PV-ES clusters need to provide reliable, predictable power regulation capability. Unit daily output characteristics (I_P), including regularity, predictability, and alignment with system load, become critical, so this metric should have maximum weight.

The Analytic Hierarchy Process (AHP) quantifies these qualitative judgments for each level. Taking the station area level as an example, construct a judgment matrix $\mathbf{A}_{400V}$:

$$\mathbf{A}_{400V} = \begin{bmatrix} 1 & a_{SO} & a_{SP} \\ 1/a_{SO} & 1 & a_{OP} \\ 1/a_{SP} & 1/a_{OP} & 1 \end{bmatrix} \quad (23)$$

where a_{SO} represents the importance scale (1–9 scale) of I_S relative to I_O , similarly for a_{SP} , a_{OP} . The matrix satisfies reciprocity $a_{ij} = 1/a_{ji}$.

By solving the judgment matrix's maximum eigenvalue $\lambda_{\max}$ and its corresponding normalized eigenvector $\mathbf{w}$, initial weights $\mathbf{W}_{400V} = [w_S, w_O, w_P]^T$ are obtained. To ensure logical consistency of judgments, compute the consistency ratio CR :

$$CI = \frac{\lambda_{\max} - n}{n - 1}, \quad CR = \frac{CI}{RI} \quad (24)$$

where $n = 3$, RI is the average random consistency index, which can be obtained from a lookup table.

If $CR < 0.1$, consistency test passes; otherwise, adjust the judgment matrix.

Similarly, construct $\mathbf{A}_{10kV}$ and $\mathbf{A}_{110kV}$, obtaining weight vectors $\mathbf{W}_{10kV}$ and $\mathbf{W}_{110kV}$, respectively. Typical weight results are shown in Table 3.

Table 3. Example of metric weights for each aggregation level based on AHP.

Aggregation Level	Weights		
	Active Sensitivity (w_S)	PV Output Characteristic (w_P)	Operator Difference (w_O)
400 V Station Area	0.12	0.25	0.63
10 kV Distribution Network	0.53	0.36	0.11
110 kV Transmission Network	0.38	0.57	0.05

The Analytic Hierarchy Process (AHP) is used to systematically translate qualitative operational priorities into a quantitative weight vector for each level. It is crucial to note that the specific weights are not fixed parameters of the method, but are inputs that can be derived from grid operator expertise, historical control performance analysis, or optimization objectives. The values in Table 3 serve as a realistic example based on typical operational logic for the purpose of this case study.

In practical deployment, these weights could be determined through:

- Expert Elicitation: System operators define judgment matrices for different forecasted grid conditions (e.g., high PV noon, evening peak).
- Offline Optimization: Weights could be tuned via historical data to maximize metrics like voltage violation reduction or control cost minimization for each level.
- Dynamic Adjustment: The framework can accommodate time-varying weight vectors $\mathbf{W}_L(t)$ updated at a slower timescale (e.g., hourly or daily) based on system state forecasts, enabling the aggregation to adapt to changing grid needs.

In addition, the ‘Operator Difference’ metric (I_O) at the 400 V level is most effective when it correlates with a unified local control authority. If a single operator’s units are dispersed across non-interconnected station areas, they would naturally fall into different local coordination domains. In such cases, the metric can be redefined to represent ‘Local Control Group ID’ rather than a broad operator ID. The proposed method is agnostic to the specific semantics of this metric; it only requires that the feature meaningfully represents a cost or constraint related to coordinated control.

3.3. Weight-Guided Improved K-Means Hierarchical Aggregation Algorithm

This step is performed independently for each voltage level $L \in \{400\text{V}, 10\text{kV}, 110\text{kV}\}$, aiming to obtain unit clustering that meets that level’s control requirements.

Step 1. Feature Preprocessing and Weighting:

Since metrics have different units and value ranges, first apply min–max normalization to each sub-vector of the original feature vector $\mathbf{v}_i^{raw}$ to eliminate scale effects:

$$\tilde{\mathbf{v}}_S^i = \frac{\mathbf{v}_S^i - \min(\mathbf{v}_S)}{\max(\mathbf{v}_S) - \min(\mathbf{v}_S)}, \quad \tilde{\mathbf{v}}_P^i, \tilde{\mathbf{v}}_O^i \text{ similarly.} \quad (25)$$

Obtain a normalized feature vector $\bar{\mathbf{v}}_i = [(\bar{\mathbf{v}}_S^i)^T, (\bar{\mathbf{v}}_P^i)^T, (\bar{\mathbf{v}}_O^i)^T]^T$.

According to level L ’s weight vector $\mathbf{W}_L = [w_S^L, w_P^L, w_O^L]^T$, we generated weight-normalized features to emphasize that level’s focus:

$$\mathbf{x}_i^{(L)} = [w_S^L \cdot (\bar{\mathbf{v}}_S^i)^T, w_P^L \cdot (\bar{\mathbf{v}}_P^i)^T, w_O^L \cdot (\bar{\mathbf{v}}_O^i)^T]^T \quad (26)$$

Weighted feature vector $\mathbf{x}_i^{(L)}$ is the comprehensive representation of unit i at level L .

