Energy Reports 16 (2026) 109477

Contents lists available at ScienceDirect

Energy Reports

journal homepage: www.elsevier.com/locate/egyr

Research paper

Revealing flexibility: A cluster-based, data-driven model for unveiling electricity consumption adaptability in households

Hossein Nasrollahi [a,*], Ioannis Lampropoulos [a], Mayke Ruesink [a], Parnian Alikhani [a], Remco C. Veltkamp [b], Wilfried van Sark [a]

[a] Copernicus Institute of Sustainable Development, Utrecht University, Princetonlaan 8a, 3584 CB, Utrecht, The Netherlands
[b] Department of Information & Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC, Utrecht, The Netherlands

ARTICLE INFO

Keywords:
Electricity demand
Clustering
Flexibility
Demand side management
Residential
Peak demand

ABSTRACT

The integration of residential demand-side flexibility into power system operations is essential for addressing grid stress during peak hours and enhancing the deployment potential of decentralized energy systems. This paper introduces a data-driven framework for identifying, evaluating, and aggregating household-level flexibility using only smart meter data. The method combines bottom-up clustering of daily electricity usage patterns with top-down, system-defined peak period detection based on dynamic thresholds. A second-stage clustering extracts feasible historical load-shifting scenarios conditioned on behavioral similarity. Each scenario is evaluated using a multi-criteria matrix that includes peak demand reduction, behavioral deviation, scenario occurrence probability, and electricity cost under dynamic pricing. The approach is demonstrated using two datasets and supports both exhaustive scenario enumeration and scalable aggregation strategies. Results show that the framework captures a spectrum of realistic flexibility options and reveals trade-offs between peak shaving potential and behavioral plausibility. By quantifying flexibility in terms of both technical effectiveness and behavioral deviation, this study offers a replicable and operationally relevant tool for grid operators and aggregators to design targeted, equitable, and dynamic demand-side management interventions.

1. Introduction

1.1. Problem definition

The shift towards a more sustainable and more resilient energy system increased the importance of demand-side flexibility (DSF), especially with the increasing penetration of renewable energy sources like wind and solar in the electricity mix (Golmohamadi et al., 2024; Lampropoulos et al., 2018). While the grid is increasingly challenged with balancing variable generation with demand, households are increasingly being perceived as important system users with the potential to contribute to system flexibility (Hussain et al., 2023), such as peak-demand reduction, grid congestion mitigation, and overall system stability improvement in the distribution networks (Lampropoulos et al., 2024).

However, realizing the potential of residential DSF in practice requires both precise identification of flexible behavior and operationally relevant modeling of its impact. The emergence of smart meters and high-resolution consumption data has created new opportunities to analyze and predict user-level flexibility (Gumz and Fettermann, 2024). Yet, most existing techniques for residential flexibility analysis remain limited in their ability to distinguish between consistent behavioral patterns and exceptional or context-dependent deviations (Li et al., 2021). Several studies analyze residential flexibility using datasets enriched with detailed contextual information (e.g., household surveys and occupant diaries) (Chen et al., 2022; Ramnath et al., 2022). In other studies, high-resolution data, often collected at sub-minute intervals, are used in non-intrusive load monitoring (NILM) methods to disaggregate appliance-level usage to quantify flexibility (Papageorgiou et al., 2025). While smart meters are technically capable of collecting such granular data, in practice, energy providers typically work with data at lower resolutions (e.g., 15-minute or hourly intervals). This is due to practical constraints such as data storage, computational overhead, and privacy concerns (Athanasiadis et al., 2023; McKenna et al., 2012; Staats et al., 2017). While such data can offer valuable behavioral insight, they are rarely available at scale, often rely on voluntary participation, and are difficult to replicate in real-world operational

* Corresponding author.
E-mail address: h.nasrollahi@uu.nl (H. Nasrollahi).

https://doi.org/10.1016/j.egyr.2026.109477
Received 18 December 2025; Received in revised form 1 April 2026; Accepted 20 June 2026
Available online 30 June 2026

Nomenclature

Abbreviation

CVI	Cluster Validity Indices
DSF	Demand Side Flexibility
DSM	Demand Side Management
DSO	Distribution System Operator
FCM	Fuzzy C-Means
HAC	Hierarchical Agglomerative Clustering
HFS	High Flexibility Scenario
LFS	Low Flexibility Scenario
NILM	Non-Intrusive Load Monitoring

Indices

d	Day index, $d \in \{1,\dots,D_h\}$
h	Household index, $h \in \{1,\dots,H\}$
k	Cluster index, $k \in \{1,\dots,K\}$
s	Selected feasible scenario
t	Time interval index

Greek symbols

$\boldsymbol{\mu}_k^{(h)}$	Centroid vector of cluster k for household h
$\boldsymbol{w}^{(\theta)}$	Weight vector indicating importance of each time step in the peak window under threshold P_θ
Δt	Sampling interval (minutes)
$\lambda(t)$	Electricity price at t (€/kWh)
$\mathcal{H}_c^{(h)}$	Set of historical days on which household h was assigned to the same daily profile cluster
$\mathcal{P}^{(h,d,\theta)}$	Set of peak-period load vectors for household h, reference day d, and threshold P_θ (kW)
τ_θ	Duration of peak window at threshold P_θ (hours)
$\mathbf{y}_i$	ith daily profile vector from the aggregated dataset
$\delta^{(h)}$	Euclidean distance (behavioral deviation) between $\phi^{(h)}$ and $\mathbf{p}^{(h,\theta)}_{\text{under-study-day}}$
ΔP_c	Peak reduction (kW)
π^h	Historical occurrence probability of the selected scenario for household h
$\phi^{(h)}$	Selected feasible peak-period load shape for household h (kW)

Symbols

C	Cost (€)
C_k	Set of daily profiles assigned to cluster k
CH	Calinski–Harabasz score
$c_d^{(h)}$	Assigned cluster label for household h on day d
D_h	Set of valid daily profiles for household h
DB	Davies–Bouldin score
$\mathbf{F}^{(\theta,h)}$	Aggregated flexibility matrix
$\mathcal{F}^{(h,d,\theta)}$	Set of feasible peak-period load scenarios for household h, reference day d, and threshold P_θ
K	Number of clusters for a given clustering configuration
K^*	Optimal number of clusters selected for a household
$K_{h,d,\theta}$	Number of clusters determined for household h, reference day d, and threshold P_θ
m	Fuzzification coefficient in FCM
N	Total number of valid daily profiles across all households
n_h	Number of feasible scenarios for household h
$N^{(\theta)}_{\text{combinations}}$	Total number of unique aggregated scenario combinations under threshold P_θ
$p_{d,i}^{(h)}$	Average power consumption of household h on day d at interval i (kW)
$p_h(t)$	Power consumption of household h at time t (kW)
$P_{\text{agg}}(t)$	Aggregated community power demand at time t (kW)
$\mathbf{P}^{(\theta)}_{\text{agg}}$	Aggregated peak-period load vector for all households under threshold P_θ (kW)
P_θ	Candidate peak threshold value (kW)
P_{high}	Upper bound of peak threshold range (kW)
P_{low}	Lower bound of peak threshold range (kW)
$P_{\text{weighted}}(\mathbf{x})$	Weighted peak-period load of vector $\mathbf{x}$ (kWh)
$\mathbf{P}^{(h)}$	Matrix of valid daily profiles for household h ($D_h \times R$)
$\mathbf{p}_d^{(h,\theta)}$	Peak-period load vector of household h on historical day d under threshold P_θ (kW)
SH	Silhouette score
$T^{(P_\theta)}_{\text{peak}}$	Set of time steps within the peak window defined by threshold P_θ (time steps)
$t_e(P_\theta)$	End time index of peak window at threshold P_θ
$t_s(P_\theta)$	Start time index of peak window at threshold P_θ
u_{ik}	Membership degree of profile i in cluster k (FCM)
$\mathbf{Y}$	Concatenated matrix of all daily profiles from all households ($N \times R$)

settings. This highlights an ongoing gap between academic approaches and the operational requirements of system operators.

Moreover, managing flexibility at the community level and particularly across users connected to the same local transformer remains unexplored, despite the evolving needs of modern electricity grids. While recent research has advanced the identification of flexibility potential at the individual level, practical frameworks that support the prioritization and coordination of users based on both their technical capacity and behavioral availability are still limited in deployment (Venizelou et al., 2023; Bouloumpasis et al., 2022). To enable effective and equitable activation of flexibility, distribution system operators (DSOs) require models that offer realistic, quantifiable indicators of user contributions under different system conditions.

To design flexibility interventions that are both behaviorally feasible and system-relevant, it is crucial to (1) characterize user demand in a way that preserves temporal and contextual variation, (2) relate household behavior to specific grid stress events such as peak demands, and (3) evaluate alternative behavioral scenarios based on a range of criteria such as cost, peak impact, and user's behavioral change. Without such models, opportunities to activate flexibility for grid support may be missed, leading to suboptimal grid reinforcement. A robust, multilevel, and context-aware approach to flexibility detection and aggregation can provide benefits for system operators by reducing the need for costly infrastructure upgrades and enabling more efficient, targeted interventions.

1.2. Literature review

Effective demand-side management (DSM) for households relies both on identifying potential load-shifting opportunities and translating these into actionable strategies (Panda et al., 2023; Lampropoulos et al., 2013). While previous studies have mostly focused on detecting high-level demand patterns (Puchegger, 2015; Macedo et al., 2015), there is a growing recognition of the need to develop models that identify behavioral variability, evaluate flexibility, and support coordinated activation at scale (Yilmaz et al., 2019b; Cruz et al., 2024).

A growing number of studies have proposed methods to estimate DSF (Hussain et al., 2023; Luo et al., 2022). These studies range from physics-based models (e.g., models based on the buildings' specific parameters (Aduda et al., 2016)) to data-driven models (e.g., load profiles analysis (Proedrou, 2021) and household surveys (Chen et al., 2022)). Recent work on smart electric vehicle charging, for example, demonstrates substantial cost and emission savings under technical constraints and user preferences (Alikhani et al., 2025), further emphasizing the potential of DSF in practice.

The use of smart meter data to identify flexible consumption patterns has been widely investigated (Shaukat et al., 2021). One common approach involves clustering-based methods (Yilmaz et al., 2019a), which group households by load profile similarities to target DSM programs or to infer their potential for peak load reduction (Wen et al., 2023). In most clustering-based DSM studies, a bottom-up approach is followed where generic DSM strategies are assigned to the identified representative daily load shapes. However, most methods do not quantify flexibility in a way that supports operational planning, such as estimating the shiftability of specific peak segments, incorporating price signals, or scenario-based analysis.

Table 1 provides a comparative overview of recent studies that employ clustering techniques for DSM, focusing on both methodological aspects and operational insights for DSM. A common approach for clustering is using the daily profiles of a household for specific days or seasons. In several studies, normalizing or scaling is conducted as a pre-processing step by dividing the load by the daily sum (Fu et al., 2018; Kwac et al., 2014). Among clustering approaches, K-means remains the most widely used due to its simplicity, scalability, and widespread implementation. However, other studies adopt fuzzy clustering (Czétány et al., 2021; Wen et al., 2023), hierarchical methods (Ahir and Chakraborty, 2022), or density-based clustering (Michalakopoulos et al., 2024) to accommodate overlapping behaviors or complex structures in the data. Only a few works explore ensemble or hybrid methods (Pelekis et al., 2023) that might better address the diversity of load profiles. Regarding the clustering evaluation metrics, internal metrics (e.g., Silhouette score (Oguz et al., 2023; Michalakopoulos et al., 2024), Davies–Bouldin index (Pelekis et al., 2023), or Calinski–Harabasz index (Michalakopoulos et al., 2024)) are used. To align metrics with DSM specific goals Pelekis et al. (2023) introduced peak performance score, though such metrics remain rare.

From a DSM objective perspective, most studies focus on understanding consumption behavior (e.g., Oguz et al. (2023)) and segmenting users for targeted DSM interventions (e.g., Wen et al. (2023)). Relatively few go beyond this to propose operational strategies for flexibility activation. Notably, Meng et al. (2023) adopt a hybrid approach that integrates top-down market signals with bottom-up user data, bridging the gap between household consumption patterns and system-level objectives. This study highlights the potential of hybrid models to establish a more complete and actionable demand response loop.

Crucially, a limitation across the literature is the absence of the capability to prioritize users for flexibility provision. Cluster membership is assumed to be indicative of flexibility potential, but only a few studies provide formal mechanisms for ranking users by flexibility potential. Another limitation across the reviewed literature is the handling of peak periods. Most studies define peak periods as fixed windows or derive them from the average cluster shape. For example, Oguz et al. (2023) define a static peak from 5–9 p.m., while others (e.g., Michalakopoulos et al., 2024) allow peaks to vary across clusters. However, these approaches are often decoupled from actual transformer-level loading (Shepherd and Mohagheghi, 2024). Finally, none of the reviewed studies provide quantitative estimates of flexibility, such as the volume of load shifting possible, the time windows in which it can occur, or the probability of activation. Flexibility remains a qualitative insight rather than a measurable resource. This study addresses these limitations by proposing a unified framework that quantifies flexibility through scenario evaluation, enables user-level prioritization based on flexibility potential, and defines peak periods dynamically using system-level demand thresholds. As a result, flexibility is modeled as a measurable and operational resource.

To guide the development and evaluation of the proposed framework, this study addresses several key research questions. First, how can residential demand-side flexibility be quantified using only standard smart meter data? Second, how can household-level flexibility be represented as a set of behaviorally feasible load-shifting scenarios derived from historically observed consumption patterns? Third, how does the definition of peak periods, through a dynamic threshold range, influence the identification, evaluation, and aggregation of flexibility across households? Finally, what trade-offs emerge between peak demand reduction, behavioral deviation, scenario probability, and cost when flexibility is aggregated at the community level?

1.3. Contributions

This paper proposes a new framework for the characterization of residential DSF from only standard smart meter data. Unlike prior studies that either focus on static clustering or rely on high-resolution or survey-based datasets, our developed methodology enables detailed scenario analysis using data at resolutions typically available to energy providers. The main contributions of this study are as follows:

- We propose a hybrid approach that combines bottom-up clustering of individual household load profiles with a top-down, system-level identification of peak periods, defining dynamic peak thresholds based on aggregated load (rather than relying on predefined time windows).
- We develop a two-stage clustering approach that initially clusters daily profiles using unsupervised clustering, and then refines these results by clustering historical peak segments conditioned on behavioral similarity. This multi-stage structure provides the basis for scenario-based flexibility for each household, given historically observed user behavior.
- We construct a flexibility matrix for every household. This matrix includes indicators such as behavioral deviation, occurrence probability, cost, and peak load reduction potential. This matrix enables comparison of potential flexibility options.
- We demonstrate the applicability of our framework on two publicly available datasets (Ausgrid (Ausgrid, 2025) and Pecan Street (Pecan Street, 2025)), representing both large and small-scale samples, and explain how the method can scale to varying sample sizes, highlighting practical scalability.

