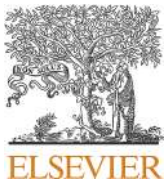
A scalable demand-side energy management control strategy for large residential districts based on an attention-driven multi-agent DRL approach

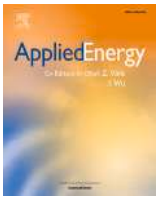
Applied Energy 393 (2025) 125993



Contents lists available at ScienceDirect

Applied Energy

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



A scalable demand-side energy management control strategy for large residential districts based on an attention-driven multi-agent DRL approach

Sabrina Savino [a,b], Tommaso Minella [c], Zoltán Nagy [d], Alfonso Capozzoli [a,*]

[a] Department of Energy (DENERG), TEBE Research Group, BAEDA Lab, Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin, 10129, Italy
[b] Department of Control and Computer Engineering, Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin, 10129, Italy
[c] Department of Mathematical Sciences, Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin, 10129, Italy
[d] The University of Texas at Austin, Austin, 301 E. Dean Keeton St., ECJ 4.200, 78712, USA

HIGHLIGHTS

- AAC-MADRL is an attention-based MADRL algorithm for scalable energy management.
- Effective energy management for districts with up to 100 buildings.
- A parameterized reward explores cooperative, competitive, and mixed scenarios.
- AAC-MADRL outperforms SAC, MARLISA, and RBC in energy storage and cost management.
- Benchmarking in the CityLearn environment for reproducibility and scalability.

ARTICLE INFO

Keywords:
Demand-side management (DSM)
Renewable energy integration
Attention mechanisms
Large district energy management
Scalability
Cooperative multi-agent reinforcement learning

ABSTRACT

The growing penetration of renewable energy sources holds great potential for decarbonizing the building energy sector. However, the intermittent and unpredictable nature of renewable generation poses significant challenges to grid stability and energy integration. Demand-side management (DSM) has emerged as a promising solution, leveraging demand flexibility to align energy consumption with periods of peak renewable generation and mitigate grid instability. To fully harness this flexibility, energy coordination across multiple buildings is essential, enabling participation in flexibility markets and optimizing energy management at district level. This paper introduces attention-actor-critic multi-agent deep reinforcement learning (AAC-MADRL), an actor-critic algorithm built upon the centralized training with decentralized execution (CTDE) framework, enhanced with attention mechanisms with the aim of enabling scalable, coordinated, and autonomous DSM in residential districts. A parameterized reward structure allows systematic testing under different cooperation scenarios – fully cooperative, competitive, and mixed – highlighting the conditions where AAC-MADRL outperforms other deep reinforcement learning (DRL) approaches, including decentralized and non-attention-based cooperative methods. Evaluated through winter and summer scenarios in districts across Alameda County, California (73 buildings) and Texas County (100 buildings) using the CityLearn platform, AAC-MADRL demonstrates substantial improvements. AAC-MADRL achieves energy cost reductions of up to 18 % in Texas and 12.5 % in California compared to the rule-based controller. Additionally, it improves self-sufficiency by 6 %–10.5 % during periods of limited solar generation and significantly reduces peak demand. The algorithm also exhibited superior computational efficiency, with deployment times 40.5 % faster than decentralized DRL and 62.5 % faster than cooperative non-attention-based DRL approaches on average.

* Corresponding author.
Email address: alfonso.capozzoli@polito.it (A. Capozzoli).

https://doi.org/10.1016/j.apenergy.2025.125993
Received 8 January 2025; Received in revised form 31 March 2025; Accepted 24 April 2025
Available online 14 May 2025

Acronyms

AAC-MADRL Attention-actor-critic multi-agent deep reinforcement learning
DSM Demand-side management
DRL Deep reinforcement learning
SAC Soft actor-critic
RBC Rule-based controller
MARLISA Multi-agent reinforcement learning with iterative sequential actions
MAS Multi-agent system
MARL Multi-agent reinforcement learning
DDPG Deep deterministic policy gradient
CTDE Centralized training decentralized execution
GAT Graph attention network
CA Alameda County, California
TX Texas County, Texas
ESS Energy storage system
BESS Battery energy storage systems
DHW Domestic hot water
PV Photovoltaic panels
SOC State of charge
KPI Key performance indicator
PAR Peak to average ratio
LF Load factor

1. Introduction

The transition toward decentralized power generation is not only reshaping the energy landscape but also requiring innovative solutions to manage increasingly complex systems [1]. As distributed energy resources (DERs) become more prevalent [2], traditional methods of supply-side control alone are no longer sufficient to maintain the delicate balance between energy supply and demand [3]. Renewable energy sources such as wind, solar, and hydro, while essential to achieving sustainable energy goals, introduce unpredictability due to their intermittent nature [4]. This variability poses significant challenges to grid stability and highlights the need for more dynamic and intelligent management strategies [5].

Demand-side management (DSM) has emerged as a critical tool to address these challenges [6]. Modern DSM technologies have evolved not only to help consumers adjust their energy usage in response to price signals or demand fluctuations, but also to ensure that the grid can respond effectively to the intermittent supply of renewable energy. Intelligent energy management systems provide the means to monitor and optimize energy consumption, reduce peak demand, and ensure that renewable resources are used efficiently [7]. Additionally, the integration of energy storage systems (ESSs) further enhances grid resilience by smoothing supply and demand imbalances, allowing excess energy to be stored for use during periods of high demand or low generation [8].

In particular, DSM is of great importance in districts with multiple buildings – each with distinct consumption patterns, energy storage systems, and renewable energy generation capabilities – that are connected either through shared infrastructure or the overarching grid [9]. When buildings are managed in isolation, the system fails to capture the benefits of collective optimization. However, when DSM strategies are applied to multiple buildings in a coordinated manner, the entire district can operate more efficiently [10]. Shared infrastructure, such as microgrids, allows buildings to communicate and coordinate their energy use, resulting in optimized energy distribution and improved system-wide outcomes [11]. By balancing both individual building needs and broader grid requirements, this approach improves energy efficiency, accelerates the integration of renewable resources, and strengthens grid stability [12,13].

To achieve this level of coordination, the underlying problem can be modeled as a multi-agent system (MAS), in which each building (or agent) acts based on local and global information. MAS systems can be categorized according to three main architectures:

- *Decentralized*: In this architecture, agents make decisions independently, which enhances scalability due to the distributed nature of computations. However, this approach often leads to non-stationary environments, as agents operate with knowledge limited to their local state rather than the overall system state. This limitation can result in sub-optimal outcomes due to the lack of global awareness [14].
- *Centralized*: Centralized methods employ a single agent to oversee and manage all agents, providing complete knowledge of the system. While this global oversight can enhance decision-making, the architecture is not scalable, as the input space increases exponentially with the number of agents, making it difficult for the centralized agent to compute actions efficiently.
- *Cooperative*: This approach seeks to mitigate the limitations of both decentralized and centralized methods by incorporating an architecture that enables agents to evaluate their policies in the context of other agents' behavior, while still maintaining individual controllers for each agent. This facilitates more effective coordination of agents' actions [15]. Notably, it has been successfully applied to coordinate power consumption in microgrids integrating solar energy and electric vehicles [16,17], or to reduce peak load while maximizing renewable energy usage through the coordination of local controllers [18].

A specialized area within the MAS domain is multi-agent reinforcement learning (MARL), which provides a robust framework for addressing multi-agent decision problems. MARL enables agents to learn and adapt to their environment through interaction and feedback, optimizing local and global objectives [19]. Several studies have demonstrated the effectiveness of MARL in the context of energy management systems. For example, a scalable deep MARL actor-critic method was proposed for managing residential energy flexibility, achieving a 47.2 % reduction in instances of voltage drops below acceptable levels (under-voltage conditions) while simultaneously lowering energy costs, all without requiring the sharing of personal data [20]. Another study addressed scalability in distributed residential energy coordination by introducing a method that isolates marginal contributions to rewards, thereby enabling prosumers to evaluate their impact on system goals while preserving privacy. This strategy reduced energy import costs, distribution network losses, battery wear, and greenhouse gas emissions [21].

In microgrid energy management, MARL has been applied to maximize renewable energy use and reduce costs, with agents controlling energy storage systems and microgrid aggregators [22]. The study showed that the implementation of algorithms such as deep deterministic policy gradient (DDPG) and Multi-Agent DDPG (MADDPG) improved coordination and outperformed single-agent methods.

In residential communities, MARL mitigated rebound peaks under renewable energy uncertainty by coordinating household energy scheduling [23]. Each building functions as an agent, intending to minimize electricity costs while maintaining user satisfaction. The method reduced the overall community energy costs and effectively mitigated rebound peaks by leveraging renewable generation and user coordination. Further research investigated the potential of surrogate models and deep reinforcement learning (DRL) in district energy management [24]. Although not explicitly focused on MARL, this study demonstrated

the benefits of coordination in reducing energy costs and peak demand in multi-building contexts. Another study compared coordinated and cooperative DRL controllers in district energy systems, showing that both approaches outperformed rule-based controllers in reducing costs, lowering demand peaks, and increasing self-consumption, thereby underscoring the benefits of intelligent control strategies [25].

1.1. Attention-based MARL in energy management systems: literature review and research gap

Many of the aforementioned MARL approaches adopt centralized training with a decentralized execution (CTDE) framework, which enhances coordination among agents [26]. Nevertheless, a significant drawback of CTDE is the mismatch between the training and execution phases: while agents benefit from centralized critics with global information during training, they must rely solely on local observations during execution. This introduces an independence assumption in agent policies, which hinders efficient collaboration and policy exploration. As a result, even during centralized training, agents struggle to fully leverage cooperative global information, limiting their potential for more effective coordination [27].

Within the CTDE framework, a typical multi-agent policy consists of decentralized actors guided by a single critic that estimates a global action-value function based on actors' states and joint actions. This setup enhances coordination by capturing the interdependent effects of all agents' actions. However, learning an accurate action-value function becomes increasingly impractical as the number of agents grows due to the exponential expansion of the state-action space [28]. Other CTDE methods employ value function factorization [29,30], which decomposes the centralized action-value function into individual agent-specific action-value functions. However, in both cases – whether using a single critic or multiple critics – training relies on global information, including the actions and states of all agents. Since this information is unavailable during execution, and cooperation remains challenging even during training due to divergent action tendencies among agents [31], this ultimately leads to suboptimal performance.

To mitigate these issues, this study integrates a multi-head attention mechanism within the CTDE framework. By employing a query-key-value structure, the attention mechanism enhances the evaluation of each agent's Q-function, improving communication and coordination among agents, as elaborated in Section 3.1.2. The approach begins with a message-exchange phase, allowing agents to share information. Subsequently, by dynamically assigning weights to received information, the attention mechanism directs agents' focus toward the most critical data for execution. This process allows agents to focus on the most relevant data, enhances the algorithm's generalization, and improves scalability by discarding redundant information [32,33], ultimately leading to more informed and efficient decision-making [34].

In the field of energy management, studies have effectively leveraged attention mechanisms to address various energy-related challenges. In energy systems control, different attention mechanisms – such as self-attention [35–37], multi-head attention [35,38,39], and graph attention networks (GAT) [40,41] – have been integrated into RL frameworks. Self-attention, originally designed for sequential decision-making, prioritizes information within local segments of the input sequence but lacks a holistic understanding of the entire sequence, which may constrain the performance of reinforcement learning (RL) algorithms in certain tasks [34]. Instead, GAT is particularly effective in spatial-temporal cooperation among agents and in handling graph-structured data [42].

Multi-head attention enhances decision-making robustness by assigning different weights to multiple attention heads, enabling the model to capture diverse aspects of the input data more effectively. In energy management, it helps agents focus on the most critical energy-related information, improving efficiency of the system. For example, a study focused on energy management in a multi-energy industrial park developed an algorithm that allowed different energy sources, storage units, and conversion devices to act as agents [43]. Each agent focused on the most relevant data, such as fluctuating energy demands and prices, to optimize long-term energy costs while ensuring user demand was met. Similarly, in grid-responsive buildings, another study employed attention-based MARL to coordinate demand response across multiple buildings [44]. The attention mechanism enabled agents to focus on the most pertinent interactions with other agents, thereby enhancing overall performance, reducing energy costs, and improving grid stability. Furthermore, this selective focus facilitated the resolution of issues such as the energy rebound effect, which arises when buildings alter their energy consumption patterns during peak periods. Instead, in the context of networked greenhouses [45], attention mechanisms were instrumental in enabling each greenhouse, acting as an agent, to efficiently manage its energy consumption. By focusing on the dynamic variability of electricity prices and renewable energy generation, the MADRL system was able to maintain optimal growing conditions while simultaneously reducing overall energy usage. In addition, attention-based MADRL algorithms have been employed in large-scale peer-to-peer (P2P) transactive energy trading, facilitating the decentralized management of energy trades by prosumers [46]. The attention mechanism enables agents to focus on pertinent information from other prosumers, thereby ensuring privacy and scalability while optimizing energy exchange within the community.

Nevertheless, the use of multi-head attention mechanisms in deep reinforcement learning (DRL) algorithms for energy system control has been limited to a relatively small scale of agents, with most studies focusing on continuous policy networks. Moreover, their energy system performance has not been evaluated in relation to the type of objective reward utilized, whether fully cooperative, fully competitive, or mixed compared with respect to well-known DRL algorithms. To this end, the development of a multi-head attention mechanism integrated with a DRL algorithm is proposed, utilizing discrete policy networks with policy sampling implemented through multinomial distributions.

1.2. Contributions and structure of this work

This paper addresses the challenge of large-scale demand-side management (DSM) in residential districts by introducing a discrete deep reinforcement learning algorithm, integrated with multi-head attention mechanism, namely Attention-Actor Critic Multi-Agent deep reinforcement learning (AAC-MADRL) inspired by [47]. Tailored to optimize coordination in DSM scenarios, AAC-MADRL has been developed to address the limitations of previous studies in this field, which primarily evaluated the efficiency of attention-based MARL in districts with a limited number of buildings (fewer than 10 buildings).

The main contributions of this paper can be summarized as follows.

- Scalability and Robustness: despite attention-based DRL algorithms having been proven to be scalable and robust [34], this study specifically evaluates the learning capacity of the multi-head attention critic as the number of agents increases. To this end, AAC-MADRL was tested on both smaller subsets and full datasets, examining its ability to handle different sizes of neighborhoods while maintaining performance. For robustness, the algorithm was evaluated under both winter and summer conditions, offering insights into its adaptability to different environmental conditions. The analysis focused on both the evaluation of performance in energy management strategies and the execution time of the deployment, to demonstrate that AAC-MADRL remains efficient across different scales and environments.
- Quantifying the effectiveness of Multi-Head Attention in Multi-Agent DRL: while multi-head attention mechanisms have shown promising results in both competitive and cooperative environments [34], their effectiveness in large-scale demand-side management (DSM) remains largely unquantified. To address this, the study systematically compares AAC-MADRL with two well-established methods: soft actor-critic (SAC) [48], a decentralized approach suited for

competitive settings, and MARLISA, a cooperative MADRL method that does not utilize attention mechanisms [49]. To ensure a structured and fair evaluation, a sensitivity analysis of the parameterized reward function is conducted, integrating both individual building-level objectives and broader district-wide goals. By adjusting the weightings of these components, the adaptability of AAC-MADRL is assessed across fully cooperative, fully competitive, and mixed reward formulations. This analysis provides a quantitative assessment of the trade-offs between performance and execution time, offering deeper insights into the impact of multi-head attention mechanisms in DSM applications, providing insights into when multi-head attention mechanisms may be beneficial for DSM applications, compared to existing methods.