Step 2. Determine Optimal Cluster Number K_L^* :

Use the Elbow Method to determine the appropriate cluster count for that level. Compute Sum of Squared Errors (SSE) for different K values (e.g., $K = 2$ to $K = 10$):

$$SSE_L(K) = \sum_{j=1}^K \sum_{\mathbf{x} \in C_j^{(L)}} \|\mathbf{x} - \mathbf{c}_j^{(L)}\|^2 \quad (27)$$

Plot $SSE_L(K)$ versus the K curve and select the elbow point corresponding to K as K_L^* .

Step 3. Improved K-Means Clustering Algorithm:

To avoid traditional K-Means’ sensitivity to initial centers and local optima, employ the K-Means++ algorithm for improved initial center selection, combined with weighted features for iteration.

Step 4. Output Hierarchical Aggregation Results:

Run Algorithm 3 separately for three levels, finally obtaining three clustering results:

- Station area level aggregation: $\mathcal{C}^{(400\text{V})} = \{C_1^{(400\text{V})}, \dots, C_{K_{400\text{V}}^*}^{(400\text{V})}\}$
- Distribution network level aggregation: $\mathcal{C}^{(10\text{kV})} = \{C_1^{(10\text{kV})}, \dots, C_{K_{10\text{kV}}^*}^{(10\text{kV})}\}$
- Transmission network level aggregation: $\mathcal{C}^{(110\text{kV})} = \{C_1^{(110\text{kV})}, \dots, C_{K_{110\text{kV}}^*}^{(110\text{kV})}\}$

Algorithm 3: Weight-Guided Improved K-Means

Input: Level L weighted feature set $\mathcal{X}^{(L)} = \{\mathbf{x}_1^{(L)}, \dots, \mathbf{x}_N^{(L)}\}$, cluster count K_L^* .

Output: Clustering partition $\mathcal{C}^{(L)} = \{C_1^{(L)}, C_2^{(L)}, \dots, C_{K_L^*}^{(L)}\}$ and cluster centers $\{\mathbf{c}_1^{(L)}, \dots, \mathbf{c}_{K_L^*}^{(L)}\}$.

1. Initialize Cluster Centers:

- Randomly uniformly select a point from $\mathcal{X}^{(L)}$ as the first cluster center $\mathbf{c}_1$.
- For each point $\mathbf{x}_i$, compute its shortest distance to selected centers

$$D(\mathbf{x}_i) = \min_{\mathbf{c} \in C_{\text{selected}}} \|\mathbf{x}_i - \mathbf{c}\|.$$

- Select the next center point according to probability $P(\mathbf{x}_i) = D(\mathbf{x}_i)^2 / \sum_{\mathbf{x}_j} D(\mathbf{x}_j)^2$. This

distribution gives a higher probability to points farther from selected centers, resulting in more dispersed initial centers.

- Repeat steps b–c until K_L^* initial centers are selected.

2. Iterative Assignment and Update:

- Assignment Step: Assign each point $\mathbf{x}_i$ to the cluster of its nearest Euclidean distance cluster center.

$$C_j^{(L)} = \left\{ \mathbf{x}_i^{(L)} : \|\mathbf{x}_i^{(L)} - \mathbf{c}_j^{(L)}\|^2 \leq \|\mathbf{x}_i^{(L)} - \mathbf{c}_k^{(L)}\|^2 \forall k, 1 \leq k \leq K_L^* \right\}$$

- Update Step: Recalculate each cluster's centroid as the new cluster center.

$$\mathbf{c}_j^{(L)} = \frac{1}{|C_j^{(L)}|} \sum_{\mathbf{x} \in C_j^{(L)}} \mathbf{x}$$

- Repeat steps a and b until cluster center changes are less than preset threshold δ (e.g., 10^{-5}) or maximum iterations reached.

Clusters within each level's results maximally align with that level's control focus and can directly guide dispatch and control strategy formulation for that level.

4. Case Study

To comprehensively validate the proposed method's effectiveness, a test case is constructed based on the extended IEEE-33 node distribution system (topology shown in Figure 3).

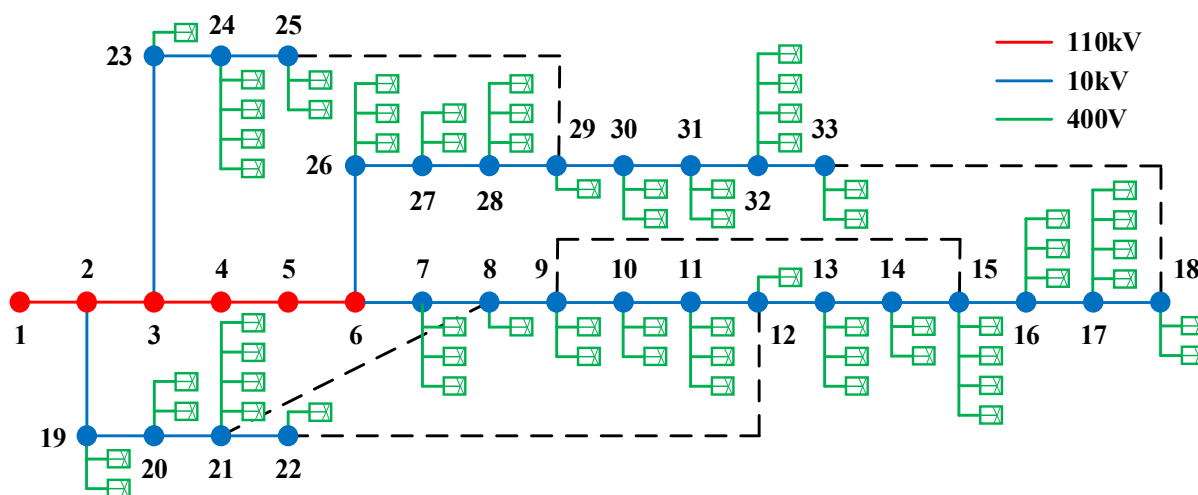


Figure 3. Topology of the extended IEEE-33 node test system and PV-ES connection points.