This paper is further organized as follows. Section 2 provides a detailed overview of all methodological steps used in this paper. Results of our investigations are presented in Section 3, followed by thorough discussion in Section 4. Section 5 concludes the paper.

Table 1
Summary of clustering methods and DSM applications.

	Clustering methodology details				DSM applications and insights				
Study	Clustering parameter	Data Resolution	Clustering method	Clustering evaluation metric	DSM objective	DSM approach	Household prioritization capability	Peak period type	Flexibility quantification
Oguz et al. (2023)	Average daily profile, Daily profiles	Dataset A: hourly data, Dataset B: 30 min data	K-means	Silhouette score	Characterization of residential energy consumption behaviors. Uncovering seasonal variations in electricity usage influenced by factors like occupant behavior and weather.	Bottom-up	Households can be prioritized for DSM by assigning them to clusters based on their alignment with specific DSM strategies.	Constant (peak period is fixed at 5:00 p.m. to 9:00 p.m.)	No
Michalakopoulos et al. (2024)	Daily profiles	Half-hourly data	K-means, K-medoids, Hierarchical agglomerative clustering, and Density-based spatial clustering	Silhouette score, Davies–Bouldin index, and Calinski-Harabasz index	Segmenting consumers into clusters based on consumption patterns to enhance market participation	Bottom-Up	Households can be prioritized for DR based on their cluster assignments.	Peak periods vary based on the clusters (i.e., early morning, noon, evening, night, late night).	No
Czétány et al. (2021)	Daily profiles	15-minute intervals	K-means, Fuzzy K-means, Agglomerative hierarchical	Elbow method, Silhouette score, Dunn index	Offering insights into consumption patterns across different seasons and building types for targeted DSM programs.	Bottom-Up	Households can be prioritized for DR based on their cluster assignments	Peak periods vary based on the clusters (e.g., evening peak).	No
Pelekis et al. (2023)	Daily net consumption	Hourly intervals data	K-means, K-medoids and Hierarchical agglomerative clustering	Peak performance score (a novel metric), Silhouette score, Silhouette score dynamic time warping, Davies–Bouldin validity index, Peak match score	Design targeted DR programs for a flexible energy community to alter prosumers consumption behavior, with the goals of minimizing reverse power flow at the primary substation.	Bottom-Up	Clusters analyzed for DR potential based on load shape, entropy, and load type.	Peak periods vary based on the clusters.	No
Wen et al. (2023)	Daily profiles	15-minute intervals	Fuzzy C-Means	Silhouette score	Identify potential customers for DSM and net metering programs to reduce peak loads, increase self-consumption, and energy saving.	Bottom-Up	Households can be prioritized for DR based on their cluster assignments	Peak periods vary based on the clusters (i.e., morning peak, pre-peak, peak, and post-peak).	No
Rafiq et al. (2023)	Daily profiles	15-minute intervals	K-means	K-elbow	Develop energy-saving strategies and customize DSM initiatives by identifying consumption patterns and influential factors (e.g., occupancy, dwelling size) for different household groups.	Bottom-Up	Households can be prioritized based on their cluster assignments.	Peak periods vary based on the clusters (e.g., cooling-included households peak at midnight in summer, cooling-excluded households peak in the late afternoon between 12 PM and 8 PM).	No
Meng et al. (2023)	Normalized average hourly consumption over the day	Half-hourly intervals	K-means	Silhouette score, and Davies–Bouldin index	Develop a multiple dynamic pricing framework for demand response to maximize retailer profit, reduce peak demand, and offer tailored prices to different customer segments based on consumption patterns.	Hybrid (Bottom-Up and Top-Down); the model uses individual smart meter data for clustering and demand modeling (bottom-up), while the retailer sets pricing constraints and optimizes tariffs based on wholesale market data and profit goals (top-down).	Households can be prioritized based on their cluster assignments.	Peak periods vary based on the clusters (e.g., morning peak, evening peak).	No
Okereke et al. (2023)	Average daily consumption	Half-hourly intervals	K-means	Mean index adequacy, Clustering dispersion index, Silhouette score, and Davies Bouldin index	Understand electricity consumption behaviors to support flexible demand management	Bottom-Up	Households can be prioritized based on their cluster assignments.	Peak periods vary based on the clusters (e.g., evening peak).	No
Ahir and Chakraborty (2022)	Daily consumption	Half-hourly intervals	Hierarchical clustering	Gap statistics	Identify customer potential for DSM and net metering	Bottom-Up	Households can be prioritized based on their cluster assignments.	Peak periods vary based on the clusters (i.e., morning peak, and evening peak).	No
This study	Daily profiles and peak-period segments	15 min intervals	K-means, Fuzzy C-Means, Hierarchical agglomerative clustering	Silhouette score, Davies-bouldin index, Calinski-Harabasz index (rank-based selection)	Quantify and rank household and community-level flexibility for dynamic DSM planning	Hybrid (Bottom-Up clustering with Top-Down peak detection and scenario aggregation)	A scenario matrix ranks users by behavioral deviation, probability, potential for providing flexibility for different peak thresholds, and impact on peak reduction	Dynamic, threshold-dependent peak periods linked to aggregated load	Flexibility quantified using a scenario analysis and different metrics are measured to analyze each scenario

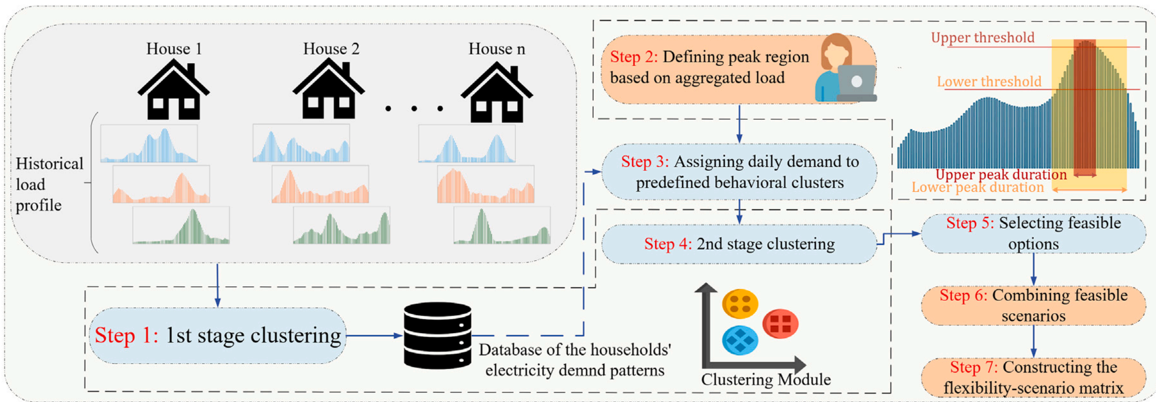

Fig. 1. Illustration of the main idea of the paper. Boxes in orange show actions at the community level, and boxes in blue show actions at each household level.

2. Methodology

2.1. Overview of the methodological framework

This study proposes a multi-stage data-driven methodology to identify, characterize, and evaluate household flexibility in response to DSM objectives. The goal is to move beyond traditional clustering of average load shapes by developing a framework that links household-level routines to peak shaving opportunities via scenario analysis. The approach is visualized in Fig. 1, and proceeds through the following sequential steps:

Step 1: Clustering daily load profiles at the household level. We begin by clustering the historical daily load profiles of each household independently. This step captures individual behavioral patterns and variability across different days. The resulting clusters represent recurring load shape typologies specific to each user and form the basis for behavioral segmentation. These clusters are stored in a database.

Step 2: Defining peak load conditions from aggregated demand. Using the aggregated load profile of the community for a selected day, we identify the peak region not through a fixed threshold, but through a flexible range between a lower threshold (P_{low}) and an upper threshold (P_{high}). This flexible thresholding allows the system operator to examine sensitivity across a range of peak definitions, acknowledging uncertainties in demand forecasting and operational limits.

Step 3: Assigning daily demand to predefined behavioral clusters. Each household's daily load profile on the day under study is matched to one of its previously identified behavioral clusters from Step 1. This step maps the current day's load to historical behavior types, allowing behavioral context to be retained in later analysis.

Step 4: Context-aware clustering of historical peak loads. To analyze feasible demand-side flexibility, we identify past days with a similar behavioral context (same cluster assignment) and extract their peak-period segments, as defined in Step 2. These segments are clustered again, this time focusing solely on peak-period variation under fixed behavioral routines. This produces a set of feasible alternative load profiles for each user.

Step 5: Selecting feasible options. From the identified peak clusters in Step 4, a subset of feasible scenarios is selected for each user, based on their operational feasibility.

Step 6: Combining feasible scenarios. The feasible options are then combined across users to estimate aggregated community flexibility.

Step 7: Constructing the flexibility matrix. For each user and each defined threshold, we construct a matrix characterizing the feasible options obtained from Steps 5 and 6. Each aggregated demand scenario is evaluated across these metrics: (i) Peak load reduction, (ii) Scenario occurrence probability, (iii) Behavioral deviation, and (iv) Scenario cost. This matrix enables the operator to filter and prioritize feasible load modifications at the user level.

The proposed framework is designed to be modular and scalable. It can support both full combinatorial exploration (as shown in smaller datasets) and constrained aggregation (as used in larger systems).

2.2. Load profile clustering

To uncover recurring patterns in electricity usage, we perform clustering on daily power profiles (Rhodes et al., 2014).

Let $\mathbf{p}_d^{(h)}$ represent the daily power consumption profile of household h on day d, sampled over R time intervals (e.g., 96 for 15-minute resolution):

$$\mathbf{p}_d^{(h)} = \left[p_{d,0}^{(h)}, p_{d,1}^{(h)}, \dots, p_{d,R-1}^{(h)}\right], \quad \mathbf{p}_d^{(h)} \in \mathbb{R}_{\geq 0}^{R} \tag{1}$$

where $p_{d,t}^{(h)}$ denotes the average power (in kW) recorded in interval t. For each household, we construct a matrix $\mathbf{P}^{(h)}$, where each row corresponds to a valid daily profile from the set of days D_h with sufficient data:

$$\mathbf{P}^{(h)} = \begin{bmatrix} \mathbf{p}_1^{(h)} \\ \mathbf{p}_2^{(h)} \\ \vdots \\ \mathbf{p}_{D_h}^{(h)} \end{bmatrix}, \quad \mathbf{P}^{(h)} \in \mathbb{R}_{\geq 0}^{D_h \times R} \tag{2}$$

To enable clustering across the entire population, the daily profiles from all households are concatenated into a single dataset:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{P}^{(1)} \\ \mathbf{P}^{(2)} \\ \vdots \\ \mathbf{P}^{(H)} \end{bmatrix} \in \mathbb{R}_{\geq 0}^{N \times R}, \quad N = \sum_{h=1}^{H} D_h \tag{3}$$

The matrix $\mathbf{Y}$ serves as the input to the clustering algorithms discussed in the following sections.

2.2.1. Clustering algorithms

Once the daily energy consumption profiles are constructed, the next step involves grouping similar profiles to uncover common consumption patterns. In this subsection the applied clustering methods are described.

(a) K-Means clustering: K-Means is a centroid-based clustering algorithm that partitions the dataset into K non-overlapping groups, minimizing the variance within each cluster (Arthur and Vassilvitskii, 2007). Given a set of daily energy profiles for each household $\mathbf{P}^{(h)}$ (Eq. (2)), the objective of the K-Means algorithm is to find K cluster centroids that best represent the data. The algorithm aims to solve the following optimization problem:

$$\arg\min_{C} \sum_{k=1}^{K} \sum_{\mathbf{y}_i \in C_k} \left\| \mathbf{y}_i - \mu_k \right\|^2 \tag{4}$$

where:

- C_k is the set of points (daily profiles) assigned to cluster k;
- μ_k is the centroid of cluster C_k, defined as the mean vector of all points within the cluster;
- $\|\mathbf{y}_i - \mu_k\|^2$ is the Euclidean distance between a daily profile $\mathbf{y}_i$ and its assigned cluster centroid.

The algorithm proceeds iteratively through four main steps. It begins by randomly initializing K centroids from the dataset. In each iteration, data points are assigned to the nearest centroid using Euclidean distance, forming clusters. The centroids are then updated by computing the mean of all points in each cluster. This assignment–update process repeats until convergence, typically defined as stabilization of centroids or reaching a maximum number of iterations. The resulting clusters provide a compact representation of recurring load profile patterns.

(b) Fuzzy C-Means (FCM) clustering: FCM is a soft clustering algorithm that allows each data point to belong to multiple clusters with varying degrees of membership. This is particularly useful when the boundaries between clusters are not well-defined, which is often the case in behavioral energy consumption patterns (Bezdek et al., 1984; Chan et al., 2024).

Unlike K-Means, which performs hard assignments of data points to exactly one cluster, FCM assigns to each point a set of membership coefficients that reflect its similarity to each cluster (Ghosh and Dubey, 2013). The algorithm seeks to minimize the following objective function:

$$J_m = \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ik}^m \, \|\mathbf{y}_i - \boldsymbol{c}_k\|^2 \tag{5}$$

where:
- $\mathbf{y}_i \in \mathbb{R}^R$ is the ith daily energy profile;
- $\boldsymbol{c}_k \in \mathbb{R}^R$ is the centroid of cluster k;
- $u_{ik} \in [0, 1]$ is the membership degree of profile i in cluster k;
- $m > 1$ is the fuzzification coefficient that controls the level of cluster overlap.

The membership degrees are constrained such that:

$$\sum_{k=1}^{K} u_{ik} = 1 \quad \forall i \in \{1, 2, \ldots, N\} \tag{6}$$

The FCM algorithm proceeds iteratively. It begins by randomly initializing a membership matrix that assigns each data point a degree of association with every cluster, ensuring that the memberships for each point sum to one. In each iteration, the cluster centroids are updated as weighted averages of the data points, where the weights are determined by the membership degrees raised to the power m. Subsequently, the membership degrees themselves are updated based on the relative distances between each data point and the updated centroids. This process repeats until convergence. The final output consists of a set of cluster centroids and a fuzzy partition matrix, indicating the degree of association between each daily energy profile and each cluster. This soft clustering mechanism is particularly effective for capturing overlapping or transitional energy usage behaviors in residential electricity consumption.