- Code Availability for Future Benchmarking: The AAC-MADRL algorithm is implemented in the CityLearn environment (aggiungi CityLearn citazione). This ensures that the algorithm can be used for future case studies and further benchmarking of demand-side management strategies for large districts of buildings.

Furthermore, all DRL-based algorithms, i.e. AAC-MADRL, SAC, and MARLISA, are benchmarked against a rule-based controller (RBC) baseline. The comparison focuses on system performance, learning stability, and energy optimization, with particular attention to DSM tasks such as energy storage management, renewable energy integration, and peak demand reduction. This analysis highlights the potential of attention-based cooperative MARL as a scalable and efficient solution for future energy management systems, particularly in contrast to rule-based or centralized approaches, which are often less scalable and computationally intensive [50].

The rest of this paper is organized as follows. Section 2 describes the methodology, while Section 3 presents the control strategies used, focusing on the structure of the proposed attention-based multi-agent DRL algorithm (AAC-MADRL). Section 4 introduces the case study and outlines the design of the control problem. Section 5 presents the results. Section 6 provides the discussion and conclusions, followed by limitations and directions for future work in Section 7.

2. Methodology

This section presents the methodological framework used in this study (Fig. 2.1) which consists of multiple steps to benchmark the proposed AAC-MADRL against existing controllers for district energy management, including SAC, MARLISA, and rule-based controllers. This structured approach enables a thorough evaluation of the algorithm's performance in terms of effectiveness, scalability, and adaptability within complex energy systems. The following subsections detail each step of the process.

2.1. Data selection

The effectiveness of AAC-MADRL for energy management is assessed using two neighborhood datasets that represent the heating season in Alameda County, California (CA), and the cooling season in Travis County, Texas (TX) [51]. These locations were chosen not only for their representative residential energy usage patterns but also for their contrasting environmental conditions, which lead to diverse building characteristics, energy systems, and operational behaviors. In Section 4.1 further details on the selected datasets are provided.

Additionally, this study utilizes a synthetic dataset (Section 4.1) to capture time-varying energy prices, including peak and off-peak periods. This dataset ensures that the control strategies are assessed within realistic economic scenarios, emphasizing the importance of balancing energy efficiency with cost-effectiveness.

Together, the combination of diverse climatic zones, variations in energy storage and consumption patterns, and dynamic pricing creates a comprehensive environment for testing and validating the proposed algorithm across various operational conditions.

2.1.1. Data stratification

To evaluate the scalability and performance of the AAC-MADRL algorithm, the selected neighborhoods' data were split into smaller subsets. In Alameda County, California (CA), the dataset was split into two subsets of 10 and 34 buildings, in addition to the full set of 73 buildings. For the Texas (TX) neighborhood, the dataset was divided into four subsets: 10, 25, 50, and the complete dataset of 100 buildings. This approach enabled a systematic assessment of the algorithm's scalability across various neighborhood sizes, ensuring that the methods remained efficient and adaptable, regardless of scale.

To ensure a fair comparison, the same methodology was applied to both SAC and MARLISA DRL algorithms and the RBC baseline controllers. By using progressively larger datasets, a comparative assessment of AAC-MADRL's performance against these different control strategies was conducted regarding computational efficiency and energy management capabilities within increasingly complex environments.

2.2. Rationale of the control problem

For all DRL algorithms, the reward function of each agent is defined by a parameterized approach, as given by the following expression:

$$r_i = (1-\beta)\cdot \mathbf{I} - \beta \cdot \frac{\mathbf{C}}{N^{\gamma}} \tag{2.1}$$

where N is the number of agents in the system, $\mathbf{I}$ represent agent-specific term, while $\mathbf{C}$ is the collective-interest term.

In this context, the parameter γ is employed to achieve a balance between the numerical values attributed to the individual and collective components. Specifically, the value of γ was selected based on the relative amplitude of each building's load within the overall district, for both the CA and TX datasets. In the case of the TX County, the value of γ was set to 3.5 for districts comprising up to 50 buildings and reduced to 2.8 for districts with 100 buildings. This reduction in γ for larger district sizes reflects the increased variability in building loads across the district. Instead, for the CA dataset, γ was set to a constant value of 2.8 across all district sizes, as the load variability was lower (see Fig. A.3 of Appendix A).

Instead, the β-parameterization enables the exploration of various strategies, ranging from purely individual objectives ($\beta = 0$) to entirely collective goals ($\beta = 1$) with intermediate values representing mixed strategies. This methodology is designed to assess the efficacy of cooperative and decentralized architectures to the specific objectives of the reward function. The goal is to develop a comprehensive framework for identifying the conditions under which one architecture may prove more advantageous than the other, thereby enhancing the practical applicability of each architecture across different scenarios.

2.3. Simulation environment setup

This study employed the CityLearn Simulation Environment, an open-source platform specifically designed for urban-scale energy management simulations [52]. Built on the OpenAI Gym framework, CityLearn provides a platform for the training, deployment, and benchmarking of a wide range of controllers, including multi-agent reinforcement learning (MARL) and rule-based controllers.

The environment simulates real-world conditions, incorporating dynamic energy pricing, PV generation, and the operation of key energy systems such as heating, ventilation, and air conditioning (HVAC), as well as electrical and thermal storage components. The hourly timestep allows for precise management of each building's energy systems, encompassing the charging and discharging processes of batteries and domestic hot water (DHW) storage, which are critical control variables in the optimization problem. For further details, please refer to [53].

In this study, the neighborhoods of CA and TX County, along with their respective building subsets, were integrated into the simulation environment. Additionally, the customized reward function and the AAC-MADRL algorithm were benchmarked against pre-existing DRL

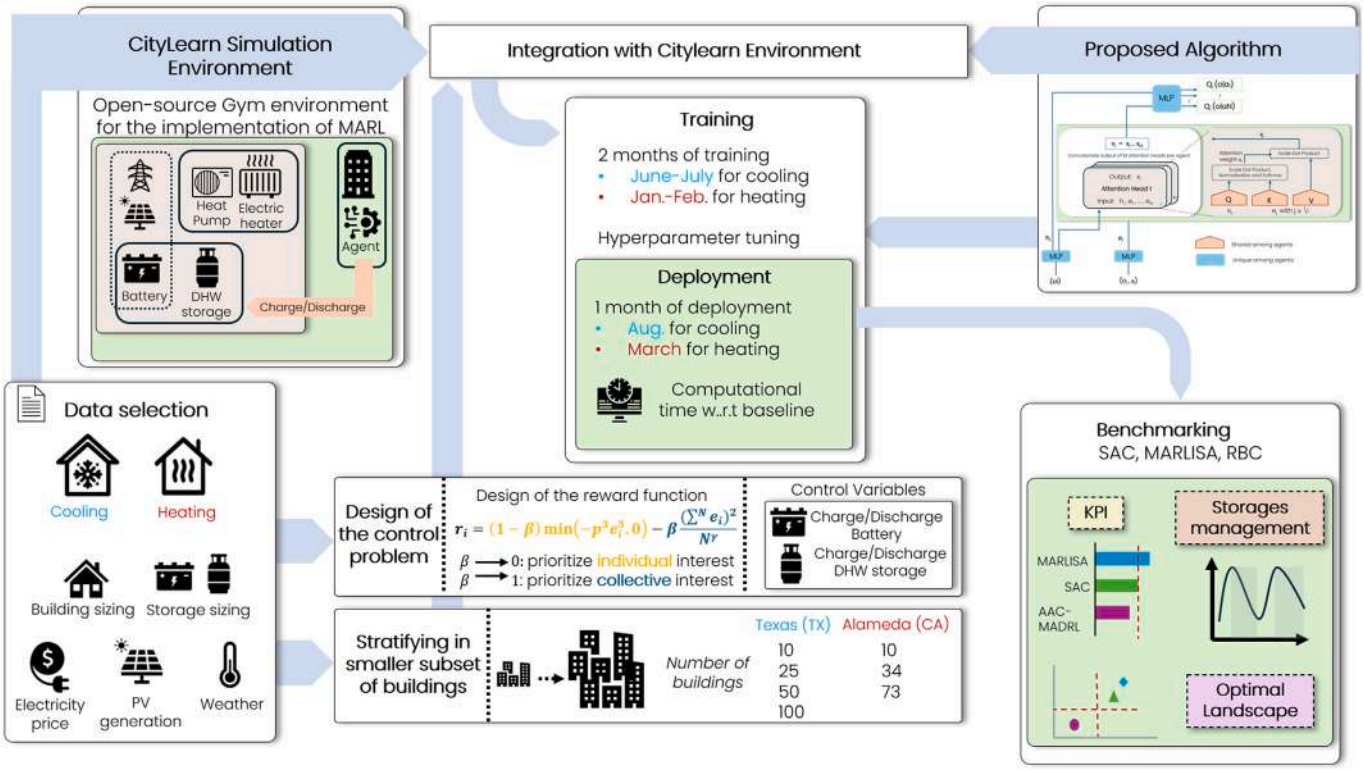


Fig. 2.1. Methodological framework.

algorithms (SAC, MARLISA) and a rule-based controller within the CityLearn environment. AAC-MADRL will be made available in the environment to enrich the benchmarking process.

2.4. Training and deployment process

The agents were trained by adjusting the value of $\beta \in [0, 1]$ in Eq. (2.1) by increments of 0.25 to explore the trade-off between individual and collective optimization goals. This approach allows for a comprehensive examination of the impact of varying degrees of cooperation and competition on energy management outcomes. Furthermore, the agents were trained across different neighborhood sizes.

The TX dataset underwent training during June and July, which aligns with the cooling season, whereas the CA dataset was trained over January and February, thus during the heating season. The training process was consistent across all DRL algorithms, including soft actor-critic (SAC), MARLISA, and the AAC-MADRL. Each agent interacted with the simulation environment for a minimum of 10 and a maximum of 15 episodes with a learning rate of $lr = 1\mathrm{e}-3$ for SAC and AAC-MADRL, while $lr = 3\mathrm{e}-4$ was used in MARLISA.

Following training, the agents were deployed for evaluation in August for the TX county case and in March for the CA dataset. Computational time was also measured for all algorithms, comparing them to a rule-based approach across different neighborhood sizes. This analysis provided valuable insights into the efficiency and scalability of the AAC-MADRL control algorithm (Table 5.1).

2.5. Evaluation: key performance indicators (KPI)

To assess the effectiveness of energy management, cost efficiency, and grid stability, several key performance indicators (KPIs) were employed to evaluate the trade-offs between individual and collective optimization goals across varying neighborhood sizes and values of β, as outlined in Table 2.1.

Table 2.1
KPI.

KPI	Formula	Units
District cost	$\sum_t^T min(-\mathbf{e_t}\mathbf{p_t}, 0)$	$
District load variance	$\frac{\sum_t^T (\mathbf{e_t}-\bar{\mathbf{e}})^2}{T}$	kWh^2
District import	$\sum_t^T min(-\mathbf{e_t}, 0)$	kWh
1-load factor	$\frac{\bar{\mathbf{e}}}{Peak}$	[–]
Self-sufficiency	$\frac{\sum_t^T \min(\mathbf{load_t}, \mathbf{PV_t})}{\sum_t^T \mathbf{load_t}}$	[–]
Daily peak average	$\frac{\sum_{day}^{n_{day}} Peak_{day}}{n_{day}}$	kW
Daily PAR average	$\frac{\sum_{day}^{n_{day}} \frac{Peak_{day}}{\bar{\mathbf{e}}_{\mathbf{day}}}}{n_{day}}$	[–]

In this context, $\mathbf{e_t}$ denotes the net electricity consumption of the district at time step t, while T represents the entire deployment period, set to one month. The use of net district consumption is particularly important when considering the practicalities of implementing energy management strategies in larger neighborhoods. As the size of the district increases, it becomes less feasible to evaluate each building individually. Consequently, a comprehensive assessment of the overall behavior of the district provides a clearer understanding of how these strategies work at scale.

Each KPI focuses on different aspects of energy management and provides valuable insight into the effectiveness of control strategies in balancing individual energy needs with the collective interests of the neighborhood. For example, the *District Cost* reflects the system's ability to minimize expenses for each building while accounting for fluctuations in energy prices imposed by the grid. Although each building aims to reduce its own costs, this individual optimization could negatively impact overall district load stability as each agent tends to request energy when prices are low, leading to demand peaks. In contrast, the *District Load Variance* is closely related to the collective optimization goal, measuring

fluctuations in the district load. The objective is to ensure that buildings avoid importing or supplying energy to the grid simultaneously, thereby coordinating to maintain district load stability. The lower variance suggests that the district load experiences fewer fluctuations. Furthermore, the *1-load factor* metric measures the ratio of the average electricity consumption to the peak demand for the entire month within the neighborhood. The *District Import* quantifies the total energy imported from the grid, while the *Self-sufficiency* evaluates the capacity of the neighborhood to meet its energy requirements through the utilization of its renewable energy sources, including photovoltaic generation and storage systems. A higher level of self-sufficiency implies a reduced reliance on the grid and improved overall stability. The *Daily Peak Average* indicates the average district daily peak over the simulation period and is related to both reward terms as minimizing peaks is the objective of both components. Finally, the *Daily PAR Average* evaluates the smoothness of daily energy import with lower PAR values indicating more consistent consumption.

3. Algorithmic framework and controllers for demand response

3.1. Multi-agent attention actor critic

In their seminal work [47], the authors introduced the Actor-Attention-Critic framework, which integrates attention mechanisms to improve cooperation among agents. A principal advantage of this approach is its potential to facilitate agent learning of effective policies in multi-agent environments by focusing on the most relevant information from other agents. The algorithm builds upon the actor-critic paradigm, where each agent has its own actor-network. Instead, the critic network employs a shared attention mechanism to aggregate information from all agents and agent-specific multi-layer perceptrons (MLPs) that evaluate each agent's Q-value function (Fig. 3.1). For the purposes of this discussion, the overall structure will be referred to as the critic network.

This section presents an overview of the operational mechanisms of each network within this framework.

3.1.1. Actor network

The actor-network of agent i, designated as π^i, represents the optimal policy that has been learned. The actor-network receives observations o specific to the agent's local environment and maps them to a probability distribution over possible actions a. This distribution guides the agent in selecting actions based on the policy network parameterized by θ. The parameters of the actor-network are updated and optimized by maximizing the expected cumulative reward $J(\pi^i_\theta)$ via gradient descent:

$$\nabla_\theta J(\pi^i_\theta) = \mathbb{E}_{a\sim\pi^i, o\sim D}\left[\nabla_\theta \log(\pi^i_\theta(a_i|o))(\alpha log(\pi^i_\theta(a_i|o)) - A_i(o,a))\right] \tag{3.1}$$

where α is the entropy parameter, while A_i represent the advantage function, defined as:

$$A_i(o,a) = Q^\Psi_i(o,a) - \mathbb{E}_{a_i\sim\pi^i(o)}[Q^\Psi(o,(a_i,a_{\setminus i})] \tag{3.2}$$

where $Q^\Psi_i(o,a)$ represents the action-value function for agent i, while the term $\mathbb{E}_{a_i\sim\pi^i(o)}[Q^\Psi(o,(a_i,a_{\setminus i})]$ represents the expected value of Q_i over all possible actions of agent i. By leveraging this comparison during the upgrade of its policy the agent can determine whether a specific action leads to an increase of the Q-value over the average of its possible actions. During training, the actor learns to associate the agent's observations o with actions that maximize the overall reward based on feedback from the critic.