Based on actual meteorological data (irradiance, temperature), generate 30-day, 1 s resolution raw output sequences for PV units using a combination of physical models and random perturbations. Then, derive three coarser scale sequences (5 min, 15 min, hourly) via sliding window averaging.

Simulate ES charge/discharge behavior considering strategies such as smoothing PV fluctuations and peak–valley arbitrage, generating second-level data for SOC, charge/discharge power, etc., and similarly deriving multi-scale sequences.

Randomly assign units to 3 virtual operators (A, B, C). Simulate differences in operator data systems: Operator A data in CSV format provides second-level and hourly data; Operator B provides 5 min and 15 min data via JSON API; Operator C uses binary format, providing second-level and hourly data.

Randomly inject the following into complete data: (a) 5% random outliers; (b) 10% random points missing. (c) Simulate one 6 h communication interruption, causing complete data loss for 1 unit during that period.

4.1. Multi-Source Multi-Scale Data Fusion Effectiveness Analysis

First, evaluate the performance of the data fusion stage (Section 2). Taking the key attribute “PV active output” as an example, compare data quality after processing by different methods. Quality assessment requires ground truth values as a baseline; use theoretically calculated values from simulation models before noise and outlier removal as truth.

Figure 4 shows the results of the data processing on the PV output sampled values. The black curve represents the actual PV output value, the red curve represents the noisy sampled value, and the blue curve represents the value after using the proposed method. It can be observed that the PV output curve processed by the above multi-source heterogeneous data processing method can effectively remove noise. Moreover, the missing data for Unit 2 during the specific period can be effectively imputed, showing a high degree of agreement with the actual photovoltaic output values.

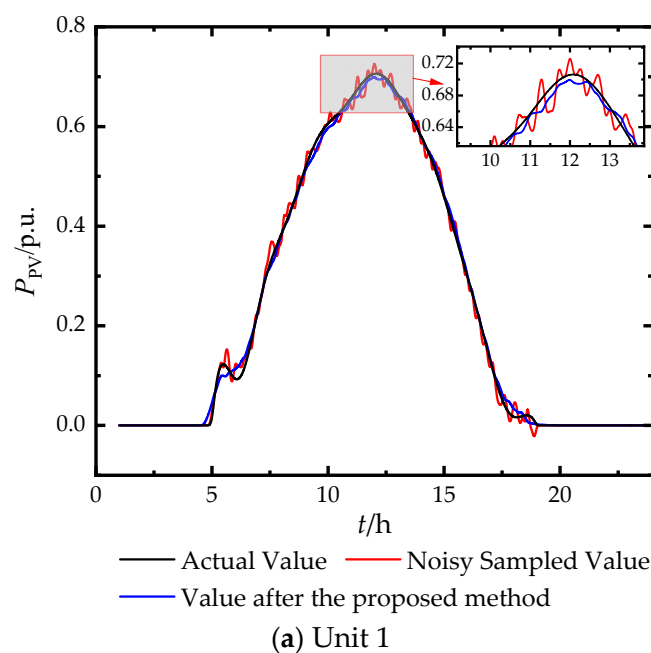


Figure 4. Cont.