(c) Hierarchical Agglomerative Clustering (HAC): HAC is a bottom-up clustering technique that starts by assigning each data point to its own cluster and then iteratively merges the closest pairs of clusters based on a defined similarity measure (Zepeda-Mendoza and Resendis-Antonio, 2013). The core assumption is that data points in close proximity within the feature space are more similar, and hence, should be grouped together. The result of the HAC process is a dendrogram (a tree-like structure) that visually encodes the sequence and distances at which clusters are merged.

A key strength of HAC lies in its flexibility to define the inter-cluster distance using different linkage criteria. Common linkage strategies include: single linkage, which computes the distance between the closest pair of points from two clusters; complete linkage, which considers the farthest pair; average linkage, which calculates the mean distance between all pairs across clusters; and Ward linkage, which minimizes the increase in total within-cluster variance upon merging. Each criterion has different implications for the shape and compactness of the resulting clusters (Ambroise et al., 2019). The inter-cluster distance between clusters C_i and C_j is formally expressed as:

$$d_{ij} = \text{linkage}(C_i, C_j) \tag{7}$$

where the linkage function is determined by the selected criterion.

Different clustering methods applied in this study provide insights into the characteristics of residential electricity demand. K-Means has a scalable and efficient algorithm and is especially suitable when load shapes are well-separated; however, its performance can be limited by sensitivity to outliers. FCM permits soft clustering with fuzzy assignment of each observation. This allows a better understanding of overlapping behaviors, which are common in behaviorally flexible households. However, compared to K-means, it is more computationally intensive. HAC, on the other hand, is useful where electricity use behaviors form a nested structure. In Section 2.2.2, we discuss the internal evaluation metrics used to assess the quality and validity of clustering results.

2.2.2. Cluster validity indices

The quality of clustering results depends not only on the choice of algorithm but also on the selection of an appropriate number of clusters (Jain et al., 2022). To systematically evaluate clustering performance, we employ a set of internal Cluster Validity Indices (CVIs) that assess the compactness and separation of clusters based on intrinsic properties of the data (Jain et al., 2021). Three commonly used CVIs are applied in this study:

(a) Silhouette (SH) score: The SH score quantifies how well each data point fits within its assigned cluster compared to other clusters (Rousseeuw, 1987). It provides a normalized score for each data point between $[-1, 1]$. $SH \approx 1$ indicates that the data point is well matched to its own cluster and poorly matched to others; $SH \approx 0$ suggests that the point lies on or near the decision boundary between two clusters; and $SH < 0$ shows potential misclassification (Satre-Meloy et al., 2020; Al-Otaibi et al., 2016).

(b) Calinski–Harabasz (CH) score: The CH score (also known as the variance ratio criterion) calculates the ratio between a measure of separation and a measure of compactness (Caliński and Harabasz, 1974). The separation represents the inter-cluster dissimilarities while the compactness characterizes the intra-cluster similarities (Satre-Meloy et al., 2020). A higher CH value suggests that the clusters are dense and well-separated, which corresponds to a desirable clustering structure.

(c) Davies–Bouldin (DB) score: The DB score calculates the average similarity between different clusters and its most similar cluster. Here, similarity is defined as the ratio of within-cluster scatter to between-cluster separation (Davies and Bouldin, 2009). Lower DB values indicate better clustering performance (McLoughlin et al., 2015; Yildiz et al., 2018).

2.2.3. Selection procedure for clustering configuration

To determine the suitable clustering configuration for each user, a rank-based aggregation strategy is used. For each clustering method, clustering is performed over a predefined range of cluster numbers (2 to 10). The internal CVIs are then computed for each configuration. Rankings are applied separately to each index: higher values of SH and CH scores receive lower ranks, while lower values of DB score are preferred. The three ranks are then summed to produce an overall score for each configuration. The final clustering method and optimal number of clusters K^* for each user are selected as those corresponding to the configuration with the lowest total rank. This approach balances the strengths of multiple CVIs.

2.3. Aggregated load profile and threshold range for peak period detection

An adaptive way to define peak demand periods is fundamental for effective DSM planning, especially under high renewable energy integration and demand-side uncertainties (Binyet et al., 2022). Traditionally, peak detection relies on applying a fixed power threshold to the aggregate load profile. In this context, all time intervals exceeding that threshold are considered part of the peak window (Nasir et al., 2021; Rizvi et al., 2023). This approach, while simple, imposes a constraint on a system that can also be analyzed dynamically (Arafat et al., 2025; Silva et al., 2020). In this paper, we propose a more flexible approach based on a range of peak threshold values. This approach is motivated by two key factors: (1) operational uncertainty: grid operators and aggregators often face uncertainty in load forecasts, user responsiveness, and network constraints (Silva et al., 2022), and (2) DSM strategy design: different DSM strategies may perform differently depending on the time a peak is defined and how narrow the peak period is defined (Wynn et al., 2022). This threshold-range approach allows us to investigate how the potential flexibility in the community changes as a function of the selected power threshold.

This step begins by aggregating household load across the community. For each time step t, the total community power demand is calculated as:

$$P_{\text{agg}}(t) = \sum_{h=1}^{H} p_{d,t}^{(h)} \tag{8}$$

Instead of applying a single threshold to detect peak periods, we define a continuous range of candidate peak thresholds bounded by:

- P_{low}: the lower threshold value for considering a time step part of a peak;
- P_{high}: the upper threshold value for considering a time step part of a peak.

The determination of P_{low} and P_{high} should be guided by system operators based on operational considerations. P_{low} may be set to reflect the threshold above which demand begins to place noticeable stress on the distribution grid relative to normal operating conditions, while P_{high} can be chosen to represent critical loading levels approaching network capacity limits. Framing the bounds in this way allows DSOs to tailor the interval $[P_{\text{low}}, P_{\text{high}}]$ to the characteristics of their networks and to capture the relevant range of peak demand conditions for congestion management and planning purposes.

Within this interval $[P_{\text{low}}, P_{\text{high}}]$, we evaluate multiple threshold levels $P_\theta \in [P_{\text{low}}, P_{\text{high}}]$ to examine how the characterization of the peak period varies. For each selected threshold P_θ, the peak period is defined as the set of consecutive time steps t for which the aggregated power $P_{\text{agg}}(t)$ satisfies $P_{\text{agg}}(t) \geq P_\theta$. Based on this condition, we extract three key descriptors: the start time $t_s(P_\theta)$, which marks the first time step exceeding the threshold; the end time $t_e(P_\theta)$, which marks the last respective time step; and the duration of the peak window, calculated as $\tau_\theta = (t_e - t_s + 1) \cdot \Delta t$, where Δt denotes the sampling interval. Fig. 2 illustrates the conceptual structure of this approach.

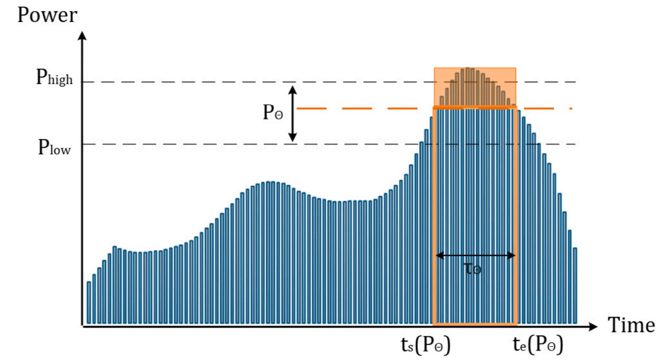

Fig. 2. Illustration of the threshold-range peak detection method. The aggregated load is plotted alongside the lower and upper threshold bounds. For each threshold within the range, the corresponding peak window is extracted, and its start time, end time, and duration are computed.

2.4. Daily load profile assignment to household-specific clusters

In this step, the power consumption profile of the days under study for each household is assigned to one of the cluster centroids obtained specifically for that household in the first clustering stage (Section 2.2). This enables us to represent individual daily behavior using a limited set of recurring load shape typologies and track how these patterns evolve across different days, particularly during high-demand periods.

Let $C^{(h)} = \{\mu_1^{(h)}, \dots, \mu_K^{(h)}\}$ denote the set of centroids derived from the clustering of household h's daily profiles. Each daily profile $y_d^{(h)}$ is assigned to the cluster whose centroid is closest in Euclidean distance:

$$c_d^{(h)} = \arg \min_{k \in \{1,\dots,K\}} \left\| y_d^{(h)} - \mu_k^{(h)} \right\|_2 \tag{9}$$

This procedure is repeated for all households and all relevant days. The result is a mapping from each household–day pair to a discrete cluster label $c_d^{(h)} \in \{1, \dots, K\}$, which compactly summarizes the daily behavior of each user.

2.5. Clustering historical peak loads

Household electricity demand exhibits significant variability, influenced by day-to-day behavioral routines, appliance use, weather, and external signals (Ahire and Patil, 2023; Satre-Meloy et al., 2020). While clustering has traditionally been used to extract general load profiles, recent studies emphasize the value of context-aware clustering, where historical behaviors are grouped not just globally but within similar operational contexts (Kallel et al., 2025; Pullinger et al., 2024; Dent et al., 2014; Xu et al., 2017).

To assess intra-user behavioral variability during peak periods, we cluster historical peak load segments by conditioning on two key dimensions: the daily profile type, which represents the household's routine behavior as determined in Section 2.4, and the community-level peak period associated with each peak threshold P_θ, as defined in Section 2.3 to reflect grid-side stress. This procedure is repeated for each peak threshold $P_\theta \in [P_{\text{low}}, P_{\text{high}}]$ (each threshold defines a different peak window $T_{\text{peak}}^{(P_\theta)}$, which alters the time steps used to extract and cluster historical peak behavior) and each household.

Let $\mathcal{H}_c^{(h)}$ denote the set of historical days on which household h was assigned to the same cluster that is defined in Section 2.4:

$$\mathcal{H}_c^{(h)} = \left\{ j \in \mathcal{D}_{(h)} \mid c_j^{(h)} = c_d^{(h)} \right\} \tag{10}$$

This conditioning restricts the analysis to historically comparable daily profiles.

From each selected historical day j, we extract the segment of the load curve that corresponds to the peak window defined by threshold P_θ, denoted $T_{\text{peak}}^{(P_\theta)} = \{t_s^{(P_\theta)}, \dots, t_e^{(P_\theta)} - 1\}$. The peak-period load vector for user h on historical day d, under threshold P_θ, is defined as:

$$\mathbf{p}_d^{(h,\theta)} = \left[p_h^{(d)}(t) \right]_{t \in T_{\text{peak}}^{(P_\theta)}} \tag{11}$$

The collection of all such vectors yields the peak segment dataset for household h, on reference day d, under threshold P_θ:

$$\mathcal{P}^{(h,d,\theta)} = \left\{ \mathbf{p}_d^{(h,\theta)} \mid d \in \mathcal{H}_c^{(h)} \right\} \tag{12}$$

For each household h, and for each target day d and peak threshold P_θ, we apply the clustering methodology from Section 2.2 to the dataset $\mathcal{P}^{(h,d,\theta)}$. The clustering is performed independently per household to capture user-specific behavioral patterns. The number of clusters is also selected separately for each user-day-threshold combination, following the procedure introduced earlier. This results in clusters that reflect the diversity of peak-period load shapes for each user under consistent behavioral and system conditions.

Each cluster obtained in this step is interpreted as a distinct flexibility scenario for household h, conditioned on reference day d and peak threshold P_θ. We denote the set of resulting scenarios as:

$$\mathcal{S}^{(h,d,\theta)} = \left\{ s_1, s_2, \dots, s_{K_{h,d,\theta}} \right\} \tag{13}$$

where $K_{h,d,\theta}$ is the number of clusters determined for that specific household-day-threshold configuration.

2.6. Identifying feasible scenarios

In this step, we evaluate the flexibility potential of each household by comparing its current peak-period power profile to its historically observed alternative scenarios. In this paper, we incorporated a time-weighted evaluation that reflects the varying criticality of different time steps within the peak period. This allows for the identification of scenarios that meaningfully reduce demand during grid-critical moments, even if accompanied by minor increases elsewhere in the peak window. Each historical scenario $s \in \mathcal{S}^{(h,d,\theta)}$, represents a possible peak-period load shape.

Let $\mathbf{p}^{(h,\theta)}_{\text{under-study}}$ denote the under-study day peak-period power vector of household h under threshold P_θ. To assess the operational desirability of each scenario, we introduce a weight vector $\boldsymbol{w}^{(\theta)}$, where $w_t^{(\theta)} \in [0,1]$ represents the importance of time step $t \in T^{(P_\theta)}_{\text{peak}}$ to the grid operator.

The weighted peak-period load of any vector $\mathbf{x} \in \mathbb{R}^{R_{\text{peak}}}$ is calculated as:

$$P_{\text{weighted}}(\mathbf{x}) = \sum_{t \in T^{(P_\theta)}_{\text{peak}}} w_t^{(\theta)} \cdot x(t) \cdot \Delta t \tag{14}$$

Using this, we define the set of feasible scenarios $\mathcal{F}^{(h,d,\theta)} \subseteq \mathcal{S}^{(h,d,\theta)}$ as those that reduce the weighted power score relative to the current-day profile:

$$\mathcal{F}^{(h,d,\theta)} = \left\{ s_s \in \mathcal{S}^{(h,d,\theta)} \mid P_{\text{weighted}}(s^{(h,d,\theta)}) < P_{\text{weighted}}\left(\mathbf{p}^{(h,\theta)}_{\text{under-study-day}}\right) \right\} \tag{15}$$

This condition ensures that only those scenarios that reduce peak-period demand in a time-sensitive manner are selected for further analysis. This enables the discovery of subtle yet meaningful demand-side flexibility options that may be overlooked by aggregate energy comparisons. The resulting feasible scenario set $\mathcal{F}^{(h,d,\theta)}$ serves as the foundation for constructing user-specific flexibility scenarios, which are explored in Section 2.7

2.7. Aggregating feasible scenarios

Having identified the set of feasible flexibility scenarios for each household under each peak threshold, the aggregated flexibility potential across the full population can be calculated by combining feasible household-level scenarios into community-wide load profiles. This aggregation enables the evaluation of system-level demand shaping under various combinations of user behavior, thereby bridging the gap between local flexibility and operational planning.