3.1.2. Critic network

The critic network evaluates the actions undertaken by all the N agents, based on their collective state $\mathbf{o} = (o_1, o_2, \dots o_N)$ and actions $\mathbf{a} = (a_1, a_2 \dots a_N)$ vectors. Rather than directly concatenating the observations and actions of all agents, the algorithm integrates the attention heads within the critic network. Initially, the state-action space of each agent, represented by the tuple (o,a), is processed through a one-layer MLP to generate the embedding $e_i = g_i(o,a)$. For the discrete algorithm (as described by Section 3.2), each agent state space o is encoded in parallel by a state encoder – also a one-layer MLP – to produce the query vector $h_i = h_i(o)$. These encoders are agent-specific, ensuring that the parameters are not shared across agents. The resulting embeddings e_i and h_i are subsequently processed by the attention mechanism utilizing a query-key-value framework that consists of shared matrices $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$. This shared structure enables the critic to consistently process and compare information from different agents.

Specifically, the query vector, $(\mathbf{Q}^\mathbf{T}h)_i$ interacts with the key vectors $(\mathbf{K}^\mathbf{T}e)_{j\neq i}$ of all other agents j to compute similarity scores. These similarity scores indicate the relevance of the information from agent j to

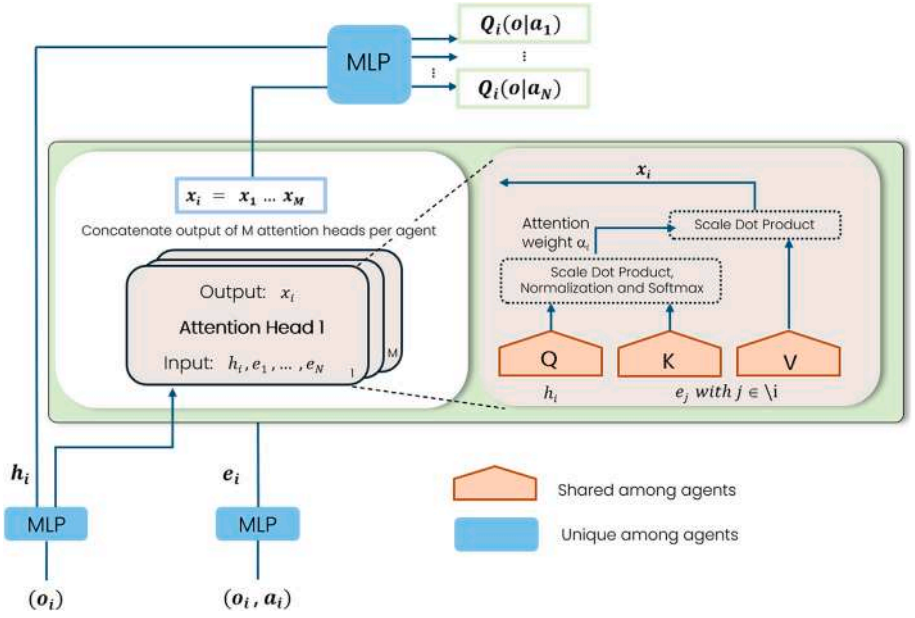


Fig. 3.1. Method for calculating $Q^\Psi_i(o|a_m)$ using the attention mechanism for agent i.

agent i. The scores are then scaled according to [54] and passed through the softmax function σ, resulting in the attention weight for each agent α_i:

$$\alpha_i = \sigma\left(\frac{(\mathbf{Q}^{\mathbf{T}}h)_i(\mathbf{K}^{\mathbf{T}}e)_{j\neq i}}{\sqrt{|\mathbf{K}^T e|}}\right) \tag{3.3}$$

Next, the attention vector x_i for agent i is computed as a weighted sum of the value vectors $\mathbf{V_j}$:

$$x_i = \sum_{j\neq i} \alpha_{ij}\mathbf{V_j} \tag{3.4}$$

where the value vectors $\mathbf{V_j} = \mathbf{V}e_j$ are obtained from the embeddings e_j using the shared transformation matrix $\mathbf{V}$, representing the information provided by each agent. Of notice, with n attention heads, each head can focus on a distinct weighted mixture of $(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ and the resulting n attention vectors x_i for agent i are concatenated into a single vector.

Finally, the Q-value for each action a_m of agent i is computed by combining the attention-augmented information x_i with the state embedding h_i through a dedicated two-layers MLP f_i specific to agent i:

$$Q_i(o|a_m) = f_i(h_i, x_i) \tag{3.5}$$

This structured approach ensures that the critic effectively evaluates the overall performance of joint actions while selectively emphasizing the most relevant information from other agents, thereby improving cooperative decision-making in multi-agent scenarios. Fig. 3.1 illustrates the aforementioned attention mechanism.

The critic update involves adjusting the network parameters to minimize the mean squared error between each agent's current $Q_i^{\psi}(o,a)$, and the target Q-value y_i. This difference is calculated for each agent, to minimize a cumulative loss that aggregates the mean square errors of all agents. The critic loss function is defined as:

$$\mathcal{L}_Q(\psi) = \sum_{i=1}^{N} \mathbb{E}_{(o,a,r,o')\sim D}\left[(Q_i^{\psi}(o,a) - y_i)^2\right] \tag{3.6}$$

where y_i represents the target Q-value for agent i, computed as follows:

$$y_i = r_i + \gamma \mathbb{E}_{a'\sim\pi_{\bar{\theta}}(o')}\left[Q_i^{\bar{\psi}}(o',a') - \alpha\log(\pi_{\bar{\theta}_i}(a_i'|o'))\right] \tag{3.7}$$

where r_i is the reward for agent i, and γ the discount factor. Next observations o' are computed based on the current actions, while next actions a' are estimated using the agent's current policy network. The Q-value is computed using the target critic networks $Q^{\bar{\psi}}$. The term $\alpha\log(\pi_{\bar{\theta}_i}(a_i'|o'))$ represents a soft update: the reward is adjusted based on the log-probability of the action under the target policy, balancing the immediate reward r_i with the likelihood of actions taken.

The target networks $Q^{\bar{\psi}}$ and $\pi_{\bar{\theta}_i}$, which are used to stabilize the learning process, are updated gradually preventing large fluctuations in value estimates and help ensure steady learning.

3.2. AAC-MADRL: multiple actions and action variability per agent

This section details the modifications made to the publicly available resources introduced by [47] to meet the specific requirements of this study. To the best of the authors' knowledge, the existing public repository supports only a single action per agent, which is insufficient for the needs of this research. Consequently, the algorithm was extended to support multiple discrete actions per agent and to accommodate customized action spaces. This modification allows the number of actions to vary across agents, reflecting the diverse energy systems present in each building.

The primary architectural changes were applied to the policy networks, which have been tailored to output probability distributions over categorical actions. For each action, only one value is selected from a set of possible k categories, where the number of categories k is flexible and can vary depending on the type of action being modeled. This modification ensures that the policy network produces outputs with variable dimensionality, depending on the number of actions and categories assigned to each agent. During training, each action is sampled from its categorical distribution to encourage exploration and help agents discover optimal strategies. The sampled actions are then converted into one-hot encoded vectors, which are passed to the attention-based critic network. This critic evaluates the joint state-action values based on these one-hot representations, essential for computing Q-values across discrete combinations. In the deployment phase, the agent selects the most probable category for each action dimension. This approach supports a highly adaptable action space, allowing both the number of categories and type of action to vary across agents without necessitating changes to the underlying network architectures. The control resolution can be easily tuned by adjusting the number of categories, enabling alignment with the specific operational characteristics of each device. As a result, each agent operates with a tailored set of action categories suited to the systems it manages – such as energy storage systems (ESS), thermal storage, or heating and cooling devices – which may differ across buildings.

3.2.1. Integration with the CityLearn simulation environment

Further adaptations were made to ensure the seamless integration of the algorithm within the CityLearn simulation environment, which inherently operates within a continuous action space. Specifically, actions related to heating and cooling devices are represented in the range [0, 1], while actions related to the charging and discharging of energy storage systems are mapped within the range [−1, 1], where negative values correspond to discharging and positive values to charging.

The AAC-MADRL framework selects a discrete category for each action, where each category corresponds to a fixed control value within the valid operational range of the target device. These values are uniformly spaced and predefined. For example, energy storage systems operating in the interval [−1, 1] are discretized at 0.1 intervals, resulting in the action set $[-1.0, -0.9, \ldots, 0.9, 1.0]$, with each category index assigned directly to one of these values (e.g. Category 0 = −1.0, Category 1 = −0.9, etc.). This mapping guarantees that each control action remains valid and interpretable in the context of the environment. Furthermore, since the environment's reward functions are computed based on these values, the mapped actions must be stored in this form when added to the replay buffer.

However, the learning algorithm is structured around discrete action representations. The actions need to be represented in their discrete form before being used for network updates. This design serves both algorithmic and practical goals. From an algorithmic perspective, the Attention-Critic network is designed to operate with discrete policies, where the critic must evaluate the value of all possible actions. This is only feasible when the action set is limited. From a practical perspective, discretization prevents unnecessary fine-grained adjustments that would have a negligible impact on the real-world system.

This dual approach ensures effective system operation by maintaining compatibility between the discrete decision-making framework and the continuous action space of the CityLearn simulation environment.

3.3. Benchmark overview: SAC, MARLISA, RBC

A comprehensive assessment of the proposed algorithm requires a comparison with established methodologies for energy management in multi-agent systems. This section provides an overview of the soft actor-critic (SAC) and multi-agent reinforcement learning with Iterative Sequential Actions (MARLISA), as well as the baseline control strategy,

the Rule-Based Control (RBC). In addition to evaluating the relative performance of different algorithms, this comparison seeks to grasp the distinctive characteristics of the various methods and their efficacy in multi-agent settings across a range of objective scenarios and system sizes.

3.3.1. Soft actor-critic (SAC)

Soft actor-critic (SAC) is a widely recognized RL algorithm introduced by [48], designed for continuous control problems. SAC operates within the actor-critic framework, where two separate networks, the actor and the critic, work together to iteratively improve the agent's policy. The actor learns a policy π that maps observations to actions, while the critic evaluates the effectiveness of the chosen actions by estimating the expected return through the value function $Q(s, a)$. The SAC algorithm's update rule follows a gradient-based approach:

$$\nabla_\theta J(\pi) = \mathbf{E}_\pi[\nabla_\theta \log \pi(a|s)(Q(s,a) - \alpha \log \pi(a|s))]$$

where α controls the trade-off between reward maximization and entropy, which influences exploration.

However, SAC's architecture is not inherently designed for multi-agent settings, instead, it is particularly effective for single-agent systems in continuous action spaces. While a centralized version of SAC can be adapted for environments with multiple agents, it faces scalability and efficiency bottlenecks, particularly in distributed, cooperative systems. The need for coordination among agents in multi-agent settings presents challenges that SAC does not explicitly address, making it an important comparison point for cooperative methods like AAC-MADRL.

3.3.2. Multi-agent reinforcement learning with iterative sequential actions: MARLISA

The MARLISA approach, as outlined in [49] employs a leader–follower mechanism to achieve coordination among multiple agents. In this framework, agents predict their future actions iteratively and share this information with others in the system. Each agent makes its decision after considering the predicted actions of its previous agents, but all agents execute their final actions simultaneously.

In a multi-agent setting, a key distinction from the centralized SAC is that MARLISA allows each agent to maintain its own policy and critic networks. This independence enables agents to learn separately while still exchanging crucial information with others. While the leader–follower approach promotes coordinated behavior, it also introduces communication overhead, particularly as the number of agents increases, which can negatively impact scalability. In comparison, AAC-MADRL addresses this challenge by using an attention mechanism to enable parallel decision-making across agents. By focusing on the most relevant information from other agents, AAC-MADRL eliminates the need for sequential communication, improving both scalability and coordination in multi-agent environments.

3.3.3. Rule-based control (RBC)

RBC is a non-learning approach that relies on predefined rules to manage energy systems, making it a straightforward yet inflexible solution for energy management. In this study, a time-of-use rule-based controller specifically tailored to exploit solar generation has been considered as a baseline. This controller operates based on fixed schedules for storage's charging and discharging, as well as for controlling heating and cooling devices.

The agent controller manages the energy storage systems with a charging rate of 11.0 % of maximum capacity every hour between 6 a.m. and 2 p.m. During this time frame, solar generation is typically at its peak, allowing the controller to effectively store excess renewable energy. Outside of this charging window, the system discharges at a rate of 6.7 % of its maximum capacity per hour, utilizing the stored energy to meet the building's demand during periods when solar generation is unavailable.

The controller also adjusts the power of cooling and heating devices based on the same time-of-use schedule. From 6 a.m. to 2 p.m., cooling devices are set to operate at 70 % of their capacity, while outside this time window, the cooling output is reduced to 30 %. Conversely, heating devices function at 30 % of their nominal power during the daytime hours and increase to 70 % during the rest of the day. This operational strategy is designed to efficiently meet the thermal needs of the building throughout the day.

The main advantage of RBC lies in its simplicity and low computational demand. The controller follows a deterministic rule set, ensuring consistent operation without the need for complex computation or learning mechanisms. This makes it a viable option for scenarios where real-time adaptability is not critical, or where computational resources are limited.

However, this approach has significant limitations, particularly in dynamic environments. Since RBC operates on static rules, it cannot adjust to fluctuations in energy demand, changes in pricing, or variations in solar generation that occur in real time. This rigidity often leads to suboptimal performance.

4. Case study and control problem

This section presents an overview of the case study, starting with an examination of the dataset and seasonal weather patterns associated with the district under consideration. This is followed by a description of the energy systems installed at the building level. Lastly, the control problem and its constraints are discussed.

4.1. District data and seasonality

In this study, two distinct datasets representing neighborhoods in different climatic zones were selected for analysis of energy consumption patterns and management strategies. One dataset corresponds to Travis County, Texas (TX), comprising 100 residential buildings, while the other pertains to Alameda, California (CA), with 73 buildings. These datasets were sourced from [51]. Table 4.1 provides an overview of the geometrical characteristics of the residential buildings analyzed. For each building, hourly data is provided, encompassing variables such as end-use loads, occupancy levels, solar power generation, and indoor environmental metrics.

The climatic conditions in these regions exhibit notable disparities when classified according to the ASHRAE system [55]. Travis County belongs to Climate Zone 2 A, characterized by a warm and humid climate. In contrast, Alameda is categorized as Climate Zone 3C, distinguished by its relatively mild climate. The analysis is focused on two specific seasons: the cooling season (June to August 2018) for Travis County and the heating season (January to March 2018) for Alameda, following the methodologies outlined in reference [51]. The temperature statistics for the specified periods are detailed in Table 4.2, while the PV production data are presented in Table 4.3.