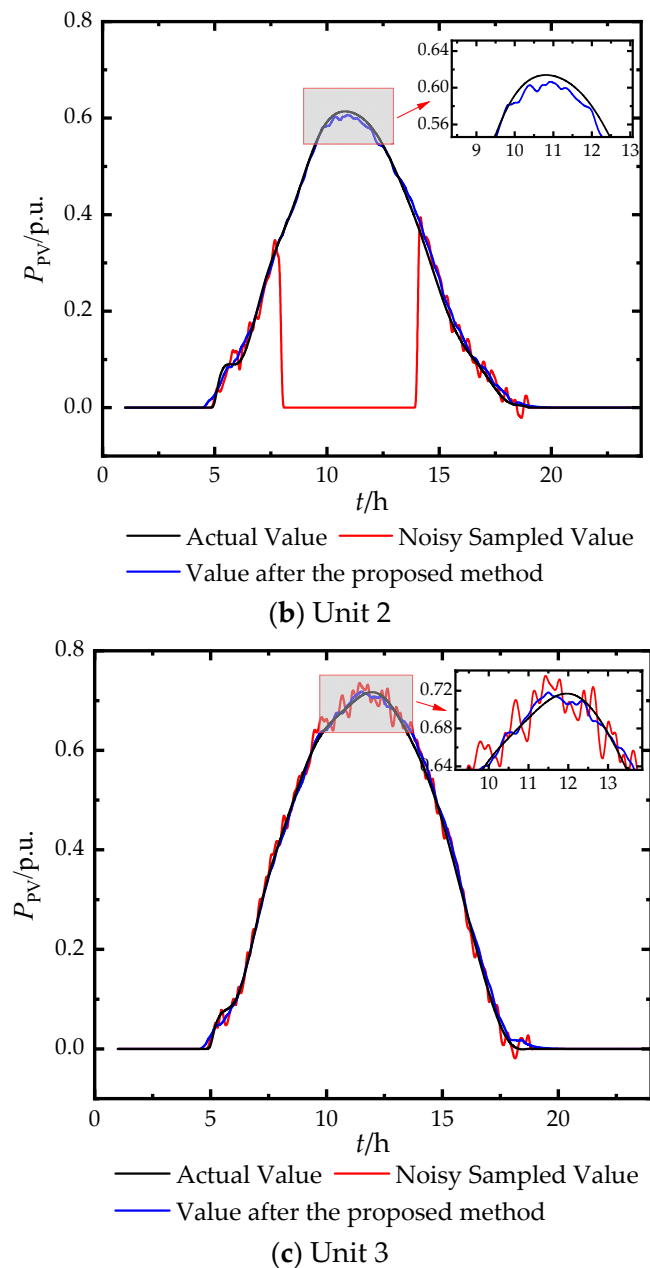


Figure 4. Multi-source heterogeneous data processing results of PV output: (a) Unit 1; (b) Unit 2; (c) Unit 3.

4.1.1. Evaluation of Missing Data Completion Performance

To quantitatively assess the proposed collaborative completion mechanism (Section 2.5), we compare it against three baseline methods for imputing the 6 h missing block in Unit 2's data: (1) Spatial Kriging using data from the five most electrically proximate neighbors; (2) Temporal Linear Interpolation between the known values before and after the gap; and (3) a Matrix Completion method (soft-impute) applied to the daily profile matrix of all units. The comparison metric is the Normalized Root Mean Square Error (NRMSE) between the imputed sequence and the ground truth (simulated actual values), calculated as

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{T} \sum_{t=1}^T (y_t - \hat{y}_t)^2}}{\max(\mathbf{y}) - \min(\mathbf{y})}, \quad (28)$$

where y_t is the ground truth and $\hat{y}_t$ is the imputed value.

The results are presented in Table 4.

Table 4. Imputation performance comparison for the 6 h missing block (Unit 2).

Imputation Method	NRMSE	Description
Proposed Method	0.078	Combines spatial correlation and multi-scale pattern refinement
Spatial Kriging	0.121	Uses spatial correlation only
Temporal Linear Interpolation	0.185	Ignores spatial information, assumes smooth transition
Matrix Completion (Soft-Impute)	0.105	Exploits low-rank structure across units and times

The results demonstrate that the proposed method achieves the lowest NRMSE, outperforming the pure spatial (Kriging) and pure low-rank (Matrix Completion) approaches. This validates the benefit of integrating spatial neighbors' information with temporal multi-scale pattern analysis for completing prolonged missing blocks in distributed PV-ES data. The added computational cost of the proposed mechanism is justified in contexts where high imputation accuracy is critical for downstream modeling.

4.1.2. Data Fusion Quality Assessment

Use Approximation Precision (AP) and Approximation Quality (AQ), commonly used in multi-source information fusion. Their definitions are based on rough set theory: for a decision attribute (here set as “unit typical output pattern”, obtained by clustering truth curves into three classes), classify objects using fused data.

Let the fused dataset for the evaluated attribute (e.g., PV active output) consist of N objects (units). For each object i , its fused data is represented as a feature vector $\mathbf{x}_i$. We first construct a normalized distance matrix $\mathbf{D}$, where the element D_{ij} is the Euclidean distance between vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, divided by the maximum distance observed in the dataset $D_{\max}$:

$$D_{ij} = \frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{D_{\max}}, \quad D_{\max} = \max_{k,l}(\|\mathbf{x}_k - \mathbf{x}_l\|_2) \quad (29)$$

The neighborhood of an object x for a given threshold α is then defined as the set of objects within a normalized distance α :

$$[x]_{R_\alpha} = \{y \in O \mid D(x, y) \leq \alpha\} \quad (30)$$

Thus, the neighborhood relation is $R_\alpha = \{(x, y) \mid D(x, y) \leq \alpha\}$. Parameter α controls the granularity of the neighborhoods: a small α creates small, precise neighborhoods, while a large α creates larger, more inclusive ones.

- Lower approximation: $\underline{R}(D) = \{x \in O \mid [x]_R \subseteq D\}$
- Upper approximation: $\overline{R}(D) = \{x \in O \mid [x]_R \cap D \neq \emptyset\}$

Then:

$$AP = \frac{\sum_{i=1}^3 |\underline{R}(D_i)|}{\sum_{i=1}^3 |\overline{R}(D_i)|}, \quad AQ = \frac{\sum_{i=1}^3 |\underline{R}(D_i)|}{|O|} \quad (31)$$

AP measures the proportion of “certain” classification parts, and AQ measures the proportion of objects that can be accurately classified relative to the total. Higher values indicate stronger discriminative capability and higher quality of fused data. Parameter α controls neighborhood size.

Table 5 shows AP, AQ values, and fusion processing time for PV output data processed by different fusion methods when $\alpha = 0.1$. Figure 5 shows the trends of AP and AQ as α varies.