For each h, we select one scenario $\phi^{(h)} \in \mathcal{F}^{(h,d,\theta)}$, and construct an aggregated peak load vector by summing the selected scenarios across all users:

$$\mathbf{P}^{(\theta)}_{\text{agg}} = \sum_{h=1}^{H} \phi^{(h)} \tag{16}$$

The total number of possible aggregated scenarios grows rapidly with the size of $\mathcal{F}^{(h,d,\theta)}$. Specifically, if each of the H households has n_h feasible scenarios, the total number of unique combinations is given by:

$$N^{(\theta)}_{\text{combinations}} = \prod_{h=1}^{H} n_h \tag{17}$$

For large-scale datasets, we adopt a bounded aggregation strategy to explore representative extremes of the flexibility space. In particular, we construct two boundary cases:

- **High Flexibility Scenario (HFS)**: For each household, the scenario that yields the lowest weighted peak-period demand is selected.
- **Low Flexibility Scenario (LFS)**: For each household, the scenario closest to the current-day profile (i.e., minimal deviation) is selected.

These two cases form upper and lower bounds on the aggregate flexibility impact for a given threshold. The HFS captures the most aggressive demand reshaping achievable under behaviorally feasible conditions, while the LFS reflects a conservative estimate rooted in minimal behavioral disruption. The resulting profiles are then used in the subsequent section.

2.8. Aggregated flexibility scenario evaluation

To move beyond identifying when demand can be shifted, DSM requires a detailed understanding of how different behavioral alternatives compare in terms of feasibility, cost, and system value. For each pair (h, P_θ), we define the flexibility scenario matrix $\mathbf{F}^{(\theta,h)}$. Each column of this matrix corresponds to an aggregated flexibility scenario, while each row represents a distinct evaluation metric. The four metrics used in this work are introduced below.

Peak load reduction: This metric quantifies the reduction in system peak demand under scenario θ:

$$\Delta P^{(\theta)}_{\text{agg}} = \max_t \mathbf{P}^{(\theta)}_{\text{base}}(t) - \max_t \mathbf{P}^{(\theta)}_{\text{agg}}(t). \tag{18}$$

Average scenario probability: The behavioral plausibility of a scenario is measured through the mean historical occurrence probability of its selected household scenarios:

$$\pi^{(\theta)}_{\text{agg}} = \frac{1}{H} \sum_{h=1}^{H} \pi_s^{(h)}. \tag{19}$$

Average behavioral deviation: To assess how strongly a scenario deviates from observed load profiles, we compute the average Euclidean distance:

$$\delta^{(\theta)}_{\text{agg}} = \frac{1}{H} \sum_{h=1}^{H} \left\| \phi_s^{(h)} - \mathbf{p}^{(h,\theta)}_{\text{curr}} \right\|_2. \tag{20}$$

Aggregated cost: System-level electricity cost under a dynamic pricing signal $\lambda(t)$ is computed as:

$$C^{(\theta)}_{\text{agg}} = \sum_{t \in T^{(P_\theta)}_{\text{peak}}} \lambda(t) \cdot \mathbf{P}^{(\theta)}_{\text{agg}}(t) \cdot \Delta t. \tag{21}$$

2.9. Implementation and dataset-specific strategy

The proposed methodology was applied to two different datasets. The first dataset is the publicly available data of the Pecan Street project (Pecan Street, 2025). The dataset contains data from 25 residential households at a 15-minute resolution in Austin, Texas (USA). The other dataset is the Ausgrid dataset, which includes 300 residential households in and around Sydney, Australia, with a 30-minute resolution (Ausgrid, 2025). Each of these datasets, based on its size, is used for a different purpose in the validation of the proposed framework.

The Pecan Street dataset is used to implement and evaluate the full methodological pipeline. Due to its limited size, it enabled an exhaustive analysis in which all possible combinations of feasible household scenarios were generated and evaluated. With this analysis, the entire flexibility space is explored, and the impact of any joint scenario is assessed across the community. However, this exhaustive approach becomes computationally intractable for larger datasets due to the exponential growth in scenario combinations.

To address scalability, the Ausgrid dataset is used. In this case, we adopted a simplified aggregation strategy that avoids full combinatorial evaluation. For each threshold, two boundary-case aggregate load profiles (LFS and HFS) are considered, as explained in Section 2.7.

All computations were carried out on a standard workstation with an Intel(R) Core(TM) i7-1255U processor (10 cores, 12 threads), 16 GB of RAM. No GPU acceleration was used.

3. Results

In this section, we first use the Pecan Street dataset to demonstrate the model's performance across the initial six steps of the framework. Due to the limited size of the Pecan Street dataset, its results mainly serve as a proof of concept. In contrast, the Ausgrid dataset provides a larger sample, better reflecting the types of datasets energy aggregators would typically work with in real-world applications. Therefore, we base our main results and the extended analysis in step seven on the Ausgrid dataset, as it allows for a more realistic evaluation of the model's performance in practical scenarios.

3.1. Overview of the model performance across steps

Fig. 3 illustrates the outputs from the first six steps of the framework. It presents the results for a summer day, for four randomly selected users. The figure shows the identified flexibility scenarios under three different peak thresholds.

Step 1 illustrates the result of first-stage clustering. For each of the four selected households, the colored lines correspond to the cluster centroids $\boldsymbol{\mu}_k^{(h)}$, which represent typical daily consumption patterns from the household's historical data. The number and shape of clusters vary across households, showing the behavioral diversity. Household 8156 displays six distinct profiles with a wide range of peak times and durations. Some clusters (e.g., 4 and 5) show sharp evening peaks, while some clusters (e.g., 0 and 3) show a longer peak duration. Additionally, smoother demand profiles are observable either during specific time windows or throughout the day (e.g., cluster 1). Household 9922 has only three clusters, with less variation in shape and peak timing. This suggests more stable daily patterns, potentially indicating consistent occupancy and appliance usage. Household 4767 exhibits six distinct clusters, ranging from extremely peaky behaviors (e.g., clusters 0 and 2) to low, flat-load profiles (e.g., cluster 4). Household 5746 has five distinct clusters, with progressive increases in peak demand. The clusters are smoother and show shifts in overall consumption level rather than timing, indicating load volume rather than schedule changes as the main driver of variability. These cluster labels serve as behavioral signatures that condition subsequent flexibility analyses. By anchoring later steps in profile types, the flexibility analyses are interpreted relative to each household's behavioral context.

Step 2 illustrates the threshold-range-based peak detection approach applied to the aggregated load profile. Each subplot shows the same daily load curve with a different horizontal line representing a selected power threshold P_θ, ranging from $P_{\mathrm{low}} = 220\,\mathrm{kW}$ to $P_{\mathrm{high}} = 340\,\mathrm{kW}$. In our model (for the Pecan Street dataset), the peak threshold step size is set to 10 kW. However, for clarity, Fig. 3 includes only three example thresholds. These thresholds span from a conservative value to a critical upper bound. According to the selected threshold, the peak window $T_{\mathrm{peak}}^{(P_\theta)}$ is identified. As P_θ increases, the peak period narrows in duration.

At the low threshold of $P_{\mathrm{low}} = 220\,\mathrm{kW}$, the peak window spans from 12:00 to 23:00. For the intermediate threshold $P_\theta = 280\,\mathrm{kW}$, the peak window spans from 14:45 to 21:00. At the high threshold of $P_{\mathrm{high}} = 340\,\mathrm{kW}$, a narrow peak window from 16:45 to 19:30 is identified.

Step 3 shows the result of assigning the under-study load profile to a specific cluster for each household. This step enables the identification of flexibility scenarios based on historically similar usage patterns. Each graph displays the load profile of a selected household for the under-study day (red line), the most similar cluster centroid to that profile (solid colored line), and all other cluster centroids (gray dashed lines). The peak windows corresponding to the three studied thresholds are also shown on the same graph with yellow color. In several cases (e.g., households 8156 and 4767), the under-study profile aligns closely with one of the cluster centroids, indicating a typical or recurring pattern. In contrast, some households, such as 9922, exhibit more distinct under-study profiles that deviate from previously observed behavior.

Step 4 in the figure visualizes the result of clustering each household's historical demand in the peak window $T_{\mathrm{peak}}^{(P_\theta)}$ (conditioned on the historical days that were assigned to the same cluster selected in the previous step). This enables the identification of user-specific flexibility scenarios during the peak period grounded in actual past behavior. The under-study day profile of the household during the peak window $\mathbf{p}_{\mathrm{curr}}^{(h,d,\theta)}$, shown by the red line, is illustrated with the full set of peak clusters $S^{(h,d,\theta)}$, represented by dashed lines. In most cases (except for household 5766 when $P_\theta = 340\,\mathrm{kW}$), the historical peak demand is represented by two clusters. Moreover, changing P_θ alters the set of identified clusters, as the corresponding peak window shifts.

Step 5 visualizes the process of identifying feasible shifting scenarios. As explained in Section 2.6, a time-weighted evaluation is used to select among the candidate scenarios identified in the previous step. As shown in Fig. 3, it is possible for a feasible scenario to have higher power consumption at certain time steps (specifically at the beginning and end of the peak period), as long as it lowers demand during the most critical time steps. This reflects the model's ability to shift demand rather than simply shaving it. In most cases, there is one feasible flexibility scenario to which the user can switch. Exceptions include Household 9922 at $P_\theta = 340\,\mathrm{kW}$, which has two possible scenarios, and Household 5746 at $P_\theta = 280\,\mathrm{kW}$, where no feasible scenario is found, since the peak demand of the under-study day is already lower than that of all candidate scenarios. These results highlight that the flexibility potential varies across households, depending on both their consumption patterns and the selected peak threshold.

The final Step 6 in our framework aggregates household-level flexibility into aggregated demand scenarios by combining the flexibility scenario sets $\mathcal{F}^{(h,\theta)}$ across all households, for each peak threshold. For a given threshold P_θ, the number of aggregated scenarios is the product of the number of feasible options per household. This number varies substantially across threshold levels. For instance, lower thresholds tend to define broader peak periods with more flexibility potential. The total number of combinations at the lowest threshold is 1440, at the medium threshold is 576, and at the highest threshold is 128.

Due to the exponential growth of possibilities with household count, full analysis becomes computationally intensive, especially when simulating large communities. To make the analysis tractable while preserving key insights, as explained in Section 2.9, we used the two boundary cases HFS and LFS for further analysis, which will be explained in the next session.

3.2. Aggregated demand scenarios across thresholds

In this section, the aggregated impact of household flexibility is explored by comparing two extreme scenarios (HFS and LFS) across a range of peak thresholds. To this end, the model is applied to the Ausgrid dataset. This subsection, along with the following ones, presents the results for a representative day in winter. At the lower bound of the threshold range, $P_{\mathrm{low}} = 440\,\mathrm{kW}$, the resulting peak window spans from 07:00 to 23:30. At the upper bound, $P_{\mathrm{high}} = 1080\,\mathrm{kW}$, the peak window narrows to just 30 min, from 17:15 to 17:45, capturing only the

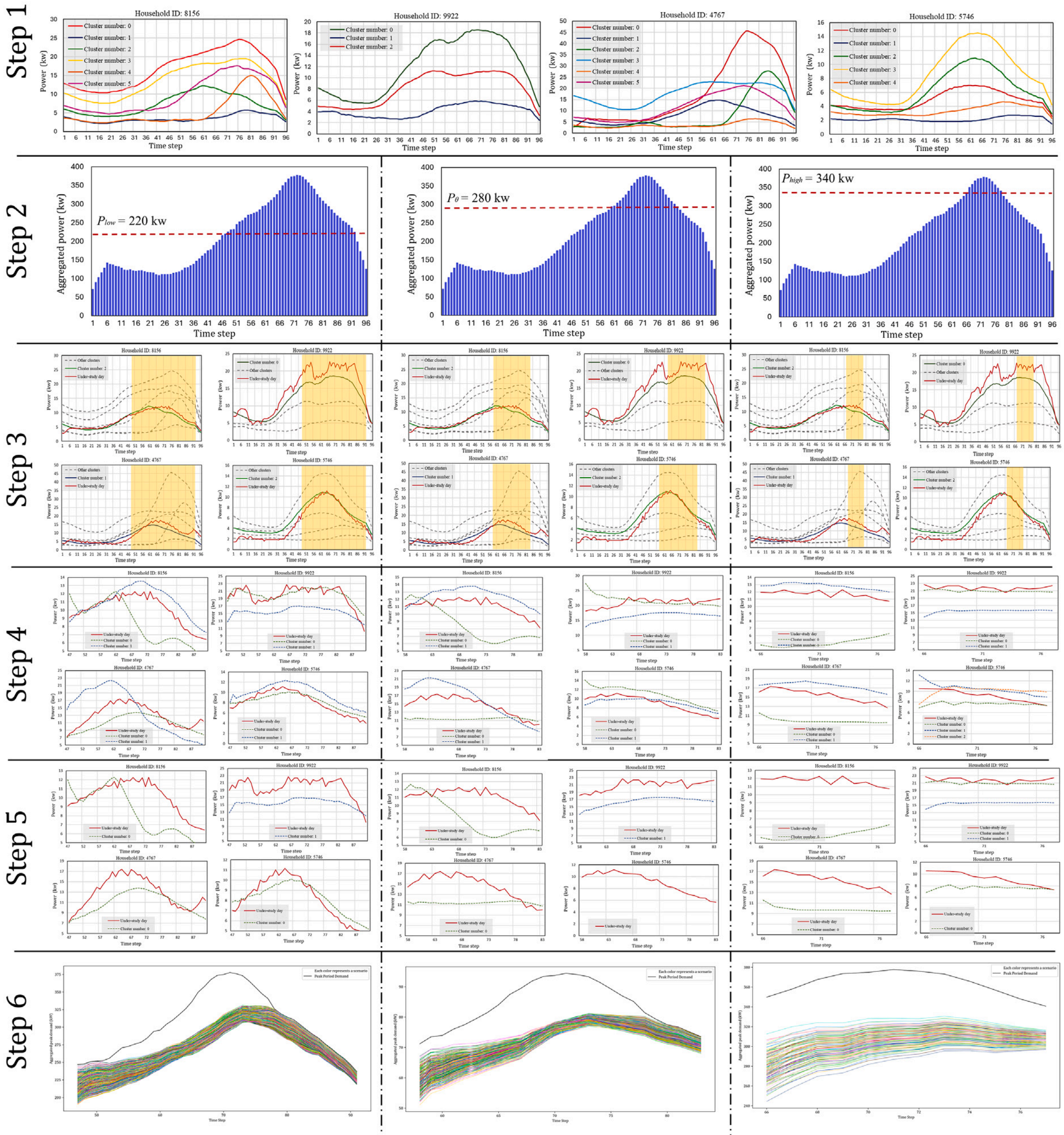

Fig. 3. Overview of the methodological pipeline applied across selected households for the Pecan Street dataset. The analysis is based on 15-minute resolution data. See the text for detailed explanations.

sharpest segment of the evening peak. For each intermediate threshold value P_{θ}, the threshold is increased by two units at a time. However, if a new threshold yields the same peak window as the previous one, it is skipped. In total, this process resulted in 62 unique threshold values being evaluated for the selected day. Fig. 4 presents the aggregated demand for the two flexibility scenarios across these peak thresholds.