Furthermore, a pricing dataset was synthesized to capture both current energy costs and forecasts for the upcoming six and twelve hours, reflecting the expenses faced by residents in both regions. This pricing structure refers to the cost of importing energy from the grid and includes a high tariff rate of $0.54/kWh during peak hours from 4 p.m. to 9 p.m., and a lower tariff rate of $0.22/kWh during off-peak hours.

The tables demonstrate the seasonal fluctuations of the case studies under consideration, which subsequently impact neighborhood energy consumption patterns. As illustrated in Figs. A.1 and A.2 of Appendix A,

Table 4.1
Buildings neighborhood roof area characteristic in m^2.

County	Average	Min	Max	St. Dev.
Travis Co., TX	193.6	34.4	470.4	89.4
Alameda Co., CA	161.9	63.4	336.4	66.4

Table 4.2
Neighborhood temperature for the respective seasons [°C].

County	Climate zone	Average	Min	Max	St. Dev.
TX	2A	30.1	21.7	42.8	4.9
CA	3C	11.3	0	25	3.7

Table 4.3
Neighborhood PV production for the respective seasons [kW].

County	Climate zone	Average	Max
TX	2A	3.67	7.18
CA	3C	2.58	7.14

cooling demands reach their peak during periods of high electricity tariffs in the summer, and increase in conjunction with solar generation. In contrast, winter heating demands are lower overall, with peaks occurring during periods of low electricity tariffs and decreasing during solar production periods. This seasonal variation suggests there may be greater potential for optimization of energy storage during winter compared to summer.

4.2. Energy system at building level

The energy systems of the buildings in both the TX and CA districts comprise photovoltaic (PV) panels, a heat pump, an electric heater, and two types of storage: a thermal storage tank for domestic hot water (DHW) and an electrical battery. Buildings are equipped with PV panels with a nominal power output of 10 kW, an electrical battery with a fixed capacity of 13.5 kWh, and a nominal power of 2.5 kW. The size of the DHW system varies by building and is designed to meet the maximum daily load [51].

Fig. 4.1 presents the energy system architecture for a prototypical building in the district, highlighting its key components and their interactions. The mathematical models of the energy devices are detailed in [56]. Adapted from [56], the diagram emphasizes the energy-related aspects relevant to this case study, illustrating how the system meets the buildings' energy demands while focusing on the controlled components – namely, the battery and DHW storage. The energy supply for each building is sourced from PV panels and the grid. Energy is directed to the heat pump to meet the building's heating and cooling demands, while the electric heater is employed for DHW loads. The electric heater can either provide hot water on demand or store energy in the DHW storage tank for later use. The electrical storage system (ESS) can be charged by the PV and electrical grid. Ideally, during periods of overproduction, surplus electricity is stored in the ESS for later use, supporting the electric heater, heat pumps, or fulfilling the electrical demands of appliances. Any excess electricity stored in the battery cannot be transferred between buildings, but it can be fed back into the grid at zero cost, establishing the grid as the primary means of communication among the buildings. This configuration enables the grid to assist the buildings in meeting their energy demands by sourcing power generated from other buildings in the district.

4.3. State and action space

4.3.1. State-space

In agent-based models, the agents learn the optimal policies by observing the influence of their actions on the environment. In the present case study, the state space for all buildings is constituted by weather data, district-level conditions, and building-specific states. To account for inherent variability and facilitate the learning process, observations are subjected to standardization. To augment the predictive capabilities of the controllers, six-hour and twelve-hour weather forecasts, as well as forecasts for particular district variables, have been integrated, as outlined in Table 4.4. Weather observations comprise both the outdoor dry bulb temperature and the direct solar radiation, along with their respective 6- and 12-hour forecasts. Weather conditions are estimated using a generic forecasting model with a predefined prediction error, which increases with the forecast time horizon for both temperature and solar radiation. The error starts at 2.5 % for 6-hour-ahead predictions and reaches 10 % for 24-hour-ahead forecasts for solar radiation. For outdoor temperature predictions, the accuracy is ±0.3 °C for 6-hour-ahead, ±0.65 °C for 12-hour-ahead, and ±1.35 °C for 24-hour-ahead forecasts [53]. District-level variables are common to all buildings and include day type (ranging from 1 to 7 for the days of the week), hour of the

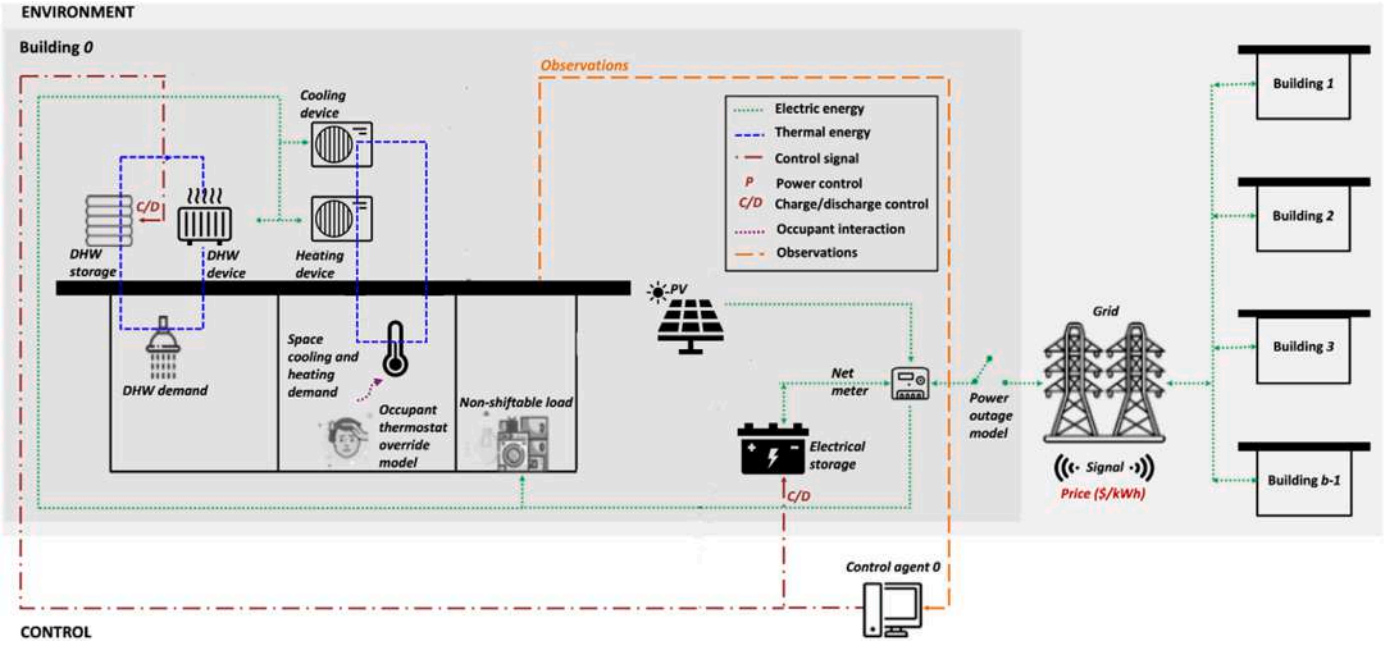


Fig. 4.1. Energy system architecture for a representative building within the district. Modified from [56].

Table 4.4
State-space description.

Variable	Unit
Weather	
Outdoor dry bulb temperature	[°C]
Outdoor dry bulb temperature forecast (6 and 12 hr ahead)	[°C]
Direct solar irradiance	[W/m^2]
Direct solar irradiance forecast (6 and 12 hr ahead)	[W/m^2]
District	
Day type	[–]
Hour of day	[h]
Electricity price	[$/kWh]
Electricity price forecast (6 and 12 hr ahead)	[$/kWh]
Building	
Load (Net electricity consumption)	[kW]
Non-shiftable load	[kW]
Solar generation	[kW]
DHW storage SoC (state of charge)	[–]
Battery SoC	[–]

day, electricity price, and their forecasts with horizons of 6 and 12 h. Building-specific states capture the distinctive energy characteristics of each agent. These include the building load (comprising non-shiftable load and cooling/heating demand), PV electricity generation, and the state of charge (SOC) of both the battery and the DHW storage system.

4.3.2. Action-space

The case study investigates the optimization of energy management in a cluster of buildings by controlling the charging and discharging processes of energy storage systems. Each building is equipped with domestic hot water (DHW) storage and a battery, allowing agents to manage these resources flexibly to achieve their desired objectives. Specifically, the controller of each building takes actions represented by a two-dimensional vector, which corresponds to the management of both the DHW and electrical storage systems. In the simulation environment, the action space is continuous, with each controller operating in the range $[-1, 1]$. These values correspond to the proportion of the storage systems' capacities to be charged (positive values) or discharged (negative values). Agents regulate the percentage of storage utilization or accumulation, with DHW storage showing notable variability across buildings in both CA and TX, meaning the same action can result in different energy quantities depending on the building.

To improve control over the charging and discharging processes, the action space used in the AAC-MADRL has been discretized, allowing for a more realistic regulation of storage levels. Although the underlying environment operates in a continuous space, partitioning it into distinct classes, as detailed in Section 3, facilitates more effective management of energy storage.

4.4. Design of the reward function

The control problem is framed using a reward function parameterized by β, which represents the trade-off between individual and collective optimization goals. The reward function comprises two distinct components:

- Individual component

$$\min(-p^3 e_i^3, 0)$$

This term is concerned with the minimization of energy costs for each building individually. In this context, e_i represents the net electricity consumption, while p reflects the dynamic pricing of electricity imports from the grid. The objective is to incentivize buildings to reduce their energy imports, particularly during periods of higher prices. The cubic transformation amplifies the penalty for both large energy imports and high price periods, encouraging buildings to significantly reduce consumption when grid prices peak. Moreover, the choice of the cubic transformation was based on the results of [49], which show that increasing the exponent of net electricity consumption improves agent performance by flattening the demand curve more effectively, with no further improvement beyond an exponent of 3.

- Collective component

$$\frac{\left(\sum_1^N e_i\right)^2}{N^\gamma}$$

This component is designed to promote grid stability by flattening the overall district load profile. The squared term penalizes the total net electricity consumption of the neighborhood, motivating buildings to not only minimize energy imports but also prevent excessive energy exports, which could destabilize the grid. By encouraging efficient use of energy storage and maintaining a balanced energy load, this component reduces the risk of disruptions that could negatively impact the district, in addition to destabilizing the grid. It helps maintain stability during periods of high demand, ensuring that the energy supply remains reliable and minimizing the potential for power outages or grid strain. This approach benefits both the grid and the local community by fostering a more stable energy system. The exponent γ is adjustable to fine-tune the balance between individual and collective contributions.

The overall reward function for each agent is thus expressed as follows:

$$r_i = (1-\beta)\min(-p^3 e_i^3, 0) - \beta \frac{\left(\sum_1^N e_i\right)^2}{N^\gamma} \tag{4.1}$$

5. Results

5.1. Pareto-like front

This section investigates the trade-offs between import costs and district load smoothness in multi-building districts, using Pareto-like graphs to analyze results for both winter and summer seasons across varying neighborhood sizes in the Texas (TX) and California (CA) datasets (Figs. 5.1, A.4 and A.5 of Appendix A). Each point on the graphs corresponds to a distinct configuration, characterized by the control method and the weighting factor β in the reward function. The analysis assumes relative optimality among the three DRL methods. To enhance visualization, configurations deemed relatively Pareto-optimal are highlighted with darker colors, distinguishing them from non-Pareto points. Notably, the AAC-MADRL algorithm consistently outperforms the other methods across various district sizes and values of β, producing configurations that dominate in terms of trade-offs between the individual and collective components of the reward function.

The Pareto-like graph results can be physically interpreted by associating the two components of the reward function with the average cost per building within the district and the variance of the district load, as illustrated in Fig. 5.2(a) and (b). The AAC-MADRL consistently occupies a position in the lower left of the graph across all values of β and neighborhood sizes. This positioning indicates an effective balance between cost and load variance, thereby achieving near-optimal performance across different configurations. However, as β approaches 1 and the size of the neighborhood increases, the performance of all algorithms deviates from the optimal region, resulting in greater load variance and higher costs. However, the underlying cause of this trend can be traced back to the structure of the reward function. When $\beta = 1$, the reward focuses exclusively on the second term of Eq. (4.1) that lacks any consideration of the behavior of the individual agent. Consequently, each

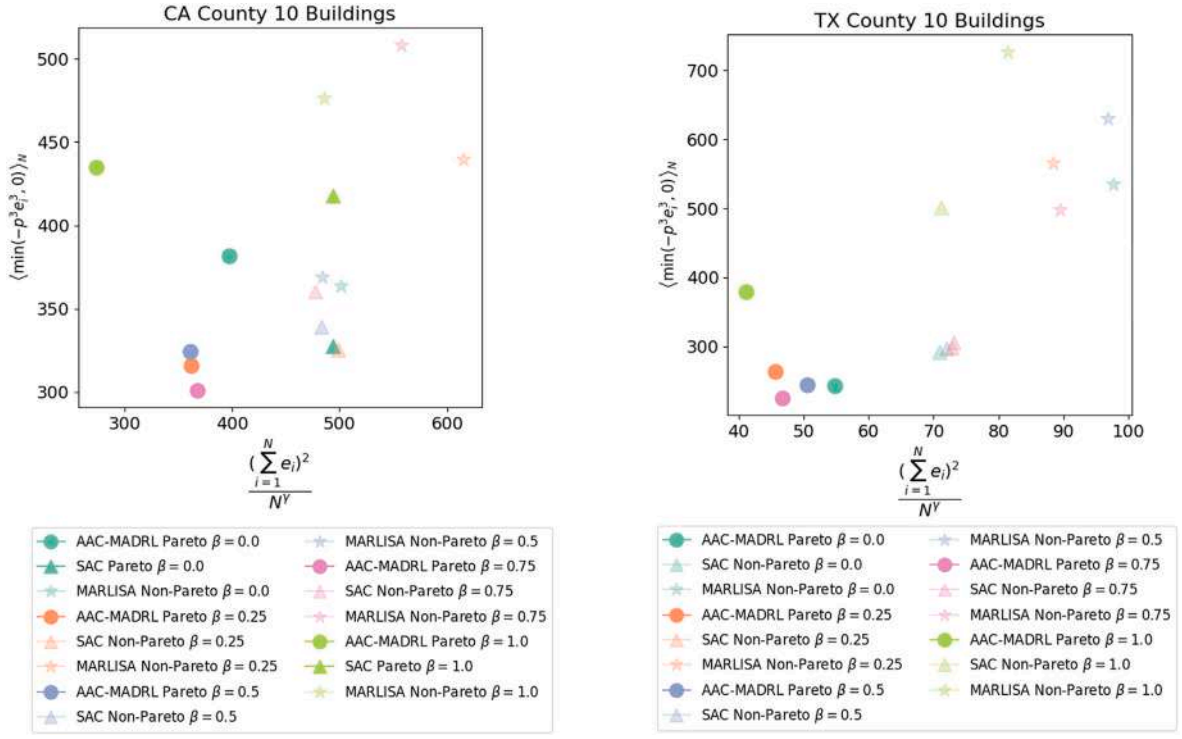


Fig. 5.1. Pareto-like front for CA and TX County in districts of 10 Buildings.