Table 5. Comparison of PV output data quality under different fusion methods ($\alpha = 0.1$).

Fusion Method	AP ¹	AQ ²	Fusion Time (s)
Raw Noisy Data	0.65	0.71	-
Traditional	0.78	0.83	2.1
Proposed (Stage1: IG coarse selection)	0.85	0.88	1.4
Proposed (Full: IG + SbT)	0.92	0.94	3.8

¹ AP stands for Approximation Precision. ² AQ stands for Approximation Quality.

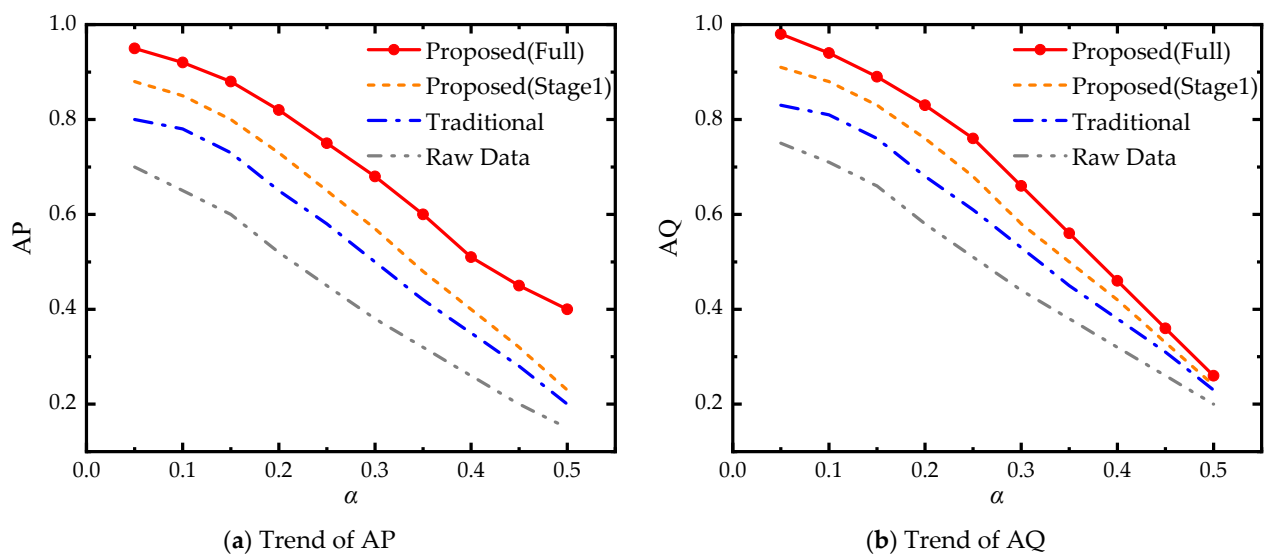


Figure 5. Trend of AP and AQ for different fusion methods across α values: (a) AP; (b) AQ.

The experimental results demonstrate that the proposed multi-source multi-scale data fusion method exhibits significant advantages in enhancing data quality. As shown in Table 5, with the key parameter α set to 0.1, the complete two-stage fusion method (Proposed Full) achieved approximation precision (AP) and approximation quality (AQ) of 0.92 and 0.94, respectively, which are substantially higher than the 0.78 and 0.83 obtained by the traditional fusion method. While the scale coarse selection using only information gain (Proposed Stage1) was the fastest in terms of processing speed, its data quality (AP = 0.85, AQ = 0.88) still surpassed that of the traditional method, confirming the effectiveness of information gain as a rapid screening tool.

Figure 4 further reveals the robustness of the proposed method under different α parameters. Parameter α controls the granularity of the similarity measure in the fused feature space. As shown in Figure 4, the AP and AQ values for all methods vary with α . For very large α (e.g., >0.25), neighborhoods become too inclusive, blurring the distinctions between decision classes and causing AP and AQ to decline. Across a wide range of α from 0.05 to 0.3, the AP and AQ curves of the proposed method consistently remained at the topmost position and exhibited the gentlest decline, indicating its insensitivity to neighborhood size selection and excellent robustness. The value $\alpha = 0.1$ used in Table 5 was chosen as it lies centrally within this stable, high-performance region for the proposed method. It represents a balanced setting where neighborhoods are specific enough to capture class structure but inclusive enough to provide meaningful lower and upper approximations. The key conclusion from the sensitivity analysis is not the absolute value at a single α , but the consistent outperformance and broader plateau of the proposed

method, which validates its robustness. Overall, the deep mining and fusion of multi-scale information are key to enhancing the intrinsic representational capacity of the data, laying a solid foundation for subsequent high-quality clustering analysis.