The top panel in Fig. 4(a) shows a wide spread in the colored demand curves under HFS. This shows a considerable variation in the aggregated demand depending on the threshold level, particularly at narrower peak periods. The greatest gap between the colored lines and the baseline appears to occur at $P_{\text{high}} = 1080\,\text{kW}$, suggesting that with an increase in peak threshold (shorter duration of change in the behavior), more peak reduction can be achieved. In contrast, the top panel in Fig. 4(b), representing LFS, shows a much narrower band of colored lines. These lines mostly follow the shape of the black curve, with only minor downward deviations. This indicates that under the LFS, aggregated demand remains closer to the original baseline, even when threshold settings change. While some thresholds result in minor reductions during narrower peak periods, the effect is considerably smaller than in the HFS.

Turning to the heatmaps in the lower panel of Fig. 4(a) again displays more structural changes in demand. A vertical band of lower demand emerges between approximately 17:00 and 19:30, particularly for thresholds above 1030 kW. This region represents the most effective demand shifting. The blue zone extends more broadly for lower thresholds, indicating that longer peak windows allow households more time to shift their load, resulting in lower demand over a broader portion of the day. It is also noteworthy that as the threshold decreases, the corresponding peak window expands, eventually encompassing a large share of the daily load profile. While this broader window captures

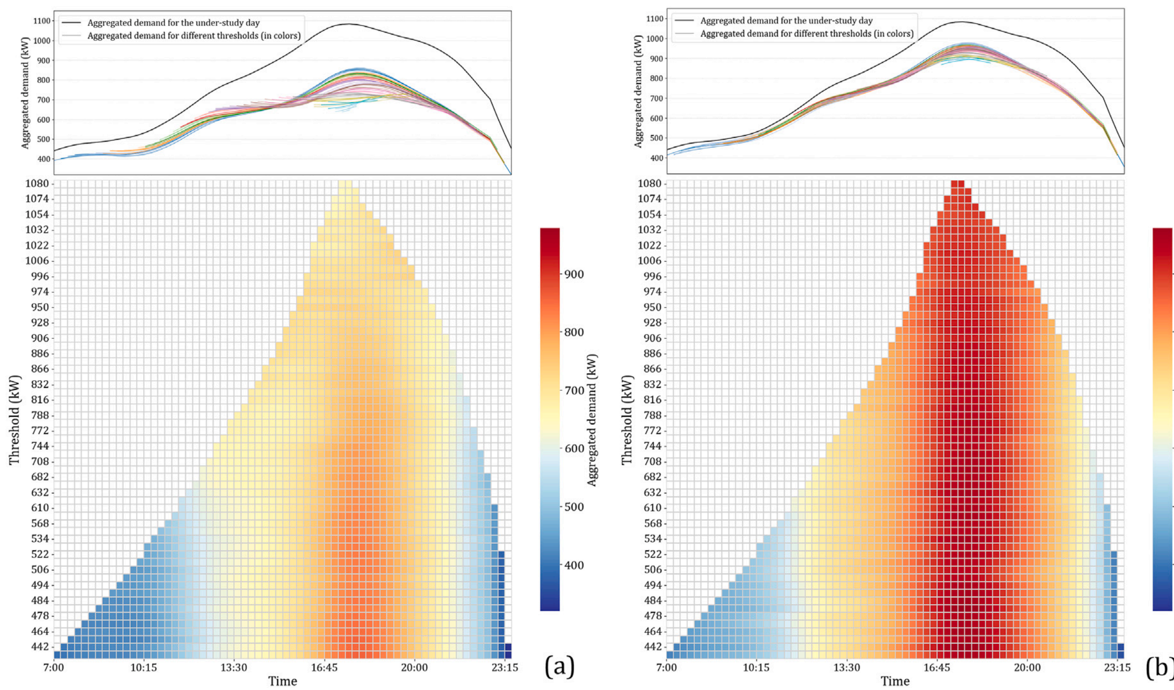

Fig. 4. Aggregated demand under two boundary scenarios for a winter day. (a) HFS. (b) LFS. Top: Overlaid load curves for different peak thresholds. The black line represents the baseline demand on the selected day. Bottom: Heatmaps of aggregated demand as a function of time and threshold.

more time intervals where load shifting is technically feasible, the result is a dispersion of flexibility across too many time steps. Consequently, the potential for concentrated demand reduction during system-critical periods diminishes. Although total energy shifted may increase slightly, the effectiveness of flexibility in mitigating sharp peaks plateaus. This issue is discussed in more detail in Section 3.3. In contrast, the heatmap in the lower panel of Fig. 4(b) remains predominantly in the red-orange color range, especially during the late afternoon and early evening hours. As mentioned before, this subplot also confirms that demand is largely unchanged despite variations in the threshold.

Another notable contrast between HFS and LFS lies in the temporal profile of demand reduction. In the HFS, the flattening effect occurs predominantly during the central peak hours, while in the LFS, the overall load shape is preserved and only marginal shifts occur. Besides, in the lower panel of Fig. 4(a), a triangular zone of reduced demand can be observed that is aligned with the peak, while the lower panel of Fig. 4(b) shows only a gradual softening of demand at the edges of the peak period. This implies that the aggregated benefit of flexibility is maximized when user-level shifts are concentrated during the most critical timeframes.

Another distinction between the two scenarios is that in HFS, the presence of diverse flexibility resources facilitates not only the shaving of peak demand, but also the redistribution of load to less critical hours. This is evident from the lighter blue and yellow tones appearing outside the critical peak periods, particularly under intermediate threshold levels, suggesting that demand is not merely curtailed but actively shifted across time. In contrast, the low flexibility scenario (Fig. 4(b)) exhibits a more constrained pattern of flexibility utilization. Here, the reductions in demand are concentrated around the peak hour, with little change in demand outside that interval.

In summary, while both scenarios show consistent structural behaviors in response to varying thresholds (e.g., the timing of the critical peak, the duration of the peak period, and the direction of demand reshaping), the extent of demand reduction is significantly higher in the HFS. In the next section, the result of the model for reducing maximum demand is analyzed in more detail.

3.3. Peak period sensitivity and flexibility envelope

Fig. 5 illustrates the peak period sensitivity and the flexibility envelope boundaries across varying threshold values P_θ. The x-axis represents the duration of the detected peak period D_θ in minutes, ranging from 30 to 990 min, while the y-axis shows the power level in kilowatts. The dashed black line denotes the dynamic threshold P_θ, which decreases as longer peak durations are considered. The blue line shows the maximum aggregated demand observed in the HFS, while the purple line represents the minimum LFS across different thresholds.

Three distinct regions can be observed, separated by the position of the threshold compared to the envelope boundaries defined by HFS and LFS. In the green window, where $\text{LFS}_{\max} < P_\theta$, the threshold lies above the LFS, meaning that users can remain within their low-flexibility bounds without violating the peak constraint. It is noteworthy that in this window, the maximum demand for HFS shows a gradual increase from 647 kW to 728 kW, while the LFS remains relatively flat around 910 kW.

In the orange-shaded region, where $\text{HFS}_{\max} < P_\theta < \text{LFS}_{\max}$, the threshold lies between the bounds of the two scenarios and demand flexibility becomes more constrained. The HFS curve rises at a faster rate in this window, and the LFS curve also starts increasing more notably.

In the red-shaded region, where $P_\theta < \text{HFS}_{\max}$, the threshold lies below both flexibility bounds. This condition enforces stricter constraints on all users, requiring more significant modifications in demand profiles to meet the threshold requirement. Accordingly, the gap between the HFS and LFS curves narrows, and both curves show a converging trend. The HFS curve continues its upward trend, reaching 859 kW, while the LFS curve steadily increases approximately to 961 kW.

Fig. 5 provides critical insight into how the maximum demand level evolves under different peak period configurations. The observed trend indicates that, while shorter peak durations enable substantial maximum demand reductions, this potential diminishes with longer durations. Therefore, this figure illustrates the feasible region for peak demand targets under varying operational constraints. It is important to note that this analysis was performed only once, without separating the

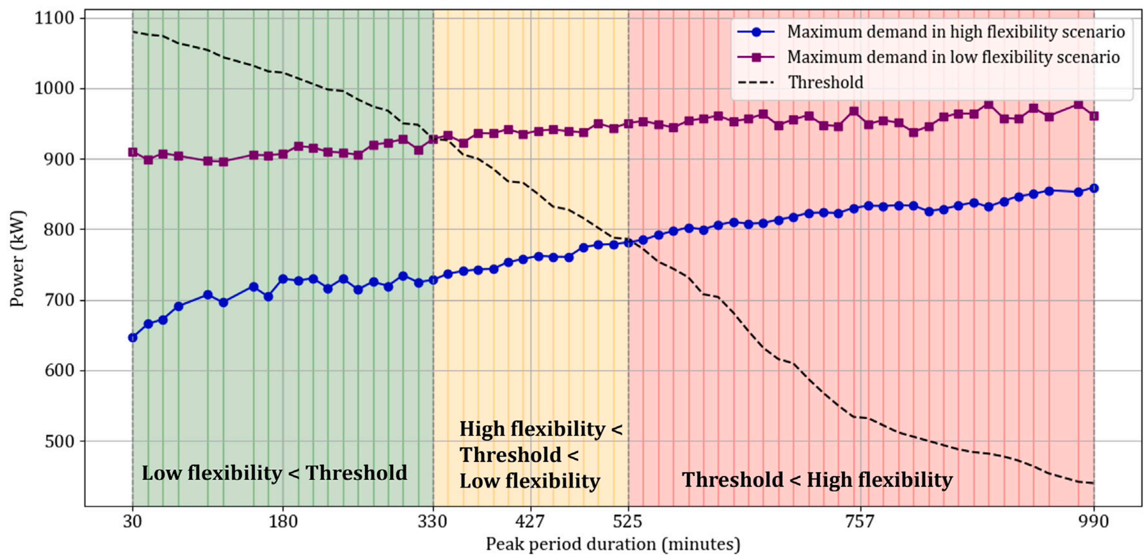

Fig. 5. Aggregated maximum demand across varying peak period durations. The blue line shows the maximum demand in the HFS, while the purple line indicates the maximum demand in the LFS. The black dashed line represents the dynamic threshold level P_θ. Shaded regions denote three threshold selection regimes: green (both scenarios satisfy the threshold condition), yellow (only the high flexibility scenario satisfies the threshold), and red (neither scenario satisfies the threshold).

HFS and LFS. Since inflexibility is defined based on the feasibility of the flexibility scenario's existence per user and threshold, it is determined independently of which boundary scenario is considered.

3.4. User-level inflexibility patterns

Understanding the distribution of inflexibility across users is key to designing effective demand response strategies and characterizing the limits of household adaptability. This section investigates individual user behavior across varying peak thresholds, focusing on how often each user is classified as inflexible within the 62 evaluated thresholds. A user is defined as *inflexible* at a given threshold if no feasible shifting scenario exists for that user under the corresponding peak window.

Fig. 6 shows a histogram of users categorized by how many times they are identified as inflexible across all 62 thresholds. The x-axis groups users into three inflexibility frequency bands: 0–20, 21–41, and 42–62. The largest portion of the population (over 200 users) is in the 0–20 category, indicating that most users are only occasionally inflexible. This suggests a generally high level of flexibility in the system, with many users capable of adjusting their demand under most threshold conditions. However, 52 fall into the 42–62 range, meaning that about 17% of households are unable to provide flexibility in most cases and can be considered as a limitation for the system.

While the histogram provides an aggregate perspective, Fig. 7 presents the same ranges dynamically across increasing threshold levels. As the threshold value rises from 440 kW to 1080 kW, the number of users in each inflexibility band varies.

The trends remain relatively stable for all ranges up to a threshold of 900 kW. Beyond 900 kW the number of inflexible users in the most flexible category (0–20) begins to increase, while the number of users in the most inflexible category (42–62) decreases. The intermediate group (21–41) remains largely constant throughout. This behavior indicates that user-level flexibility is not static; users may transition between inflexibility categories depending on the level of the threshold. In particular, higher thresholds appear to enable a subset of previously inflexible users to become flexible, and vice versa. Despite these category shifts, the total number of inflexible users (i.e., the sum across all three ranges) remains relatively stable across all thresholds, ranging between 65 and 77 users, with an average of 71. This implies that while the composition of inflexible users changes, the aggregate inflexibility level of the system remains largely constant.

3.5. Aggregated flexibility capacity and peak reduction potential

In this section, the aggregated flexibility capacity and the resulting peak demand reduction are investigated across different peak period durations and flexibility scenarios. Fig. 8 summarizes these results by jointly presenting the aggregated flexibility capacity (left axis, in kW) and the peak demand reduction (right axis) as functions of the peak period duration.

Fig. 8 illustrates two key outcomes across increasing peak period durations: the total aggregated flexibility capacity and the corresponding peak demand reduction. In both scenarios, the capacity increases as the peak period duration lengthens. This is expected, as a longer peak window contains more time steps in which flexible load can be activated. In the HFS, the aggregated flexibility capacity increases from 879 kW at the shortest duration of 30 min to 11,665 kW at the maximum duration of 990 min. A similar trend, though at a consistently lower magnitude, is observed in the LFS, where the capacity grows from 350 kW to 6254 kW across the same duration range.

In contrast, the peak demand reduction exhibits a decreasing trend as the peak period expands. In the high flexibility scenario, the initial reduction reaches 435 kW for the shortest window but declines to 223 kW for the longest duration. The low flexibility scenario shows a similar pattern, decreasing from 172 kW to 121 kW.