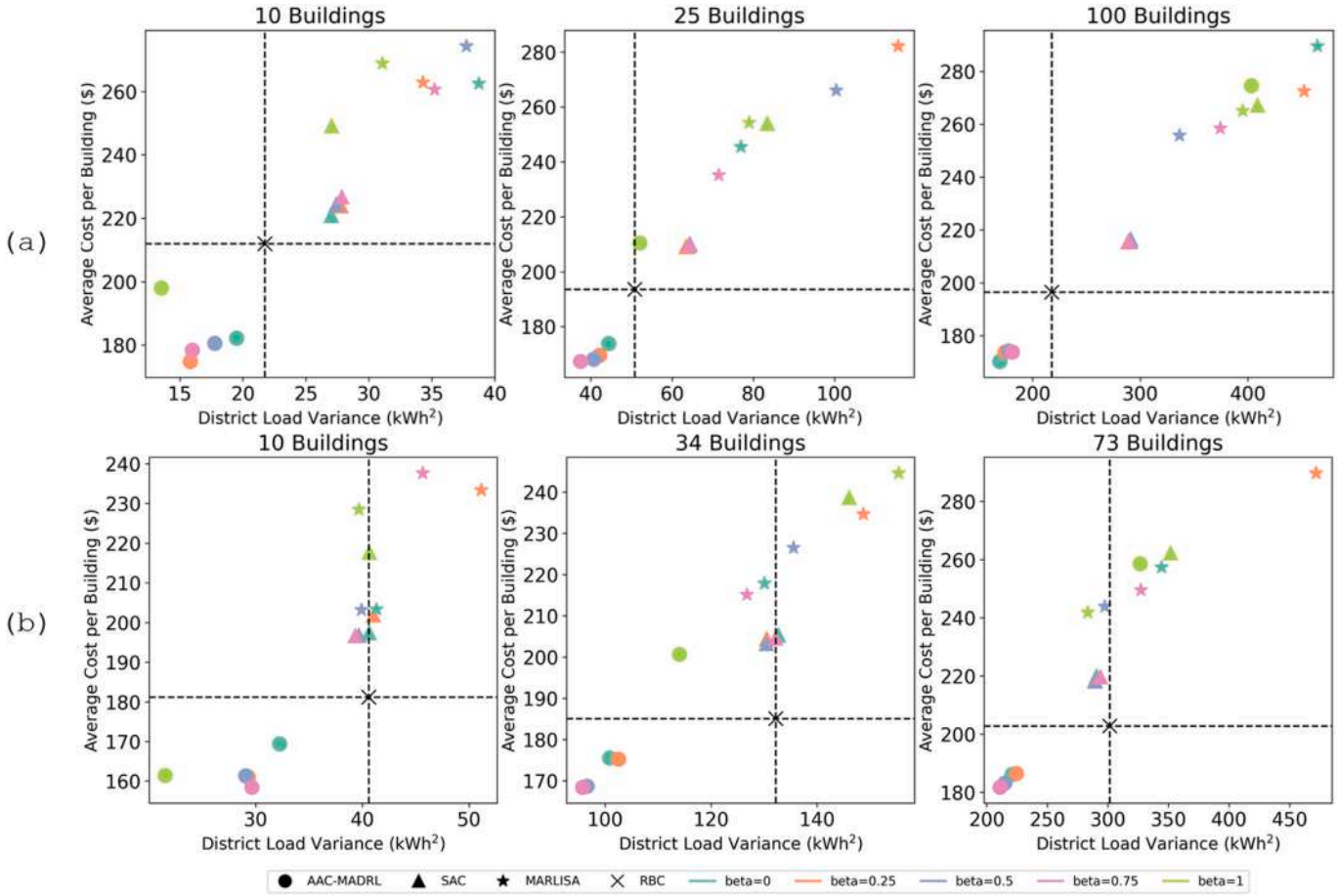


Fig. 5.2. Average cost per building in the district and load variance trade-off for different district sizes and β values. (a) TX County. (b) CA County.

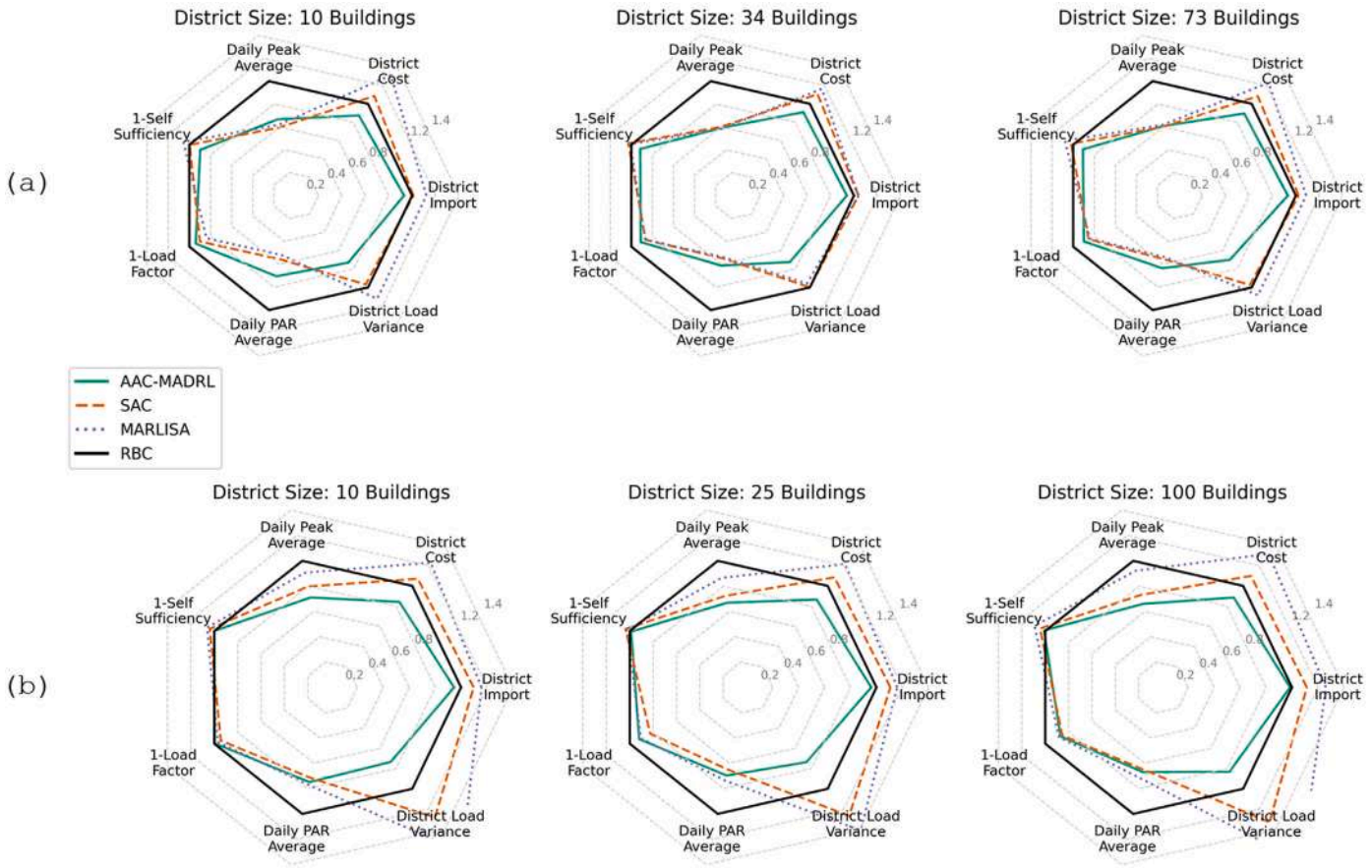


Fig. 5.3. KPIs for different building sizes with $\beta = 0.75$.

agent receives the same reward regardless of its specific impact, resulting in the algorithms rapidly converging to a steady reward value. As a result, the optimization dynamics is mainly influenced by the size of the district rather than the actions of individual buildings. This makes it challenging for the algorithms to adapt to fluctuations in the district's internal conditions, ultimately resulting in suboptimal outcomes for both objectives.

This effect is particularly evident in scenarios with high variability in building loads, where all agents receive the same reward despite inherent differences in their actions. This limitation is evident in the TX dataset, which exhibits greater variability in building loads compared to the CA dataset (Fig. A.3 of Appendix A). However, for districts with up to 10 buildings, the AAC-MADRL algorithm demonstrates superior performance by effectively optimizing for self-sufficiency, storage management, and thus minimizing both imports and costs (see Fig. A.6 of Appendix A).

5.2. Performance evaluation over different strategies

This section assesses the efficacy of the proposed cooperative AAC-MADRL, compared to alternative control strategies within the context of district optimization. In light of the insights from Section 5.1, the evaluation is centered on configurations with $\beta \leq 0.75$. Fig. 5.3(a) and (b) illustrate the results at $\beta = 0.75$ across a range of building sizes. Fig. 5.3(a) represents the Texas (TX) dataset, while Fig. 5.3(b) represents the California (CA) dataset. The figures provide a comprehensive overview of the performance of each algorithm to key performance indicators (KPIs) under different climate conditions.

The radar plots demonstrate that AAC-MADRL consistently exhibits superior performance relative to other algorithms across a range of climate scenarios and district sizes, achieving more favorable outcomes for key performance indicators (KPIs). In comparison to the reference baseline control (RBC), all deep reinforcement learning algorithms have been observed to reduce the average daily peak, the daily peak-to-average ratio (Daily PAR), and the 1-load factor (1-LF) metrics. In particular, AAC-MADRL exhibits a pronounced increase in the peak reduction relative to RBC within the TX dataset, achieving reductions of 18 %–38 %, compared to 1 %–13.5 % for MARLISA and 20 %–28.5 % for SAC. Instead, the daily peak reduction for the CA dataset is more consistent across the DRL algorithms reaching approximately 40 % relative to the RBC. However, AAC-MADRL achieves these performance improvements while also enhancing cost efficiency by reducing grid imports, leading to greater self-sufficiency. Specifically, AAC-MADRL reduces costs by 9 %–12.5 % in various district sizes in CA, while achieving a reduction of 10 %–18 % in TX. Additionally, AAC-MADRL improves load stability by reducing fluctuations by approximately 30 % in CA and 20 % in TX. A more detailed presentation of the results, categorized by values of β and strict size, is available in Table A.1 and A.2 of Appendix A.

This enhanced performance can be attributed to the optimized use of storage resources. Figs. 5.4 and 5.5 illustrate the district load profiles respectively for the last three days of August (TX) and March 8–10 (CA),

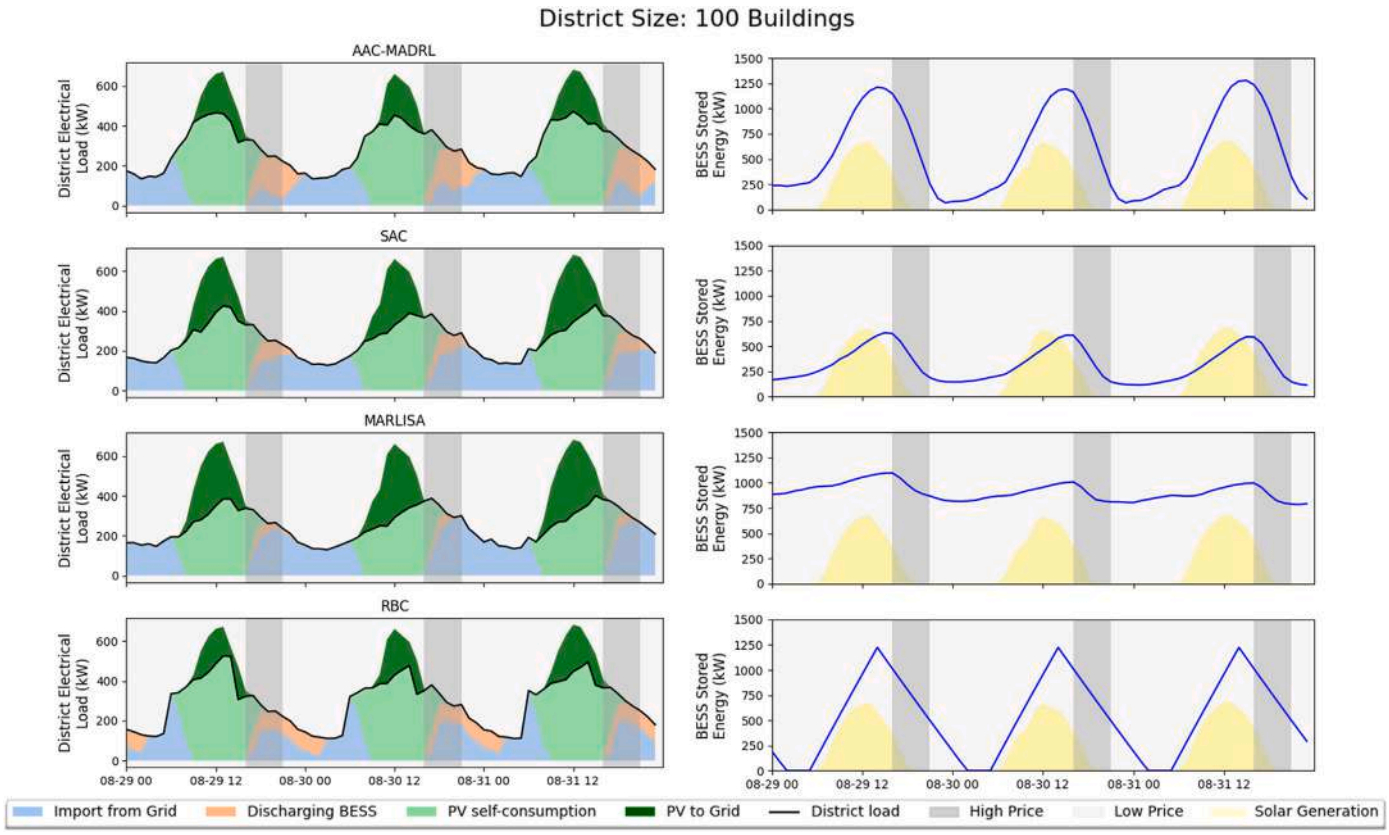


Fig. 5.4. TX district load profile for control strategic, $\beta = 0.75$.