It is worth noting that the proposed two-stage fusion framework introduces an incremental computational cost compared to using IG alone (3.8 s vs. 1.4 s in Table 5, primarily due to the SbT construction and search). However, this trade-off is justified and strategically aligned with the requirements of distributed PV-ES aggregation modeling. First, the fusion process is typically executed as an offline or periodic pre-processing step (e.g., daily or hourly) to prepare high-quality feature datasets for downstream scheduling, planning, or control algorithms. In such contexts, a processing time of a few seconds is acceptable, especially when compared to the timescales of market-clearing or dispatch cycles. Second, the quality gains from the SbT stage—reflected in the improved AP and AQ metrics—directly enhance the discriminative power and representational accuracy of the fused data. As demonstrated in Section 4.2, this higher-quality input translates into superior clustering compactness (lower SSE for core metrics at each hierarchical level, Table 6), which is critical for ensuring that aggregated clusters are both internally coherent and aligned with specific grid-control objectives. Therefore, while the IG stage provides a fast and effective initial screening, the SbT stage delivers the precision necessary for building reliable aggregation models in heterogeneous environments. For systems with extreme real-time constraints, the IG-only variant offers a viable rapid alternative; however, for most planning, analysis, and operational scenarios where model accuracy is paramount, the complete two-stage approach offers a favorable balance between computational effort and model fidelity.

Table 6. Comparison of core metric clustering effectiveness across levels (SSE, lower is better).

Aggregation Level	Core Metric	Traditional	MS-Only ¹	Proposed
400 V Station Area	Operator Difference SSE ²	0.24	0.28	0.12
10 kV Distribution Network	Active Sensitivity SSE	0.21	0.19	0.15
110 kV Transmission Network	PV Output Characteristic SSE	0.23	0.18	0.14

¹ “MS-Only” refers to the aggregation method that solely considers the fusion of multi-source data. ² SSE stands for Sum of Squared Errors.

4.2. Distributed PV-ES Hierarchical Aggregation Effectiveness Analysis

Next, evaluate the aggregation stage (Section 3) effectiveness. First, obtain metric weights for each level via AHP, with results consistent with the Table 3 example. Then, run three comparative methods for aggregation.

For each level, we care about the intra-cluster cohesion of its core metric (the metric with maximum weight). Compute the within-cluster Sum of Squared Errors (SSE) for the feature dimensions corresponding to that core metric under each method. Results shown in Table 6.

It can be seen that the proposed method achieves the lowest SSE across all three levels—0.12 for operator difference at the 400 V level, 0.15 for active sensitivity at the 10 kV level, and 0.14 for PV output characteristics at the 110 kV level—demonstrating its superior ability to produce tightly clustered groups that align with each level’s primary control objective. In contrast, the MS-Only approach, which uses multi-scale fused data without hierarchical weighting, shows mixed results; it slightly outperforms the traditional MF-Trad method in some metrics but underperforms in others, such as operator difference at the 400 V level. This pattern underscores the essential role of the hierarchical weight-guidance mechanism in steering the clustering process toward domain-specific priorities, thereby ensuring that the aggregated clusters are both internally coherent and practically meaningful for grid operation.

To assess the impact of the prescribed weights on the clustering outcome, we conduct a sensitivity analysis around the example weights in Table 3. For each voltage level L , we define a baseline weight vector $\mathbf{W}_L^{base}$ and generate perturbed vectors $\mathbf{W}_L^{pert}$ by independently varying each weight component by ± 0.2 (subject to the sum constraint $\sum w = 1$). For each perturbed vector, we run the weight-guided K-Means algorithm and compute the resulting core metric SSE. We then calculate the maximum relative deviation of the SSE from its baseline value:

$$\Delta SSE_L^{max} = \max_{\mathbf{W}_L^{pert}} \left(\frac{|SSE(\mathbf{W}_L^{pert}) - SSE(\mathbf{W}_L^{base})|}{SSE(\mathbf{W}_L^{base})} \right) \quad (32)$$

The results, presented in Table 7 below, show that while clustering compactness (SSE) varies with the weights, the proposed method consistently produces lower SSE compared to traditional and MS-Only methods across all tested weight perturbations. This indicates that the hierarchical weight-guidance mechanism is effective across a range of plausible priority assignments, not just the specific example. This analysis validates the need for careful, context-specific weight determination but also confirms the robustness of the overall aggregation framework to reasonable variations in these inputs.

Table 7. Sensitivity of core metric SSE to weight perturbations.

Aggregation Level	Baseline SSE	Max Relative Deviation ΔSSE^{max}	Range of Perturbed SSE Values
400 V Station Area	0.12	16.7%	[0.10, 0.14]
10 kV Distribution Network	0.15	13.3%	[0.13, 0.17]
110 kV Transmission Network	0.14	14.3%	[0.13, 0.16]

4.3. Validation of Clusters' Operational Utility

To demonstrate the practical, system-level benefits of the generated clusters beyond internal metrics, we conduct simulation-based validation experiments targeting the core control objective at each voltage level.

- 400 V Station Area Level: Local Coordination Consistency

We simulate a local demand response event and measure the synchronization of units within a cluster using the Cluster Response Consistency Coefficient (CRCC), defined as the average Pearson correlation between the commanded power change and the actual change in each unit. A CRCC closer to one indicates a highly coordinated response. As shown in Table 8, our operator-difference-weighted clusters achieve the highest CRCC, confirming effective identification of groups with consistent control behavior.

$$CRCC = \frac{1}{N_c} \sum_{k=1}^{N_c} \rho(\Delta \mathbf{P}_{cmd}, \Delta \mathbf{P}_{act}^{(k)}) \quad (33)$$

- 10 kV Distribution Network Level: Voltage Control Efficiency

We simulate a high-PV midday scenario inducing overvoltage. A control strategy preferentially curtails clusters with higher aggregate active power sensitivity. We evaluate the Voltage Regulation Economy Index (VREI), defined as the total curtailed energy (MWh) divided by the achieved reduction in the maximum voltage deviation (p.u.). A lower VREI indicates more efficient control. As shown in Table 8, clusters formed by our method achieve the lowest VREI, enabling a ~40% reduction in energy curtailment cost for the same voltage correction.