Comparing the two scenarios, the high flexibility setting consistently outperforms the low flexibility case in both aggregated flexibility capacity and peak demand reduction. The gap between the scenarios is most pronounced for shorter peak durations (e.g., below 300–500 min), where concentrated flexibility within a narrow window enables highly effective peak shaving. As the peak period becomes longer, however, the available capacity must be distributed over a wider interval, reducing its instantaneous impact and flattening the peak shaving effect.

These patterns indicate that the system's ability to reduce peak load is governed not only by the total available aggregated flexibility capacity but also by its temporal alignment with the peak intensity. This highlights the importance of selecting an appropriate peak

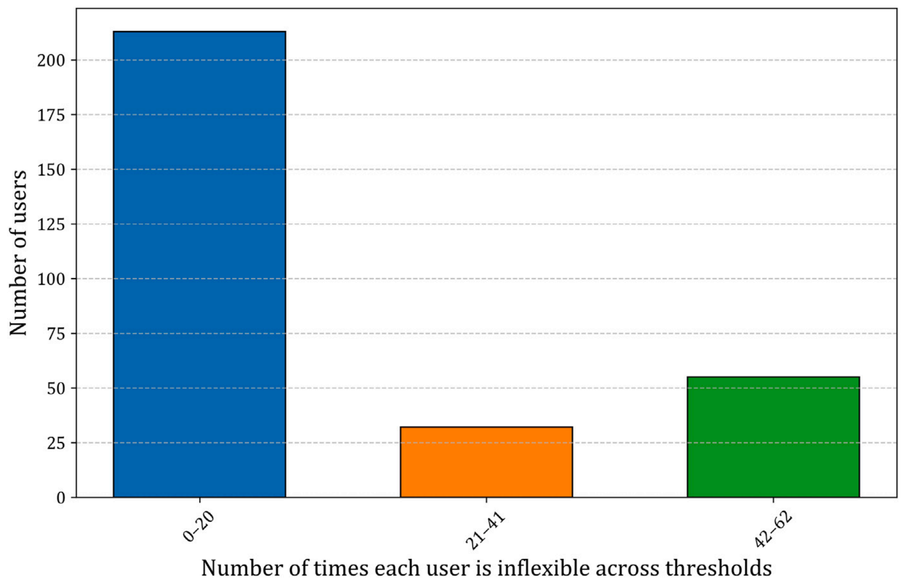

Fig. 6. Histogram showing the total number of users within each inflexibility range (0–20, 21–41, 42–62) over all 62 evaluated thresholds.

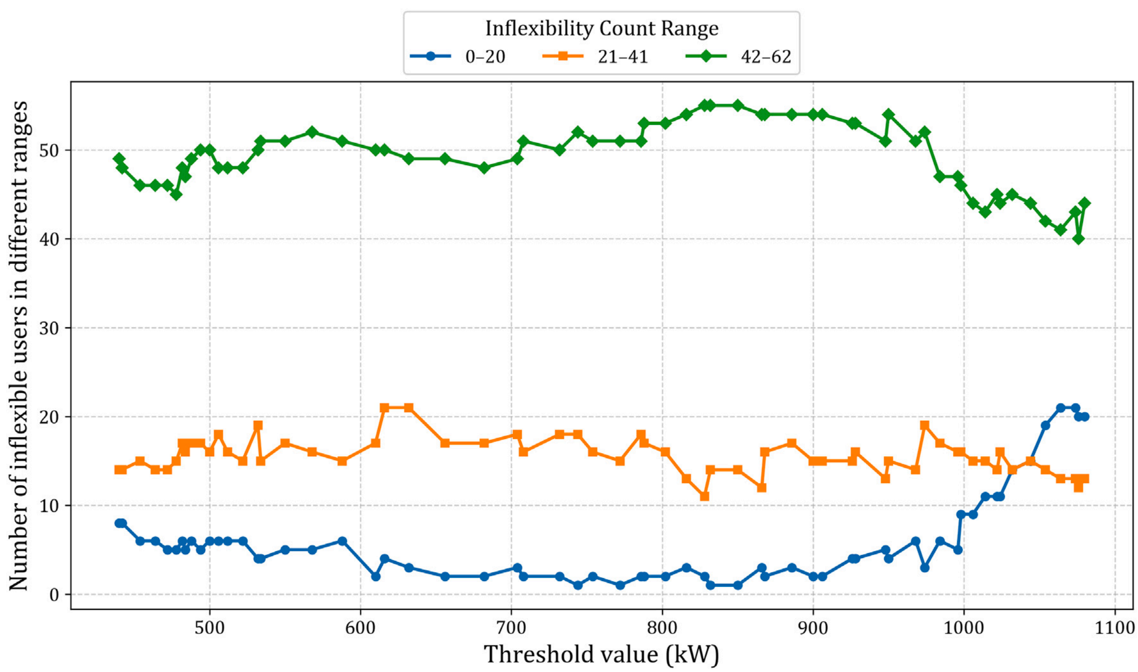

Fig. 7. Threshold-wise distribution of users according to the number of times they are identified as inflexible. Each line represents the count of users who are inflexible within one of three frequency ranges (0–20, 21–41, 42–62), across increasing peak thresholds.

period duration to balance flexibility utilization and peak reduction performance.

3.6. Cost analysis

This subsection examines the economic implications of flexibility activation by evaluating how reshaped load profiles translate into cost savings under dynamic pricing. While previous analyses focused on aggregate demand, flexibility volumes, and peak reduction potential, here we assess the resulting electricity costs under varying peak period durations.

Fig. 9 presents the aggregated cost outcomes for both high and low flexibility scenarios across varying peak durations. The figure uses two vertical axes: the left axis displays the total daily electricity cost during the peak period in euros, while the right axis shows the percentage reduction relative to the under-study day. The green and brown lines represent the costs of the HFS and LFS, respectively, while the blue line shows the baseline cost of the original profile. Shaded regions indicate the cost difference between each scenario and the benchmark, and the lines with triangle markers denote the relative cost savings.

As the peak period duration increases, the absolute cost of both flexibility scenarios rises steadily, reflecting the expansion of the peak window. The HFS consistently achieves greater cost reductions, reaching 41% for the shortest durations and decreasing to 23% for the longest. In contrast, the LFS yields smaller savings, starting at 16% and gradually declining to 12%.

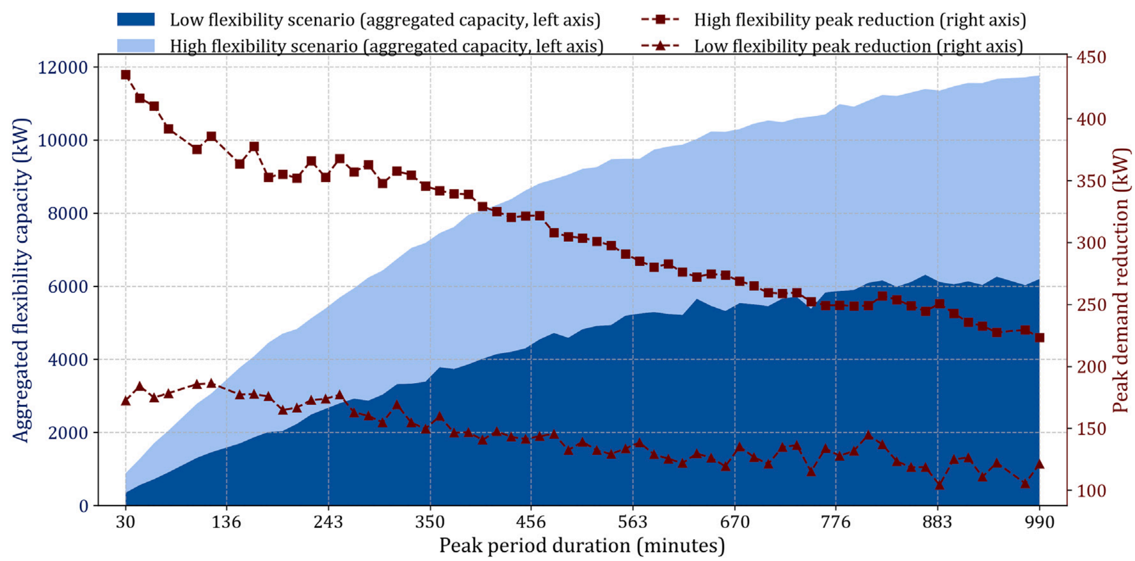

Fig. 8. Aggregated flexibility capacity (shaded, left axis) and peak demand reduction (dashed lines, right axis) across varying peak period durations.

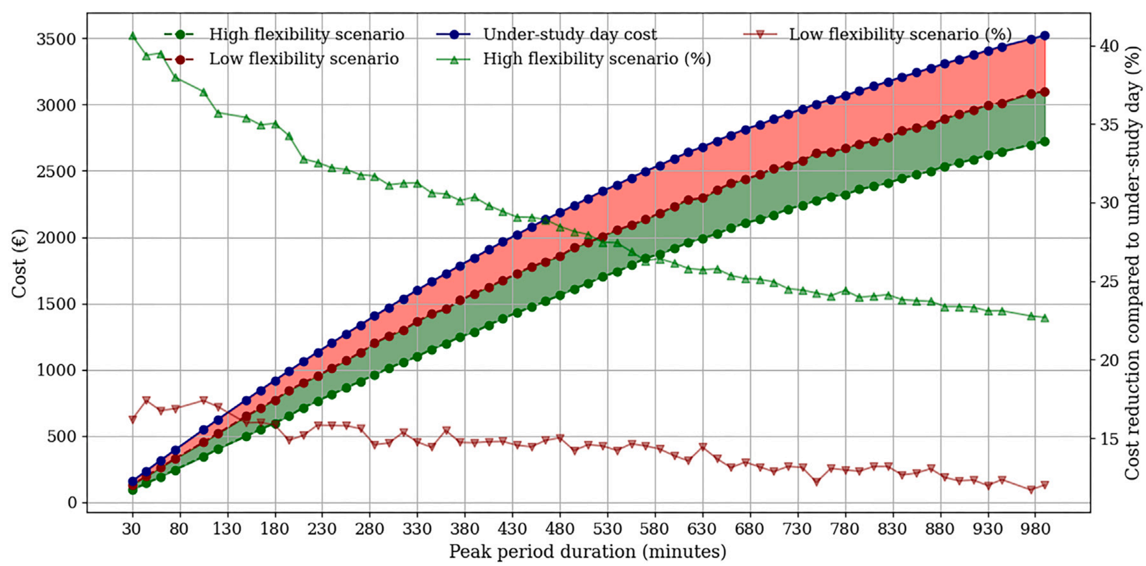

Fig. 9. Aggregated cost and relative cost reduction across varying peak period durations. Solid lines show the total cost for HFS (green), LFS (brown), and the original profile (blue). The lines with triangle markers indicate the percentage cost reduction for each scenario (right axis).

The widening cost gap between the HFS and LFS for longer peak windows highlights the added value of behavioral diversity and extensive shifting capacity. However, the narrowing percentage gap in cost savings indicates diminishing marginal returns when extending the peak window, suggesting that economic benefits saturate beyond a certain point.

3.7. Behavioral plausibility analysis

This subsection evaluates the behavioral feasibility of the flexibility scenarios by examining two metrics: the average scenario probability and the average behavioral deviation across the 62 evaluated peak thresholds. Fig. 10 reports these two indicators, with solid lines representing average scenario probability (left axis) and dashed lines representing average deviation (right axis). Green curves correspond to the HFS, while red curves correspond to the LFS.

The HFS shows a higher average probability in most peak durations, along with a consistently larger behavioral deviation. The average probability in the HFS increases from 0.63 at a 30-minute window to 0.69 at 990 min, with a more pronounced rise in mid-range thresholds (500–880 min).

The LFS exhibits a different dynamic: although it starts at a higher probability (0.66), its probability fluctuates notably, dipping to around 0.62 in several intervals (e.g., 400–600 min). This variability indicates that no universal trend describes how scenario probability evolves with changing peak duration; instead, it is highly situation-specific and depends on the structure of each household's consumption profile.

Behavioral deviation further differentiates the two scenarios. In the HFS, deviation begins around 4 kW and increases to nearly 39 kW, reflecting the substantial shifts required to meet peak reduction targets under longer windows. In the LFS, deviation is lower (2–21 kW) and grows more gradually.

Taken together, the cost and behavioral plausibility analyses underscore the complex trade-offs involved in designing effective flexibility scenarios. While longer peak periods tend to offer greater flexibility capacity, and in this case, slightly improved behavioral alignment (particularly under HFS), they also result in increased deviation from typical user behavior and diminished marginal economic returns. The

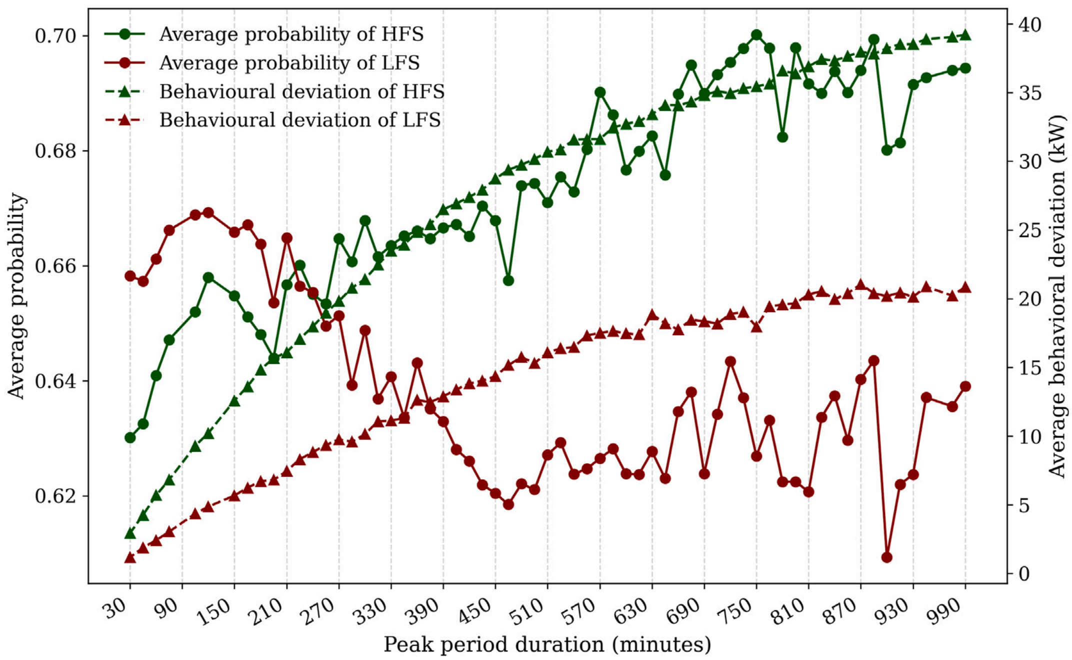

Fig. 10. Scenario-level probability and behavioral deviation across varying peak durations. Solid lines represent average scenario probability (left axis), while dashed lines represent average behavioral deviation (right axis), for high (green) and low (red) flexibility scenarios.

observed variability in scenario probability emphasizes that behavioral feasibility cannot be assumed to follow a consistent trend across all thresholds or households. Instead, it depends on how well specific scenario configurations align with historical usage patterns at the individual level. These findings suggest that flexibility interventions must be tailored not only to system-level goals such as peak reduction and cost savings but also to the diverse behavioral profiles of end-users.