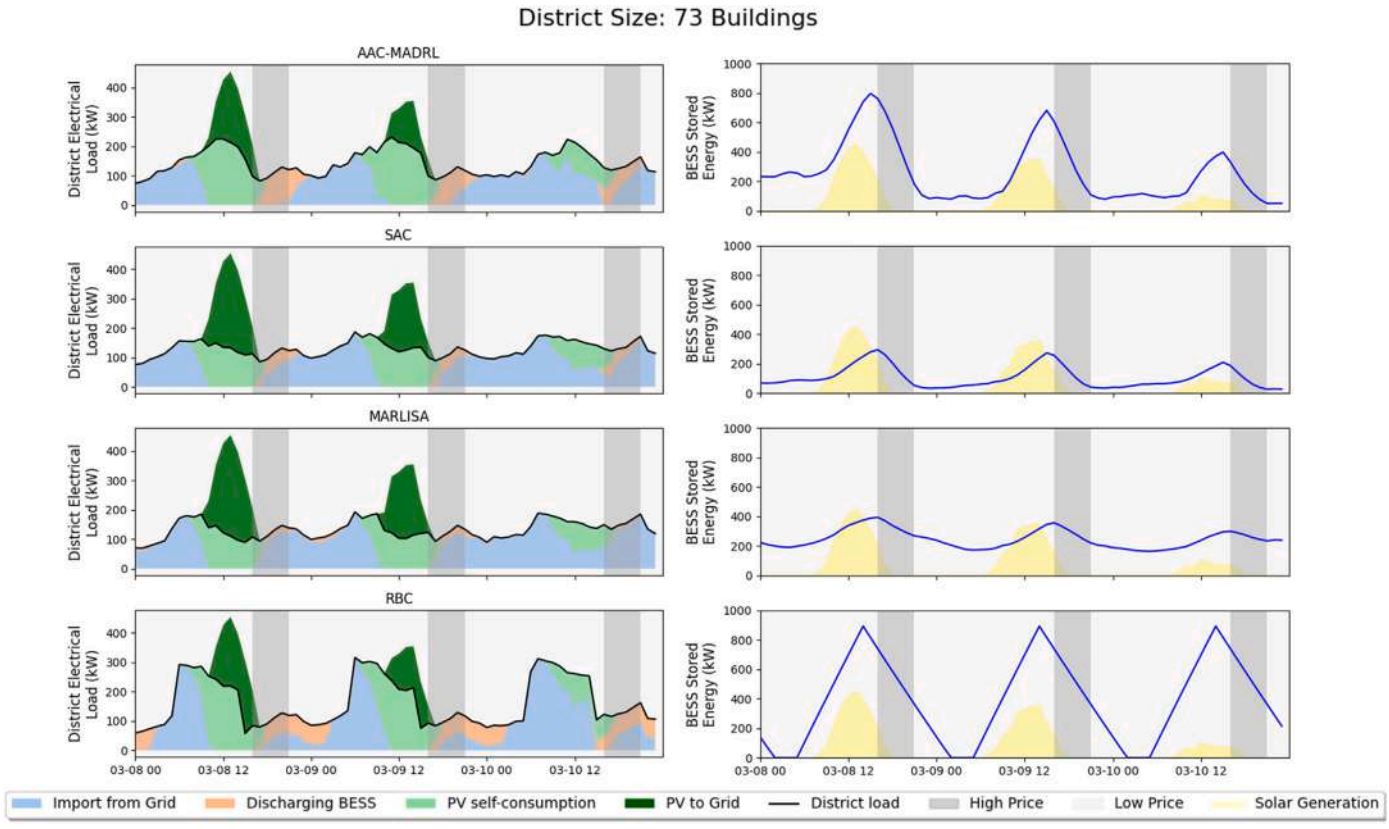


Fig. 5.5. CA district load profile for control strategic, $\beta = 0.75$.

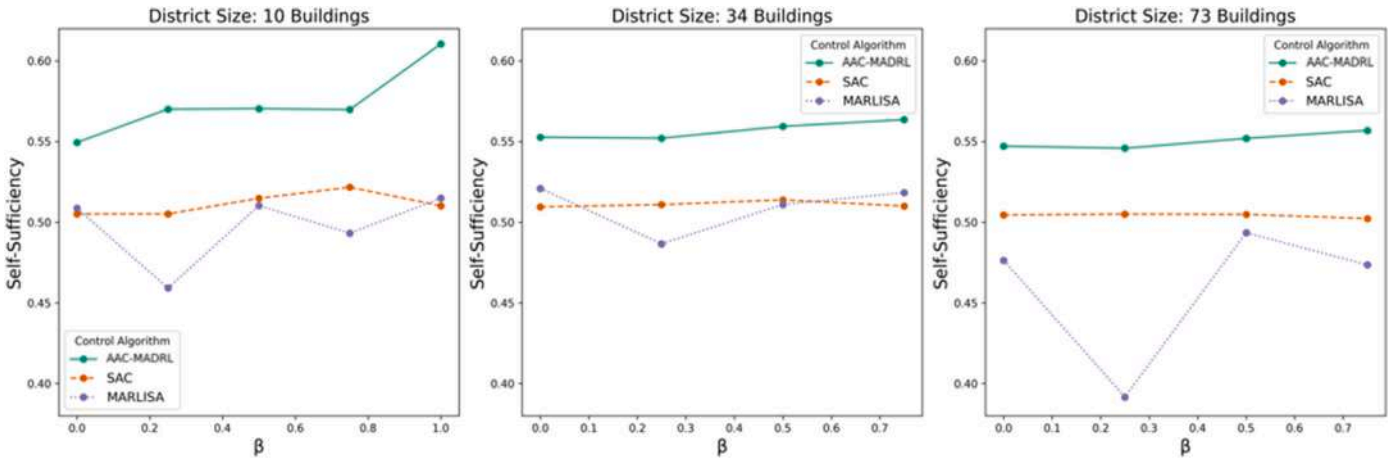


Fig. 5.6. Self-sufficiency at different β values and district sizes for CA County. The AAC-MADRL demonstrates higher levels of self-sufficiency, with these levels increasing as β increases. This is due to the reward structure placing greater emphasis on collective outcomes as β rises, thereby aligning the agents' actions more closely with district-wide objectives.

Table 5.1
Per-agent actions computation time (in seconds) for different algorithms (SAC, AAC-MADRL, MARLISA, and RBC) across counties and district sizes.

County	District size	SAC (s)	AAC-MADRL (s)	MARLISA (s)	RBC
TX	10	0.03	0.02	0.05	0.02
	25	0.08	0.04	0.13	0.03
	100	0.29	0.16	0.43	0.16
CA	10	0.03	0.02	0.05	0.02
	34	0.1	0.06	0.15	0.04
	73	0.2	0.12	0.36	0.09

highlighting the differences in the proportion of photovoltaic energy used by the district versus that exported to the grid. These figures also show the levels of energy stored in the battery energy storage systems. The AAC-MADRL demonstrates better performance during both winter and summer periods, effectively utilizing PV production to facilitate greater energy storage during low-price periods and enhanced energy discharge during high-tariff periods compared to other DRL approaches, allowing for less import needs. Furthermore, during the winter months, AAC-MADRL results in a more pronounced reduction in the overall load of the district in comparison to the baseline reference control (RBC), which tends to overcharge the battery from the grid even when it is not cost-effective. This optimized strategy enhances self-sufficiency in the case of Alameda County for the AAC-MADRL (Fig. 5.6) achieving increases of 6 %–10.5 %. In contrast, this improvement is not observed in the Texas case study, where AAC-MADRL's self-sufficiency levels remain comparable to those of the RBC during the summer months, primarily due to higher photovoltaic (PV) generation (Fig. A.1 and Section 4.1). However, this dynamic changes in winter, when the DRL algorithms have greater flexibility to optimize operations. The AAC-MADRL, for instance, adjusts the battery charging to maintain sufficient energy levels to meet the load, thereby avoiding expensive imports from the grid. In contrast, the SAC and MARLISA algorithms are unable to learn the optimal strategy in situations where there is a margin for optimization, resulting in poorer performance compared to the RBC in both summer and winter.

In addition to these performance benefits, AAC-MADRL also demonstrates faster deployment times compared to SAC and MARLISA algorithms (Table 5.1), with compatible training times (Table A.3). The table demonstrates that computational time increases for all control strategies with district size, partly due to the computational complexity introduced by the simulation environment. As district size grows, more state metadata must be processed, and synchronization delays also increase. For DRL algorithms, the increase in computational time is further augmented by the reward design, which includes a global term that sums the electricity consumption of all agents. All simulations are run on a computing platform equipped with an Intel Core i9 processor at 3.70 GHz with 128 GB of RAM.

Additional results for other β values, which demonstrate consistent results across various objectives, are available in the Appendix A (Figs. A.7–A.9).

6. Discussion and conclusion

The results demonstrate that the AAC-MADRL cooperative architecture outperforms both decentralized approaches (SAC) and cooperative DRL models without attention (MARLISA). This advantage holds across various settings, including distinct seasonal conditions, different neighborhood sizes, and a range of β values that govern the cooperative weight in the reward structure. In particular, the attention-based model demonstrates consistent performance across a range of reward structures, from fully cooperative ($\beta = 1$, where all algorithms perform suboptimally beyond 10 buildings) to mixed ($0 < \beta < 1$) and fully independent reward structures ($\beta = 0$). This adaptability is critical in energy management applications, where environmental conditions and district configurations vary.

One key insight from this study is that the attention mechanism in AAC-MADRL enables a level of agent coordination that enhances performance over the decentralized architecture, even when $\beta = 0$. In this fully independent reward structure, where agents would typically act without regard for others, the coordination facilitated by the attention mechanism still leads to improved outcomes. While a decentralized DRL framework might encourage each agent to optimize solely for its gain, this approach neglects the inherent interconnectedness of district energy systems. In such settings, the agents are not isolated entities but share a common grid infrastructure and often rely on the same energy storage or distribution resources. Consequently, even in scenarios where agents are incentivized to operate independently, as under $\beta = 0$, coordination based on awareness of other agents' actions offers a natural advantage. The attention mechanism effectively facilitates this by allowing each agent to focus on relevant interactions, thus optimizing resource allocation, load balancing, and overall energy efficiency. Furthermore, the decentralized architecture appears less adaptable to changes in the reward structure, especially when district size increases. In contrast to AAC-MADRL, whose performance adjusts according to different values of β (see Fig. 5.6 for an illustrative example), the results for the decentralized approach remain almost invariant, regardless of whether the reward structure is fully independent, mixed, or fully cooperative. This underscores the limited flexibility of decentralized models in responding to varying levels of collaboration within the system.

Compared to cooperative RL architectures that do not incorporate attention (MARLISA), the AAC-MADRL offers additional flexibility through its capacity to adapt dynamically to inter-agent interactions. In traditional cooperative RL frameworks without attention, all agents are typically treated as equally relevant, which can result in a less differentiated decision-making process and hinder adaptability. In contrast, the attention-based architecture selectively emphasizes the influence of specific agents or groups, allowing it to capture and respond to context-dependent interactions within the district. This adaptability is especially beneficial in scenarios where load and storage demands are more dynamic, as seen in the winter dataset. An important distinction is that MARLISA prevents potential privacy concerns by relying exclusively on aggregated and predicted variables, rather than accessing direct inter-agent information. While this ensures a privacy-preserving approach, it can limit performance compared to models like AAC-MADRL which incorporate richer inter-agent information. However, a trade-off to consider with architectures like AAC-MADRL is their reliance on inter-agent communication, which, if not efficiently managed, could potentially lead to implementation challenges such as communication bottlenecks.

The cooperative AAC-MADRL enhances inter-agent coordination by leveraging shared infrastructure, adapts to diverse operating conditions through selective focus on relevant interactions, and consistently outperforms both decentralized DRL and cooperative RL without attention across a range of reward structures and district sizes (Figs. 5.3–5.5 and A.6–A.9 of Appendix A). The findings highlight the potential of attention-based architectures to drive more efficient and resilient energy management strategies in interconnected, multi-agent systems.

7. Limitations and future work

Although this study demonstrates the effectiveness of the attention-based cooperative DRL architecture (AAC-MADRL) in various district energy management settings, several limitations provide opportunities for future enhancement and exploration.

First, optimizing training time remains a significant area for improvement. Although the current implementation successfully handles moderate district sizes, training time could be further reduced, enhancing the scalability and potentially allowing real-time application of the algorithm in larger district energy systems.

Additionally, the AAC-MADRL framework's unique agent MLP structures can be decoupled from the attention mechanism. By transmitting only the embeddings, direct inter-agent communication can be minimized, preserving privacy while maintaining coordination. This modification could expand the applicability of AAC-MADRL in privacy-sensitive environments without compromising performance. An alternative strategy for improving both scalability and privacy is federated learning [57,58], which facilitates decentralized training while minimizing communication overhead.

Beyond these architectural improvements, a comparative evaluation against other attention-based DRL architectures would provide a more thorough assessment of AAC-MADRL's relative strengths and potential areas for refinement.

Finally, while the present study evaluated the algorithm in districts of varying sizes and environmental conditions, future research should investigate its performance in larger district configurations and across even more diverse datasets. Testing on larger districts would not only validate the algorithm's scalability but also provide insights into its robustness under a wider range of real-world conditions. Moreover, experimentation with datasets representing disparate climates, building types, load profiles, and price schedules could assist in determining the model's generalizability and uncover any additional optimization needs specific to regional characteristics.

CRediT authorship contribution statement

Sabrina Savino: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Tommaso Minella:** Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zoltán Nagy:** Writing – review & editing, Validation, Methodology, Investigation. **Alfonso Capozzoli:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Conceptualization.

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.

Acknowledgements

The work of Sabrina Savino is part of the project PNRR-NGEU- CUP: E14D23001820006 which has received funding from the MUR – DM 118/2023.

The work of Alfonso Capozzoli was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)–MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3–D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Appendix A

This appendix presents supplementary figures and tables that support and extend the findings discussed in the main text. These materials include further details regarding the datasets used in the study (Figs. A.1–A.3), additional results for various values of the parameter β (see Figs. A.4–A.9 and Tables A.1 and A.2), and comparison of the computational training times of the different algorithms (Table A.3).

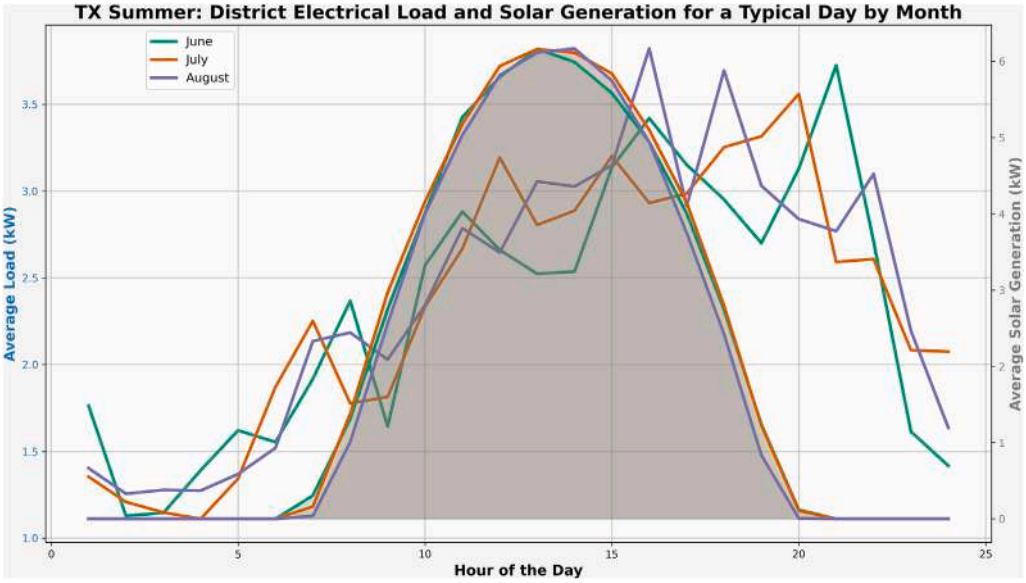


Fig. A.1. Electrical load on a typical summer day in a 100-building district, Texas, TX.