$$VREI = \frac{\sum P_{curt}}{\Delta V_{max}} \quad (34)$$

- 110 kV Transmission Network Level: Aggregate Power Predictability

We analyze the 24 h aggregate power profile of each cluster formed at this level. We evaluate the Normalized Mean Absolute Error (NMAE) of a persistence forecast, normalized by the cluster's rated capacity. A lower NMAE indicates a smoother, more predictable resource. Results in Table 8 show that our clusters exhibit the lowest NMAE, demonstrating a ~29% improvement in short-term predictability compared to the traditional method.

$$\text{NMAE} = \frac{\frac{1}{T} \sum_{t=1}^T |P_{agg}(t) - P_{agg}(t-1)|}{P_{rated}} \quad (35)$$

Table 8. Comparison of clusters' operational utility across hierarchical levels.

Aggregation Level	Key Performance Indicator (KPI)	Traditional	MS-Only	Proposed
400 V Station Area	Cluster Response Consistency Coefficient (CRCC)	0.82	0.88	0.95
10 kV Distribution Network	Voltage Regulation Economy Index (VREI) [MWh/p.u.]	141.25	124.84	85.14
110 kV Transmission Network	Aggregate Power Forecast NMAE [%]	8.7	7.9	6.2

The results in Table 8 provide clear, simulation-based evidence that the proposed hierarchical aggregation method creates functionally meaningful clusters that directly enhance system operation. The clusters not only exhibit internal cohesion (Table 6) but also deliver superior performance in voltage control efficiency, power predictability, and local coordination consistency. This validates the practical relevance of our method in improving the dispatchability and controllability of distributed PV-ES systems.

5. Conclusions

Aiming at the challenges of multi-source heterogeneous data fusion and the significant differences in control requirements across different voltage levels in distributed photovoltaic (PV) and energy storage (ES) aggregation and control, this paper proposes a hierarchical aggregation method for distributed PV and ES based on multi-source and multi-scale data fusion. Firstly, a novel two-stage data fusion framework is established. The initial stage employs Information Gain (IG) for rapid, coarse-scale selection and cross-source alignment. Subsequently, the fine-grained stage constructs Scale-based Trees (SbT) for each attribute, utilizing a consistency-based depth-first search to achieve adaptive, object-level optimal scale selection, thereby fundamentally enhancing data quality and interpretability. Building upon the high-quality fused data, three aggregation indicators—active power sensitivity, PV output characteristics, and operator differences—are constructed in accordance with the distinct control focuses of different voltage levels (400 V station area, 10 kV distribution, and 110 kV transmission networks). The Analytic Hierarchy Process (AHP) is applied to determine differentiated weighting vectors for these indicators at each hierarchical level. Furthermore, an improved K-Means++ clustering algorithm, guided by these hierarchical weights, is proposed to achieve targeted aggregation that aligns with specific operational objectives. Test results on the IEEE-33 node system demonstrate that the proposed method significantly improves data fusion metrics (AP and AQ) and achieves superior clustering compactness (lower SSE) for core metrics at each level compared to traditional approaches.

Despite the encouraging results, the proposed method has certain limitations that merit further investigation. Firstly, the current multi-scale fusion and weight determination processes operate in an offline or quasi-offline manner, which may pose challenges for direct deployment in large-scale real-time control systems requiring sub-second decision-making. The computational overhead of constructing Scale-based Trees (SbT) and performing hierarchical clustering, though manageable in our case study, needs further optimization

for real-time applications with thousands of distributed units. Secondly, the method's performance relies on the availability and quality of multi-source data; communication delays or persistent data loss in practical systems could affect fusion accuracy. Future work will focus on developing lightweight, incremental versions of the fusion and clustering algorithms to enhance real-time capability. Additionally, integrating the framework with model predictive control (MPC) or deep reinforcement learning (DRL) strategies could enable adaptive, closed-loop aggregation that responds dynamically to changing grid conditions and market signals. Testing the method in real-world demonstration projects with actual SCADA and PMU data streams will be an essential next step to validate its scalability and practical robustness.

Author Contributions: Conceptualization: S.Y. and X.H.; investigation: S.Y., X.H., and L.W.; methodology: S.Y., X.H., G.S., M.S., and J.L.; software: S.Y., Z.M., Z.L., and Z.T.; resources: S.Y., L.W., and X.H.; writing—original draft preparation: S.Y., X.H., and G.S.; writing—review and editing: S.Y., X.H., M.S., and Z.M.; supervision: S.Y. and L.W.; project administration, S.Y. and L.W.; funding acquisition, S.Y., L.W., and X.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the State Grid Corporation headquarters technology project under grant number 5419-202316463A-3-2-ZN.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: Author Shaobo Yang, Xuekai Hu, Lei Wang, Guanghui Sun, Min Shi, Zhengji Meng and Zifan Li were employed by State Grid Hebei Electric Power Company Electric Power Research Institute. 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. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