4. Discussion

This study introduced and validated a modular, multi-stage framework for analyzing residential DSF using only standard smart meter data. The results show that: First, the structure of user demand (captured via the two-stage clustering approach) significantly affects the feasibility and diversity of flexibility scenarios. Across the evaluated peak thresholds, the HFS consistently achieved substantially higher peak shaving, with reductions in maximum aggregated demand on the order of 220–430 kW, compared to 90–180 kW under the LFS (Fig. 5). In parallel, the aggregated flexibility capacity under HFS was approximately twice that of the LFS for most peak period durations, ranging from about 1–12 MW for HFS compared to 0.5–6 MW for LFS, indicating a markedly larger pool of deployable flexibility (Fig. 8). These system-level gains translated into stronger economic benefits, with relative cost reductions of approximately 23%–41% for HFS compared to 12%–16% for LFS across peak period durations (Fig. 9). However, these improvements came at the cost of increased behavioral deviation, which rose to approximately 4–39 kW for HFS, compared to 2–21 kW for LFS (Fig. 10), revealing a clear and systematic trade-off between behavioral adherence and achievable system benefits.

Second, the analysis across 62 dynamic peak thresholds demonstrated that the definition of peak periods has a critical impact on all dimensions of flexibility. Shorter peak windows enabled substantial reductions in maximum demand that were feasible under both HFS and LFS, indicating operating regimes where peak constraints can be met without requiring pronounced behavioral deviation. As the peak window expanded, meeting the same threshold became feasible only under the HFS, while the LFS increasingly violated the constraint, reflecting a growing reliance on more behaviorally demanding adjustments. Beyond a critical peak duration, the threshold fell below the flexibility envelope of both scenarios, rendering peak reduction infeasible regardless of scenario choice. This progression highlights how peak period design governs not only the magnitude of achievable peak shaving, but also the boundary between feasible, selectively feasible, and infeasible system targets, as captured by the flexibility envelope.

Third, household-level flexibility is both heterogeneous and strongly context-dependent. While most users exhibit some degree of adaptability under certain peak definitions, a persistent minority of households (approximately 17%) remains inflexible across the majority of evaluated thresholds, effectively setting a hard lower bound on achievable system flexibility. Importantly, the classification of users as flexible or inflexible is not fixed: users transition between inflexibility categories as peak period duration and threshold levels change, even though the total number of inflexible users remains relatively stable. This indicates that residential flexibility is not a binary attribute, but a conditional and scenario-dependent property that emerges from the interaction between user behavior and system constraints. By capturing these dynamics explicitly, the proposed framework enables a more nuanced identification of structurally limiting users, supports scenario-level prioritization, and allows system operators to distinguish between inherent inflexibility and threshold-induced infeasibility.

4.1. Interpretation and implications

From a grid operator's perspective, the suggested framework in this study provides a concrete tool for quantifying and ranking flexibility options. The flexibility scenario matrix developed in this study enables an operational interpretation of user flexibility by integrating multiple evaluation dimensions at the community level (possible flexible load shapes, maximum demand, peak shaving possibilities, available flexibility, and cost) and household level (potential for providing flexibility, behavioral deviation, and probability). This multi-dimensional evaluation can support more informed decision-making when it comes to selecting which users to engage, what level of flexibility to request, and accordingly how to structure financial or behavioral incentives. Grid operators can assign different weights to each metric depending on their operational priorities to derive context-specific flexibility strategies. The model also facilitates sensitivity testing with respect to peak thresholds, enabling grid operators to fine-tune peak duration definitions in ways that optimize effectiveness without imposing undue burden on households. Besides, the HFS and the LFS bounds introduce an area for which a balance between ambitious to conservative system-level gains can be expected.

For real-world applications, one of the most relevant findings is the uneven distribution of flexibility across the population and the dynamic nature of user adaptability, which suggests that one-size-fits-all DSM strategies may be inefficient or even counterproductive. Personalized or localized scenario design becomes essential, not only to respect user constraints but also to unlock hidden flexibility in segments of the population that may otherwise appear static under conventional analysis.

Another contribution of this study is its treatment of behavior change as a continuous, scenario-based process rather than a discrete, single-moment shift. Instead of assuming a binary transition from one load profile to another, the framework explores a range of historically grounded, behaviorally plausible alternatives. This approach makes the trajectory of behavior change traceable over time. One key advantage of this view is that it enables a more nuanced understanding of user adaptation. Moreover, this continuous framing helps prevent the emergence of unintended secondary peaks, which can arise when demand is simply shifted from one point in time to another. Therefore, the framework supports smoother transitions and more distributed adjustments, reducing the risk of rebound effects and ensuring grid stability.

An important practical consideration in applying this framework is its computational scalability, particularly given the multi-stage architecture and the large number of household-specific operations. In the most exhaustive setup (as explained in Section 3.1), the number of community-level scenario combinations grows exponentially with the number of users who have multiple feasible options. For instance, even with 25 households and an average of two feasible scenarios per user, the total number of joint configurations can reach 2^{25}. To manage this, we used full combinatorial analysis only on small datasets (e.g., Pecan Street) and applied bounded aggregation rules on a larger dataset, which includes 300 households, and 62 peak thresholds are considered for analysis. Runtime for evaluating all users' scenarios averaged 20 min per peak threshold, while aggregate scenario construction under bounding rules required one minute per threshold. The model is also highly parallelizable, as steps 1 to 4 can be performed independently for each household. Thereby, the methodology is computationally feasible for real-world applications.

When viewed in the context of existing residential DSM literature, the results of this study highlight several distinctions. Many prior approaches rely on fixed peak definitions, appliance-level controllability assumptions, or static user classifications derived from daily load clustering. While such methods provide useful segmentation insights, they often obscure how flexibility feasibility evolves under changing system conditions. By contrast, the proposed framework demonstrates that flexibility emerges as a conditional and threshold-dependent property, shaped jointly by user behavior and system-level peak characteristics. The identification of flexibility envelopes, feasibility regimes, and a persistent yet shifting subset of inflexible users extends beyond conventional clustering-based DSM approaches, offering a system-relevant and behaviorally grounded perspective on residential demand-side flexibility using only standard smart meter data.

The results of this study are useful for several stakeholders in the energy system. Most importantly, grid operators can use the proposed framework to identify, quantify, and prioritize demand-side flexibility under different peak conditions. Specifically, it can support operational decision-making for congestion management and the design of targeted demand response programs. In addition, policymakers and market designers working on the development of incentive mechanisms can benefit from these results, as they provide insights into how flexibility varies across users and conditions.

4.2. Methodological considerations and future directions

The proposed framework offers several methodological strengths that make it a robust tool for household flexibility assessment. Its stepwise design (from clustering to scenario evaluation) is highly interpretable. Besides, the approach is adaptable across datasets with different household counts and consumption patterns, suggesting generalizability to other contexts. Its modular structure also enables customization of threshold strategies, evaluation criteria, and aggregation rules, supporting integration with diverse DSM schemes.

Compared with similar clustering-based studies, this study extends the literature in important ways. Previous studies use clustering to segment users or classify daily load profiles, often relying on fixed peak definitions and failing to explicitly quantify flexibility in operational terms (Wen et al., 2023; Czétány et al., 2021; Michalakopoulos et al., 2024; Pelekis et al., 2023; Meng et al., 2023). In such approaches, cluster membership is typically used as a proxy for flexibility potential, without evaluating the feasibility, likelihood, or system impact of alternative consumption patterns (Wen et al., 2023; Czétány et al., 2021; Pelekis et al., 2023; Meng et al., 2023). A first key distinction lies in the ability to quantify flexibility directly from standard smart meter data. While many existing approaches either rely on high-resolution measurements, appliance-level disaggregation, or supplementary survey data (Chen et al., 2022; Ramnath et al., 2022; Papageorgiou et al., 2025), the proposed framework demonstrates that meaningful flexibility indicators can be extracted from commonly available data. A second distinction is the representation of flexibility as a set of behaviorally feasible scenarios derived from historically observed consumption patterns. In contrast to conventional methods that assign static flexibility labels or rely on representative load profiles (Wen et al., 2023; Czétány et al., 2021), the proposed framework captures intra-user variability and preserves behavioral realism. A third contribution is the use of a dynamic threshold range to define peak periods. Existing studies often rely on fixed peak windows or cluster-dependent peak definitions, which are not directly linked to system-level conditions (Michalakopoulos et al., 2024; Oguz et al., 2023). By contrast, the proposed approach explicitly connects peak identification to aggregated demand, allowing the analysis of how flexibility changes under different system stress levels and enabling a more system-relevant evaluation. Finally, the framework introduces a multi-criteria evaluation of flexibility scenarios that jointly considers peak reduction, behavioral deviation, scenario probability, and cost. While previous studies typically focus on a single objective (such as peak shaving (Wen et al., 2023)) or cost minimization (Meng et al., 2023), this approach reveals the trade-offs between technical effectiveness and behavioral feasibility, providing a more comprehensive basis for decision-making. Nevertheless, several limitations and assumptions should be acknowledged. The entire framework relies on the assumption that historically observed consumption behaviors provide a valid basis for constructing future flexibility scenarios. This assumption underpins not only the estimation of scenario probabilities, but also the identification of feasible load shapes. While this allows the model to operate without requiring intrusive sensing or forecasting, it may not always hold in the short term, particularly in contexts where household behavior is evolving due to changes such as new technologies, dynamic pricing schemes, or shifting social norms, or where underlying drivers of consumption (e.g., occupancy patterns or appliance-level usage) are not explicitly captured. However, the model is designed to adapt: as new behaviors become more frequent and are captured in the historical dataset, they are gradually incorporated into the clustering and scenario construction process. Furthermore, although the model can generate all feasible scenario combinations, computational limitations make full enumeration infeasible for large communities, and scalable optimization or approximation methods may be required.

Future research could extend the framework in several directions. First, incorporating an adaptive incentive system that targets users individually (based on their historical patterns and flexibility potential) would enhance the responsiveness and effectiveness of DSM strategies. Such a system could iteratively adjust incentive levels in response to user engagement, helping to align flexibility targets with actual user preferences and constraints. Second, integrating real-time user feedback would improve the behavioral plausibility of scenario selection by reducing dependence on historical proxies. To support real-time applications and larger populations, future work could also explore optimization strategies such as reinforcement learning, which can help balance competing goals like cost minimization, peak shaving, and users' behavior deviation. Finally, the current implementation does not explicitly model flexible appliances such as heat pumps, home batteries, or electric vehicles. Incorporating these technologies into the framework (with properly decided assumptions) would extend its relevance to increasingly electrified households and allow for more comprehensive modeling of residential flexibility potential.

Overall, the proposed framework offers several advantages, including its data-driven nature, its reliance only on standard smart meter data, its ability to represent flexibility through behaviorally grounded scenarios, and its modular and scalable structure. At the same time, it has several limitations. These include its dependence on historical consumption patterns, the absence of explicit modeling of underlying behavioral drivers and emerging flexible technologies, and computational challenges associated with large-scale scenario aggregation. These shortcomings can be addressed through the integration of additional contextual data sources, the incorporation of adaptive or learning-based methods, and the development of more scalable optimization or approximation strategies, as discussed above.

5. Conclusions

This study introduced a modular, data-driven framework for evaluating residential demand-side flexibility using only standard smart meter data. By combining behavioral clustering with dynamic, system-level peak detection, the framework enables the identification of feasible flexibility scenarios and the quantification of both user-level and aggregated system impacts without relying on intrusive end-use modeling or strong controllability assumptions. The results demonstrate that demand heterogeneity plays a central role in shaping flexibility outcomes: high-flexibility scenarios enable substantially larger peak reductions and economic benefits, while low-flexibility scenarios provide more conservative but behaviorally stable contributions.

Across the evaluated peak definitions, aggregated peak load reductions ranged from approximately 16%–40% under short peak windows, where reductions are temporally concentrated, to 8%–20% under long peak windows, where flexibility is distributed over a longer duration, highlighting the strong dependence of achievable peak shaving on peak period design. At the same time, the analysis revealed that flexibility is inherently conditional on system constraints rather than a fixed household attribute. The identification of flexibility envelopes, feasibility regimes, and a persistent minority of structurally inflexible users highlights both the opportunities and the limits of residential flexibility under realistic operating conditions.

From an operational perspective, the framework can be directly integrated into distribution system operator workflows, supporting congestion management, day-ahead demand response planning, and the formulation of quantified, scenario-based inputs for flexibility market participation. The results further suggest that effective demand response programs should move beyond one-size-fits-all designs, instead combining broad, low-deviation participation from the majority of households with targeted, incentive-driven engagement of high-flexibility users to unlock larger system-level benefits.

Overall, the proposed framework provides a scalable and operationally relevant approach for translating residential consumption data into actionable flexibility insights, supporting more informed grid planning, user-centric demand response design, and the development of flexibility-oriented market mechanisms. This work opens several directions for future research. These include incorporating adaptive, user-specific incentive mechanisms based on the outputs of the proposed framework, integrating real-time user feedback to enable real-time applications, and developing optimization approaches to balance competing objectives such as cost, peak reduction, and behavioral deviation.

CRediT authorship contribution statement

Hossein Nasrollahi: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Conceptualization. **Ioannis Lampropoulos:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Mayke Ruesink:** Writing – review & editing, Methodology. **Parnian Alikhani:** Writing – review & editing, Methodology, Conceptualization. **Remco C. Veltkamp:** Writing – review & editing, Supervision, Conceptualization. **Wilfried van Sark:** Writing – review & editing, Conceptualization.