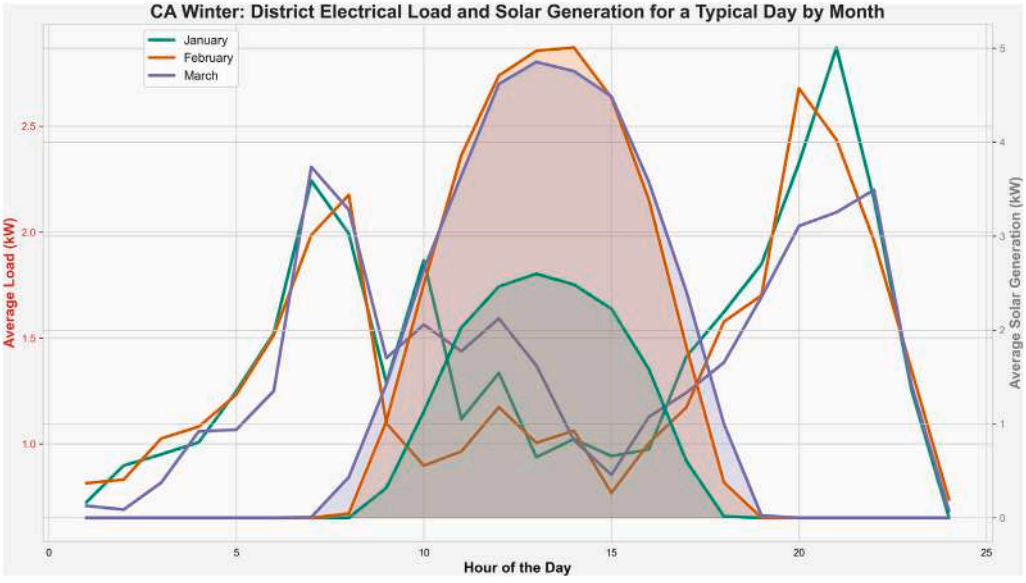


Fig. A.2. Electrical load on a typical winter day in a 73-building district, Alameda, CA.

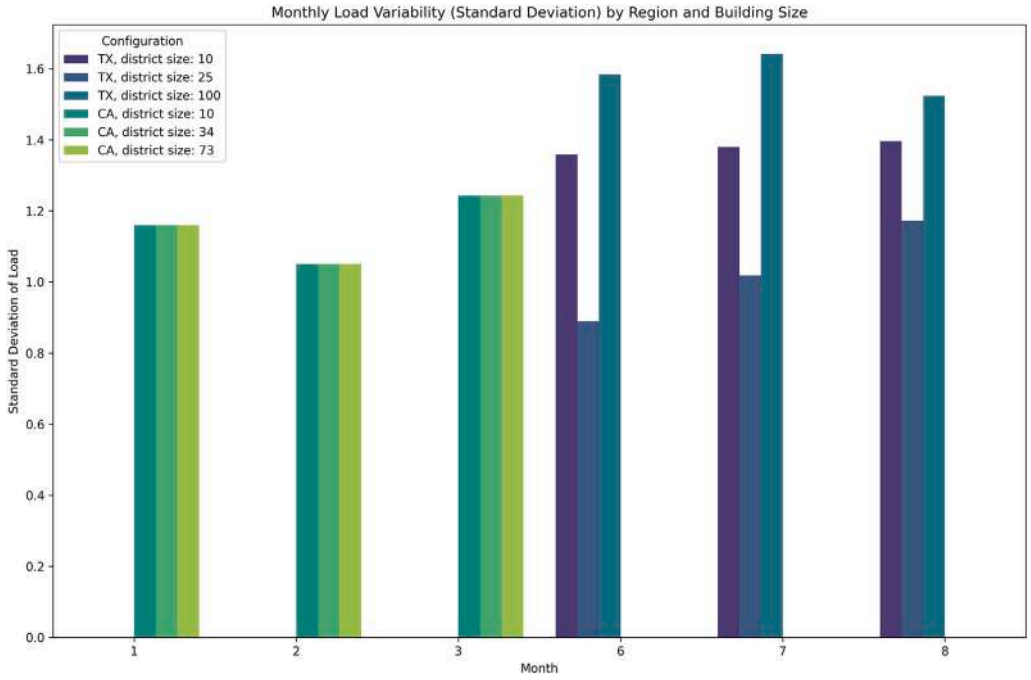


Fig. A.3. Electrical load variability by month and size for CA (January–March) and TX (June–August).

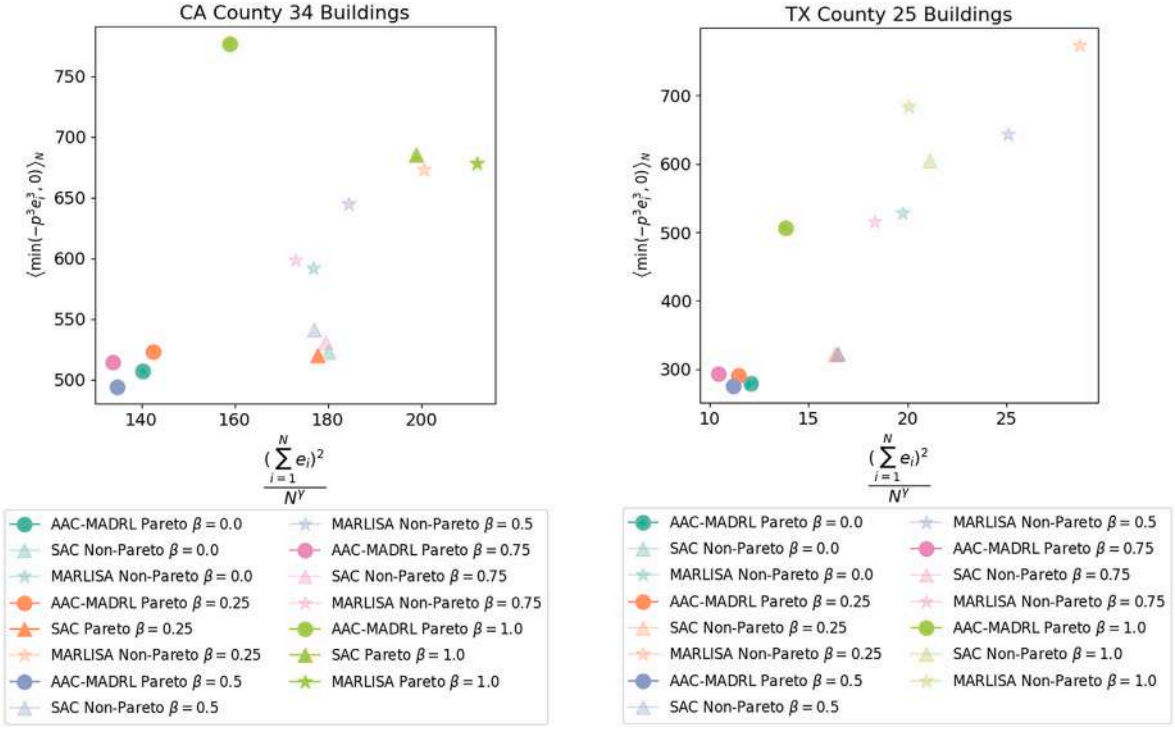


Fig. A.4. Pareto-like front for CA (34 buildings) and TX (25 buildings).

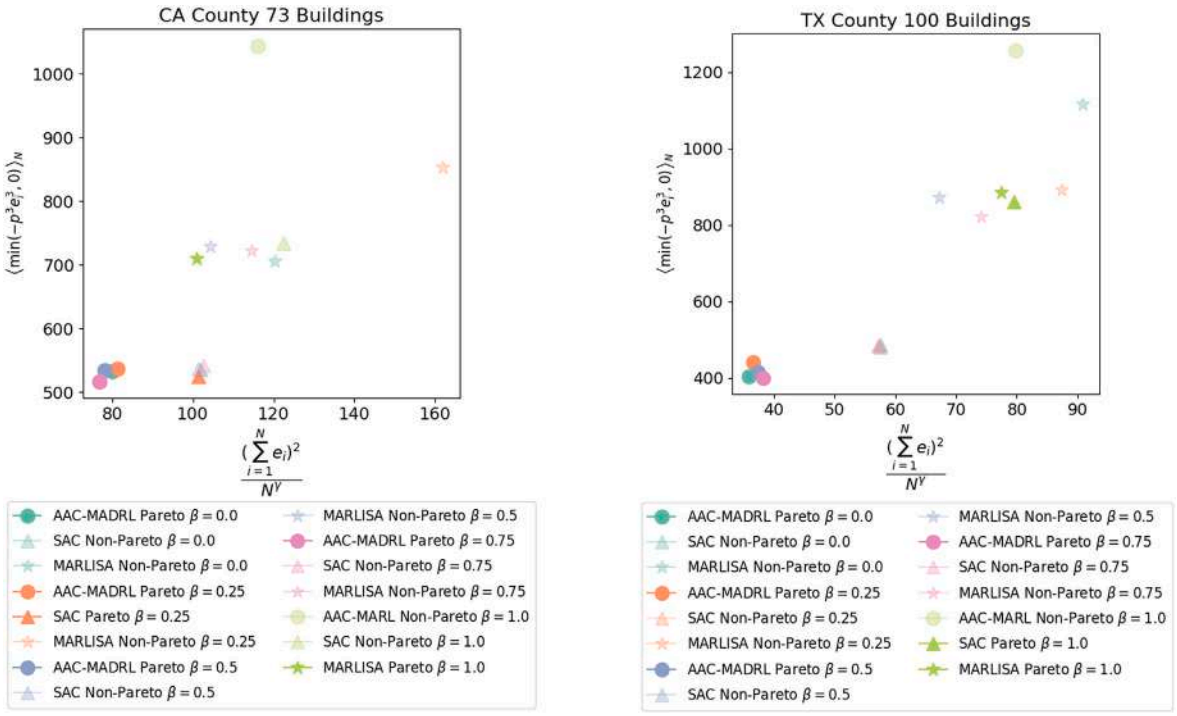


Fig. A.5. Pareto-like front for CA County and TX County.

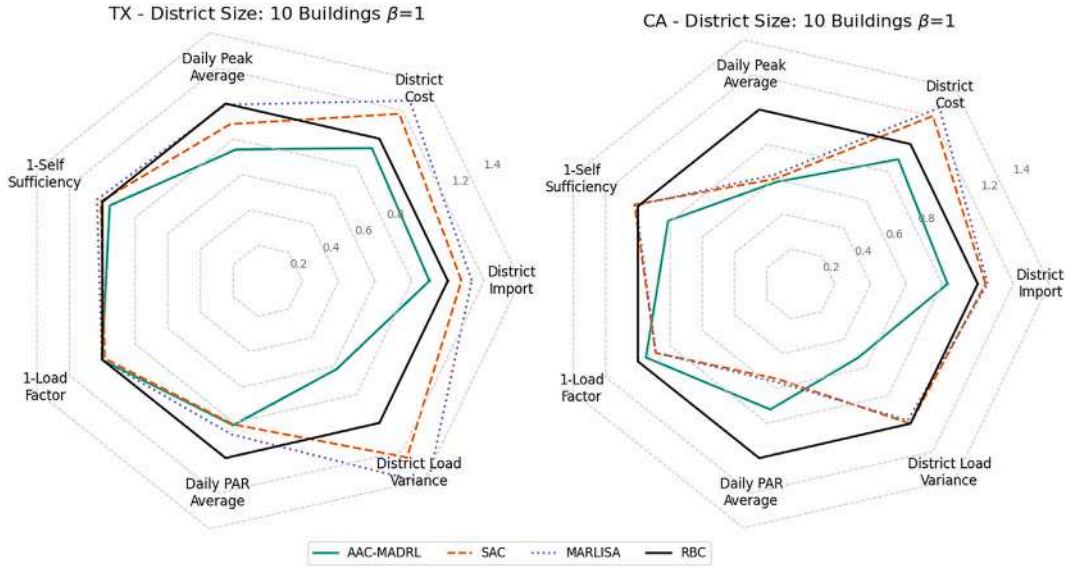


Fig. A.6. KPIs for $\beta = 1$ and district size: 10 buildings for TX and CA.

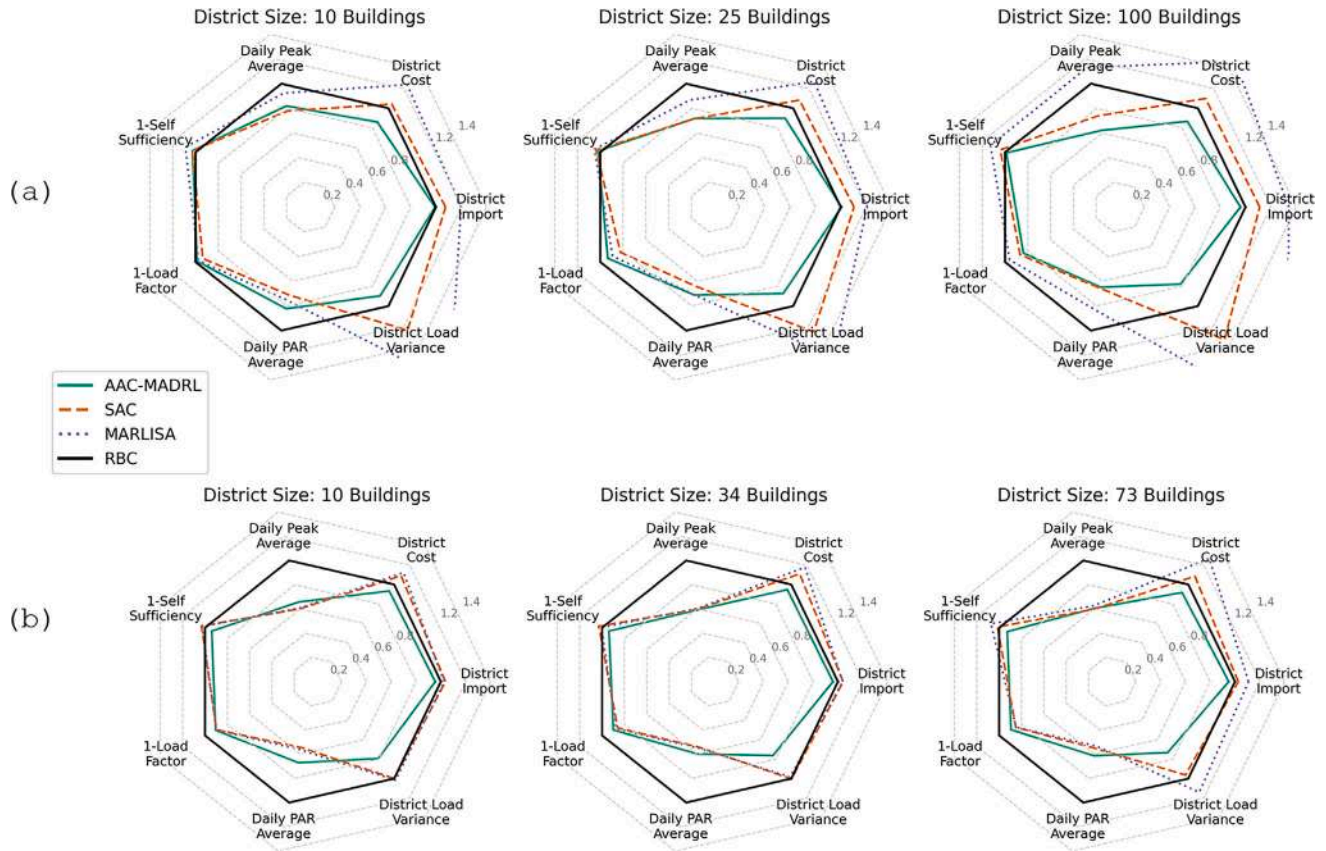


Fig. A.7. KPIs for different building sizes with $\beta = 0$.