PV	Photovoltaic
ES	Energy Storage
Ms-MsDT	Multi-source Multi-Scale Decision Table
MsIT	Multi-scale Information Table
IG	Information Gain
SbT	Scale-based Trees
AHP	Analytic Hierarchy Process
SSE	Sum of Squared Errors
AP	Approximation Precision
AQ	Approximation Quality

References

1. Li, X.; Wang, L.; Yan, N.; Ma, R. Cooperative dispatch of distributed energy storage in distribution network with PV generation systems. *IEEE Trans. Appl. Supercond.* **2021**, *31*, 1–4. [\[CrossRef\]](#)
2. Wang, J.; Sun, K.; Wu, H.; Zhu, J.; Xing, Y.; Li, Y. Hybrid connected unified power quality conditioner integrating distributed generation with reduced power capacity and enhanced conversion efficiency. *IEEE Trans. Ind. Electron.* **2020**, *68*, 12340–12352. [\[CrossRef\]](#)
3. Lorente, J.L.; Liu, X.; Morrow, D.J. Spatial aggregation of small-scale photovoltaic generation using Voronoi decomposition. *IEEE Trans. Sustain. Energy* **2020**, *11*, 2677–2686. [\[CrossRef\]](#)
4. León, L.F.; Martinez, M.; Ontiveros, L.J.; Mercado, P.E. Devices and control strategies for voltage regulation under influence of photovoltaic distributed generation. A review. *IEEE Lat. Am. Trans.* **2022**, *20*, 731–745. [\[CrossRef\]](#)
5. Jin, Y.; Wu, D.; Ju, P.; Rehtanz, C.; Wu, F.; Pan, X. Modeling of Wind Speeds Inside a Wind Farm with Application to Wind Farm Aggregate Modeling Considering LVRT Characteristic. *IEEE Trans. Energy Convers.* **2020**, *35*, 508–519. [\[CrossRef\]](#)

6. Muljadi, E.; Pasupulati, S.; Ellis, A.; Kostrov, D. Method of equivalencing for a large wind power plant with multiple turbine representation. In Proceedings of the 2008 IEEE Power and Energy Society General Meeting—Conversion and Delivery of Electrical Energy in the 21st Century, Pittsburgh, PA, USA, 20–24 July 2008; pp. 1–9.
7. Sloatweg, J.G.; Polinder, H.; Kling, W.L. Dynamic modeling of a wind turbine with doubly fed induction generator. In Proceedings of the IEEE Power Engineering Society Summer Meeting, Vancouver, BC, Canada, 15–19 July 2001; pp. 644–649.
8. El-Sharkawi, M.A. Dynamic equivalent models for wind power plants. In Proceedings of the IEEE PES General Meeting, Detroit, MI, USA, 24–28 July 2011.
9. Ali, M.; Ilie, I.S.; Milanovic, J.V.; Chicco, G. Wind farm model aggregation using probabilistic clustering. *IEEE Trans. Power Syst.* **2013**, *28*, 309–316. [[CrossRef](#)]
10. Hussein, D.N.; Matar, M.; Iravani, R. A type-4 wind power plant equivalent model for the analysis of electromagnetic transients in power systems. *IEEE Trans. Power Syst.* **2013**, *28*, 3096–3104. [[CrossRef](#)]
11. Wang, K.; Yan, J.; Yang, Z.; Hong, Y. PV Scenario Generation Based on K-means Method. In Proceedings of the 2023 IEEE 5th International Conference on Civil Aviation Safety and Information Technology (ICCSIT), Dali, China, 14–16 December 2023; pp. 1039–1042.
12. Chen, Z.; Ren, B.; Xu, J.; Wang, K.; Li, G.; Li, Q. Generalized Composite Load Aggregation Considering Different Distributed PV Systems Based on Improved K-means Algorithm. In Proceedings of the 2022 Power System and Green Energy Conference (PSGEC), Shanghai, China, 22–24 August 2022; pp. 296–301.
13. Wen, Y.L.; Hu, Z.C.; You, S.; Duan, X. Aggregate feasible region of DERs: Exact formulation and approximate models. *IEEE Trans. Smart Grid* **2022**, *13*, 4405–4423. [[CrossRef](#)]
14. Xu, T.Y.; Chen, T.; Zhang, X.; Yu, H.; Ya, C. Zonotope based wide-area aggregation and regulation method for distributed resources in virtual power plant. *Autom. Electr. Power Syst.* **2024**, *48*, 139–148.
15. Zhang, K.; Wang, X.; Yang, H.K.; Jiang, C. Distributionally robust optimal scheduling method for power system under flexibility boundaries of data center clusters. *Autom. Electr. Power Syst.* **2024**, *48*, 235–247.
16. Wang, Z.; Liu, D.; Xu, C.Y.; Weng, J.; Chen, F. Status Quo and Prospect of Multi-source Heterogeneous Data Fusion Technology for New Power System. *Electr. Power* **2023**, *56*, 1–15.
17. Su, X.J.; Zhang, C.K.; Fu, Y. Two-stage optimal placement of BESS in an unbalanced active distribution network considering multi-scenario operation. *Power Syst. Prot. Control* **2023**, *51*, 88–97.
18. Zhang, X.Y.; Lin, J.J. Scalable data fusion via a scale-based hierarchical framework: Adapting to multi-source and multi-scale scenarios. *Inf. Fusion* **2025**, *114*, 102694. [[CrossRef](#)]

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.