Declaration of generative AI in scientific writing

During the preparation of this work, the authors used ChatGPT in order to assist with text editing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This research was partially funded by the Dutch Research Council (NWO), project Long-term consumer and community empowerment in energy applications through inclusive Game design, Artificial Intelligence, and system Modelling (GAIM), grant number KICH1.ED03.20.022.

Declaration of competing interest

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

Data availability

Publicly available datasets were analyzed in this study.

References

Aduda, K., Labeodan, T., Zeiler, W., Boxem, G., Zhao, Y., 2016. Demand side flexibility: Potentials and building performance implications. Sustain. Cities Soc. 22, 146–163.

Ahir, R.K., Chakraborty, B., 2022. A novel cluster-specific analysis framework for demand-side management and net metering using smart meter data. Sustain. Energy, Grids Netw. 31, 100771.

Ahire, P.G., Patil, P.D., 2023. CAC-WOA: context aware clustering with whale optimization algorithm for knowledge discovery from multidimensional space in electricity application. Clust. Comput. 27 (1), 499–513. http://dx.doi.org/10.1007/s10586-023-03965-4.

Al-Otaibi, R., Jin, N., Wilcox, T., Flach, P., 2016. Feature construction and calibration for clustering daily load curves from smart-meter data. IEEE Trans. Ind. Inform. 12 (2), 645–654.

Alikhani, P., Brinkel, N., Schram, W., Lampropoulos, I., van Sark, W., 2025. Multi-objective optimization of electric vehicle charging considering market coupling in north-western Europe. Transp. Res. Part D: Transp. Environ. 146, 104829.

Ambroise, C., Dehman, A., Neuvial, P., Rigaill, G., Vialaneix, N., 2019. Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics. Algorithms Mol. Biol. 14 (1), 22.

Arafat, I., Guerra, E.C., Meng, J., 2025. Uncertainty impact on aggregator performance for peak shaving. IEEE Access 13, 103733–103747. http://dx.doi.org/10.1109/ACCESS.2025.3579635.

Arthur, D., Vassilvitskii, S., 2007. k-means++: The Advantages of Careful Seeding. Tech. rep., pp. 1027–1035.

Athanasiadis, C.L., Papadopoulos, T.A., Kryonidis, G.C., 2023. A review of smart meter data analytics for distribution network applications. In: 2023 IEEE Belgrade PowerTech. IEEE, pp. 1–6.

Ausgrid, 2025. Ausgrid dataset. URL https://www.ausgrid.com.au/Industry/Our-Research/Data-to-share.

Bezdek, J.C., Ehrlich, R., Full, W., 1984. FCM: The fuzzy c-means clustering algorithm. Comput. Geosci. 10 (2–3), 191–203.

Binyet, E., Chiu, M.-C., Hsu, H.-W., Lee, M.-Y., Wen, C.-Y., 2022. Potential of demand response for power reallocation, a literature review. Energies 15 (3), 863.

Bouloumpasis, I., Mirzaei Alavijeh, N., Steen, D., Le, A.T., 2022. Local flexibility market framework for grid support services to distribution networks. Electr. Eng. 104 (2), 401–419.

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Comm. Statist. Theory Methods 3 (1), 1–27.

Chan, K.Y., Yiu, K.F.C., Kim, D., Abu-Siada, A., 2024. Fuzzy clustering-based deep learning for short-term load forecasting in power grid systems using time-varying and time-invariant features. Sensors 24 (5), 1391.

Chen, X., Zanocco, C., Flora, J., Rajagopal, R., 2022. Constructing dynamic residential energy lifestyles using latent Dirichlet allocation. Appl. Energy 318, 119109.

Cruz, C., Tostado-Véliz, M., Palomar, E., Bravo, I., 2024. Pattern-driven behaviour for demand-side management: An analysis of appliance use. Energy Build. 308, 113988.

Czétány, L., Vámos, V., Horváth, M., Szalay, Z., Mota-Babiloni, A., Deme-Bélafi, Z., Csoknyai, T., 2021. Development of electricity consumption profiles of residential buildings based on smart meter data clustering. Energy Build. 252, 111376.

Davies, D.L., Bouldin, D.W., 2009. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. (2), 224–227.

Dent, I., Craig, T., Aickelin, U., Rodden, T., 2014. Variability of behaviour in electricity load profile clustering; who does things at the same time each day? In: Advances in Data Mining. Applications and Theoretical Aspects: 14th Industrial Conference, ICDM 2014, St. Petersburg, Russia, July 16-20, 2014. Proceedings 14. Springer, pp. 70–84.

Fu, X., Zeng, X.-J., Feng, P., Cai, X., 2018. Clustering-based short-term load forecasting for residential electricity under the increasing-block pricing tariffs in China. Energy 165, 76–89.

Ghosh, S., Dubey, S.K., 2013. Comparative analysis of k-means and fuzzy c-means algorithms. Int. J. Adv. Comput. Sci. Appl. 4 (4).

Golmohamadi, H., Golestan, S., Sinha, R., Bak-Jensen, B., 2024. Demand-side flexibility in power systems, structure, opportunities, and objectives: A review for residential sector. Energies 17 (18), 4670.

Gumz, J., Fettermann, D.C., 2024. User's perspective in smart meter research: State-of-the-art and future trends. Energy Build. 308, 114025.

Hussain, S., Lai, C., Eicker, U., 2023. Flexibility: Literature review on concepts, modeling, and provision method in smart grid. Sustain. Energy, Grids Netw. 35, 101113.

Jain, M., AlSkaif, T., Dev, S., 2021. Validating clustering frameworks for electric load demand profiles. IEEE Trans. Ind. Inform. 17 (12), 8057–8065.

Jain, M., Jain, M., AlSkaif, T., Dev, S., 2022. Which internal validation indices to use while clustering electric load demand profiles? Sustain. Energy, Grids Netw. 32, 100849.

Kallel, S., Amayri, M., Bouguila, N., 2025. Clustering and interpretability of residential electricity demand profiles. Sensors 25 (7), 2026.

Kwac, J., Flora, J., Rajagopal, R., 2014. Household energy consumption segmentation using hourly data. IEEE Trans. Smart Grid 5 (1), 420–430.

Lampropoulos, I., van den Broek, M., van der Hoofd, E., Hommes, K., van Sark, W., 2018. A system perspective to the deployment of flexibility through aggregator companies in the Netherlands. Energy Policy 118, 534–551.

Lampropoulos, I., Kling, W.L., Ribeiro, P.F., van den Berg, J., 2013. History of demand side management and classification of demand response control schemes. In: 2013 IEEE Power & Energy Society General Meeting. pp. 1–5. http://dx.doi.org/10.1109/PESMG.2013.6672715.

Lampropoulos, I., Nasrollahi, H., Belinskiy, A., Werning, S., Veltkamp, R.C., van Sark, W., 2024. A serious game connecting the smart meters of residential users for demand side management: Lessons learned in the preparation stage. In: 2024 IEEE International Smart Cities Conference (ISC2). IEEE, pp. 1–6.

Li, H., Wang, Z., Hong, T., Piette, M.A., 2021. Energy flexibility of residential buildings: A systematic review of characterization and quantification methods and applications. Adv. Appl. Energy 3, 100054.

Luo, Z., Peng, J., Cao, J., Yin, R., Zou, B., Tan, Y., Yan, J., 2022. Demand flexibility of residential buildings: definitions, flexible loads, and quantification methods. Engineering 16, 123–140.

Macedo, M.N., Galo, J.J., Almeida, L.A., Lima, A.C., 2015. Typification of load curves for DSM in Brazil for a smart grid environment. Int. J. Electr. Power Energy Syst. 67, 216–221.

McKenna, E., Richardson, I., Thomson, M., 2012. Smart meter data: Balancing consumer privacy concerns with legitimate applications. Energy Policy 41, 807–814.

McLoughlin, F., Duffy, A., Conlon, M., 2015. A clustering approach to domestic electricity load profile characterisation using smart metering data. Appl. Energy 141, 190–199.

Meng, F., Ma, Q., Liu, Z., Zeng, X.-J., 2023. Multiple dynamic pricing for demand response with adaptive clustering-based customer segmentation in smart grids. Appl. Energy 333, 120626.

Michalakopoulos, V., Sarmas, E., Papias, I., Skaloumpakas, P., Marinakis, V., Doukas, H., 2024. A machine learning-based framework for clustering residential electricity load profiles to enhance demand response programs. Appl. Energy 361, 122943.

Nasir, T., Bukhari, S.S.H., Raza, S., Munir, H.M., Abrar, M., Muqeet, H.A.u., Bhatti, K.L., Ro, J.-S., Masroor, R., 2021. Recent challenges and methodologies in smart grid demand side management: State-of-the-art literature review. Math. Probl. Eng. 2021 (1), 5821301.

Oguz, E., Tekdemir, I.G., Gozel, T., 2023. A demand-side management assessment of residential consumers by a clustering approach. Electr. Eng. 105 (1), 493–508.

Okereke, G.E., Bali, M.C., Okwueze, C.N., Ukekwe, E.C., Echezona, S.C., Ugwu, C.I., 2023. K-means clustering of electricity consumers using time-domain features from smart meter data. J. Electr. Syst. Inf. Technol. 10 (1), 2.

Panda, S., Mohanty, S., Rout, P.K., Sahu, B.K., Parida, S.M., Samanta, I.S., Bajaj, M., Piecha, M., Blazek, V., Prokop, L., 2023. A comprehensive review on demand side management and market design for renewable energy support and integration. Energy Rep. 10, 2228–2250.

Papageorgiou, P.G., Christoforidis, G.C., Bouhouras, A.S., 2025. NILM in high frequency domain: A critical review on recent trends and practical challenges. Renew. Sustain. Energy Rev. 213, 115497.

Pecan Street, 2025. Pecan Street, Dataport: The World's Largest Energy Data Resource, Pecan Street Inc. URL https://dataport.pecanstreet.org/.

Pelekis, S., Pipergias, A., Karakolis, E., Mouzakitis, S., Santori, F., Ghoreishi, M., Askounis, D., 2023. Targeted demand response for flexible energy communities using clustering techniques. Sustain. Energy, Grids Netw. 36, 101134.

Proedrou, E., 2021. A comprehensive review of residential electricity load profile models. IEEE Access 9, 12114–12133.

Puchegger, M., 2015. Electric load behaviour and DSM potential of office buildings. Energy Build. 100, 43–49.

Pullinger, M., Zapata-Webborn, E., Kilgour, J., Elam, S., Few, J., Goddard, N., Hanmer, C., McKenna, E., Oreszczyn, T., Webb, L., 2024. Capturing variation in daily energy demand profiles over time with cluster analysis in british homes (September 2019 – August 2022). Appl. Energy 360, 122683. http://dx.doi.org/10.1016/j.apenergy.2024.122683.

Rafiq, H., Manandhar, P., Rodriguez-Ubinas, E., Barbosa, J.D., Qureshi, O.A., 2023. Analysis of residential electricity consumption patterns utilizing smart-meter data: Dubai as a case study. Energy Build. 291, 113103.

Ramnath, G.S., Muyeen, S., Kotecha, K., 2022. Household electricity consumer classification using novel clustering approach, review, and case study. Electronics 11 (15), 2302.

Rhodes, J.D., Cole, W.J., Upshaw, C.R., Edgar, T.F., Webber, M.E., 2014. Clustering analysis of residential electricity demand profiles. Appl. Energy 135, 461–471.

Rizvi, S.S.H., Chaturvedi, K.T., Kolhe, M.L., 2023. A review on peak shaving techniques for smart grids. AIMS Energy 11 (4), 723–752.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Satre-Meloy, A., Diakonova, M., Grünewald, P., 2020. Cluster analysis and prediction of residential peak demand profiles using occupant activity data. Appl. Energy 260, 114246.

Shaukat, M.A., Shaukat, H.R., Qadir, Z., Munawar, H.S., Kouzani, A.Z., Mahmud, M.P., 2021. Cluster analysis and model comparison using smart meter data. Sensors 21 (9), 3157.

Shepherd, M.M., Mohagheghi, S., 2024. Demand flexibility: A review of quantification methods, models, and required data. Appl. Sci. 14 (23), 11276.

Silva, C., Faria, P., Vale, Z., Corchado, J., 2022. Demand response performance and uncertainty: A systematic literature review. Energy Strat. Rev. 41, 100857.

Silva, B.N., Khan, M., Han, K., 2020. Futuristic sustainable energy management in smart environments: A review of peak load shaving and demand response strategies, challenges, and opportunities. Sustainability 12 (14), 5561.

Staats, M., de Boer-Meulman, P., Van Sark, W., 2017. Experimental determination of demand side management potential of wet appliances in the Netherlands. Sustain. Energy, Grids Netw. 9, 80–94.

Venizelou, V., Tsolakis, A.C., Evagorou, D., Patsonakis, C., Koskinas, I., Therapontos, P., Zyglakis, L., Ioannidis, D., Makrides, G., Tzovaras, D., Georghiou, G.E., 2023. DSO-aggregator demand response cooperation framework towards reliable, fair and secure flexibility dispatch. Energies 16 (6), 2815.

Wen, H., Liu, X., Yang, M., Lei, B., Cheng, X., Chen, Z., 2023. An energy demand-side management and net metering decision framework. Energy 271, 127075.

Wynn, S.L.L., Pinthurat, W., Marungsri, B., 2022. Multi-objective optimization for peak shaving with demand response under renewable generation uncertainty. Energies 15 (23), http://dx.doi.org/10.3390/en15238989.

Xu, S., Barbour, E., González, M.C., 2017. Household segmentation by load shape and daily consumption. In: Proc. of. ACM SigKDD. pp. 1–9.

Yildiz, B., Bilbao, J.I., Dore, J., Sproul, A., 2018. Household electricity load forecasting using historical smart meter data with clustering and classification techniques. In: 2018 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia). IEEE, pp. 873–879.

Yilmaz, S., Chambers, J., Patel, M.K., 2019a. Comparison of clustering approaches for domestic electricity load profile characterisation-implications for demand side management. Energy 180, 665–677.
Yilmaz, S., Weber, S., Patel, M., 2019b. Who is sensitive to DSM? Understanding the determinants of the shape of electricity load curves and demand shifting: Socio-demographic characteristics, appliance use and attitudes. Energy Policy 133, 110909.
Zepeda-Mendoza, M.L., Resendis-Antonio, O., 2013. Hierarchical agglomerative clustering. In: Encyclopedia of Systems Biology. Springer, pp. 886–887.