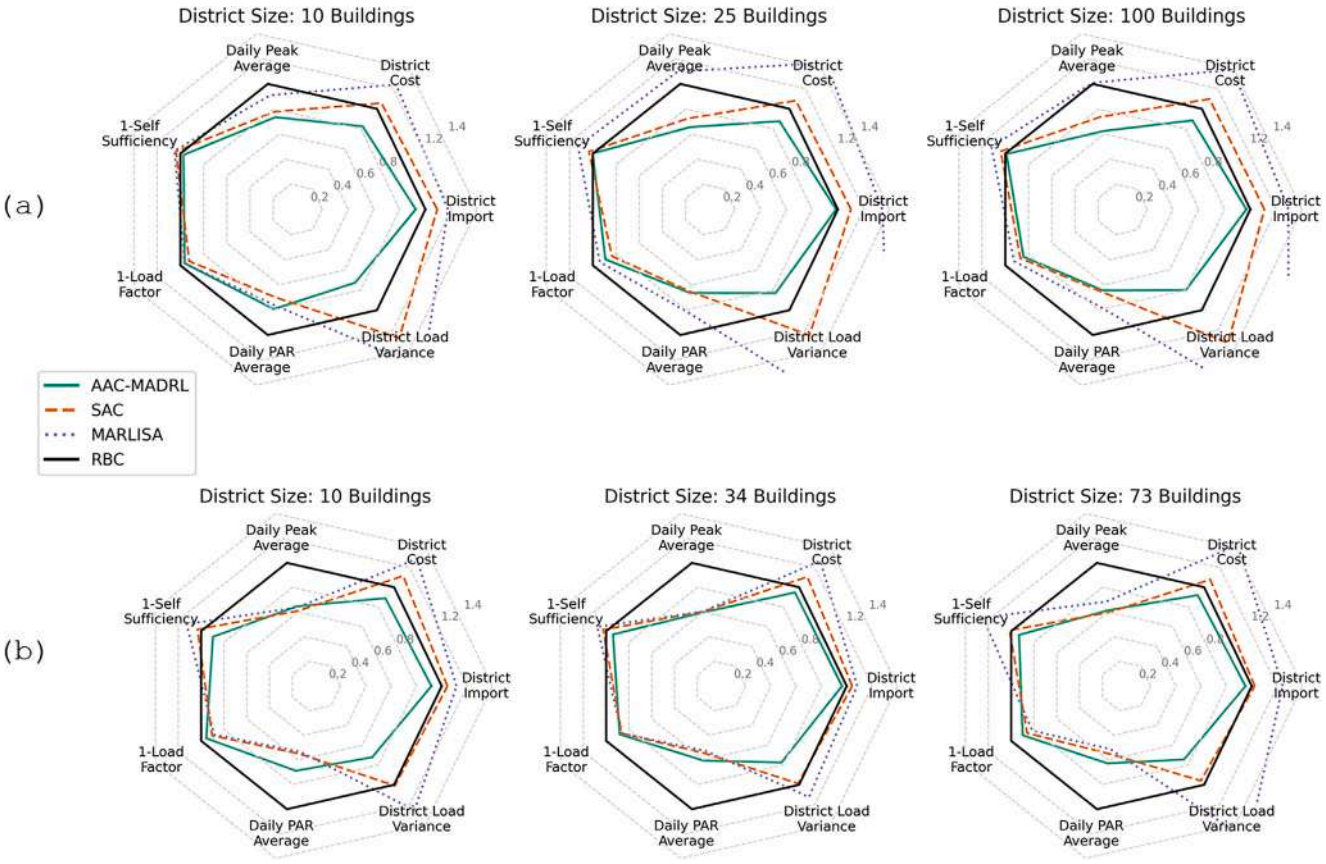


Fig. A.8. KPIs for different building sizes with $\beta = 0.25$.

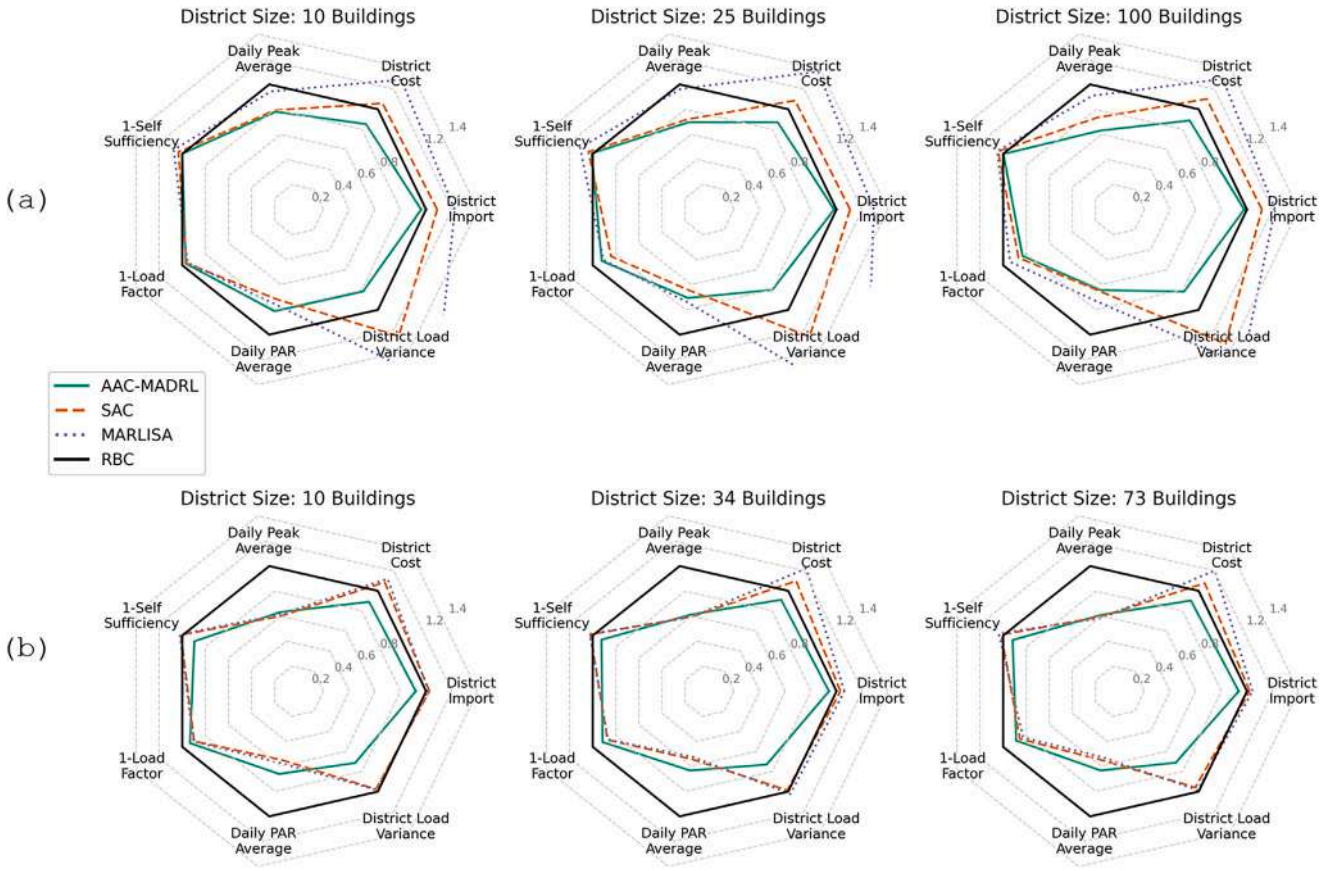


Fig. A.9. KPIs for different building sizes with $\beta = 0.5$.

Table A.1
Percentage change in KPIs relative to RBC across different β values and district sizes in TX.

KPI	Algorithm	District size 10				District size 25				District size 100			
		β_0	$\beta_{0.25}$	$\beta_{0.5}$	$\beta_{0.75}$	β_0	$\beta_{0.25}$	$\beta_{0.5}$	$\beta_{0.75}$	β_0	$\beta_{0.25}$	$\beta_{0.5}$	$\beta_{0.75}$
Cost (+/−)	AAC-MADRL	−14 %	−17.5 %	−14.8 %	−15.8 %	−10.2 %	−12.4 %	−13.1 %	−13.6 %	−13.3 %	−11.5 %	−11.3 %	−11.5 %
	SAC	+4.2 %	+5.7 %	+5.9 %	+7 %	+8.2 %	+8.1 %	+8.6 %	+8.4 %	+9.8 %	+9.7 %	+10.2 %	+9.8 %
	MARLISA	+23.8 %	+24 %	+29.4 %	+22.9 %	+26.8 %	+45.8 %	+37.4 %	+21.5 %	+47.4 %	+38.7 %	+30.2 %	+31.5 %
Load variance (+/−)	AAC-MADRL	−10.3 %	−27.3 %	−18.4 %	−26.5 %	−12.6 %	−17.1 %	−20 %	−26.2 %	−22.1 %	−20 %	−18.2 %	−16.7 %
	SAC	+24.3 %	+27.8 %	+26.1 %	+28.1 %	+26 %	+24.9 %	+26.7 %	+26.6 %	+33.7 %	+32.7 %	+33.7 %	+33 %
	MARLISA	+78.1 %	+57.7 %	+73.6 %	+62 %	+51.5 %	+127.8 %	+97.8 %	+40.8 %	+113.4 %	+107.8 %	+54.6 %	+72 %
Daily peak (+/−)	AAC-MADRL	−18 %	−26.5 %	−21.8 %	−29 %	−28.3 %	−34.4 %	−30.2 %	−33.1 %	−37.8 %	−37.5 %	−36.8 %	+34.1 %
	SAC	−22.1 %	−22.2 %	−20.6 %	−20.3 %	−28.4 %	−27.4 %	−27.8 %	−27.6 %	−26 %	−26.3 %	−26.4 %	−27 %
	MARLISA	−7.6 %	−9 %	−5.6 %	−9.3 %	−13.3 %	−9.8 %	−3.6 %	−13.5 %	−13.1 %	−0.9 %	−9 %	−8 %
Self-sufficiency (+/−)	AAC-MADRL	−2.7 %	+2.6 %	−0.1 %	−0.1 %	−2.1 %	−0.2 %	−0.9 %	+1.1 %	+1.1 %	+1 %	+0.3 %	+1 %
	SAC	−2.7 %	−3.8 %	−3.3 %	−3.8 %%	−3.7 %	−3.4 %	−3.7 %	−3.6 %%	−4.2 %	−4 %	−4.4 %	−4.4 %
	MARLISA	−9 %	−5.1 %	−7.3 %	−6.1 %	−5.2 %	−12.7 %	−10.9 %	−2.4 %	−12.7 %	−12 %	−5.1 %	−9 %

Table A.2
Percentage change in KPIs relative to RBC across different β values and district sizes in CA.

KPI	Algorithm	District size 10				District size 34				District size 73			
		β_0	$\beta_{0.25}$	$\beta_{0.5}$	$\beta_{0.75}$	β_0	$\beta_{0.25}$	$\beta_{0.5}$	$\beta_{0.75}$	β_0	$\beta_{0.25}$	$\beta_{0.5}$	$\beta_{0.75}$
Cost (+/−)	AAC-MADRL	−6.5 %	−11.1 %	−11 %	−12.5 %	−5.1 %	−5.3 %	−9 %	−9 %	−8.2 %	−8 %	−9.7 %	−10.4 %
	SAC	+9 %	+11.5 %	+8.6 %	+8.6 %	+11 %	+10.5 %	+10 %	+10.6 %	+8.5 %	+7.8 %	+7.6 %	+8.4 %
	MARLISA	+12.2 %	+28.8 %	+12.2 %	+31.2 %	+17.7 %	+26.8 %	+22.4 %	+16.3 %	+27 %	+43 %	+20.3 %	+23.1 %
Load variance (+/−)	AAC-MADRL	−20.6 %	−27.8 %	−28.5 %	−27 %	−23.7 %	−22.5 %	−27 %	−27.6 %	−26.8 %	−25.6 %	−28.8 %	−30.1 %
	SAC	+0 %	+1 %	−2.3 %	−3.2 %	+0.3 %	−1.3 %	−1.3 %	0 %	−3.7 %	−4.1 %	−4.1 %	−2.6 %
	MARLISA	+1.7 %	+25.9 %	−1.7 %	+12.4 %	−1.6 %	+12.5 %	+2.5 %	−4.1 %	+14.1 %	+56.4 %	−1.4 %	+8.5 %
Daily peak (+/−)	AAC-MADRL	−34.3 %	−35.3 %	−37 %	−33.3 %	−40.7 %	−40.2 %	−38.8 %	−40.8 %	−39.4 %	−38.5 %	−39 %	−39.3 %
	SAC	−39.7 %	−39.1 %	−40 %	−40.5 %	−40.2 %	−39.6 %	−40.6 %	−40 %	−40 %	−40.4 %	−40.3 %	−40 %
	MARLISA	−39.1 %	−36.2 %	−38.7 %	−38.1 %	−39.6 %	−39.5 %	−40.8 %	−41.4 %	−37.3 %	−31.6 %	−41.6 %	−38.2 %
Self-sufficiency (+/−)	AAC-MADRL	+6.1 %	+10.4 %	+10.4 %	+10.3 %	+6.2 %	+6 %	+7.6 %	+8.4 %	+7.2 %	+7 %	+8.2 %	+9.2 %
	SAC	−3.2 %	−3.2 %	−1.2 %	+0.3 %	−3 %	−2.6 %	−2 %	−2.8 %%	−1.6 %	−1.5 %	−1.5 %	−2 %
	MARLISA	−2.5 %	−12.7 %	−2.1 %	−5.7 %	−0.5 %	−7.7 %	−2.6 %	−1 %	−7.3 %	−24.7 %	−3.8 %	−8 %

Table A.3
Training times for each algorithm across different district sizes. The reported times represent the average training duration across various values of β.

Algorithm	District size				
	10	25	34	73	100
AAC-MADRL	6 h 25 min	23 h 55 min	13 h 17 min	1 d 13 h 2 min	1 d 17 h 30 min
SAC	6 h 35 min	21 h 37 min	11 h 12 min	1 d 6 h 36 min	1 d 7 h 35 min
MARLISA	6 h 11 min	20 h 25 min	10 h 29 min	1 d 2 h 12 min	1 d 4 h 18 min

Data availability

Data will be made available on request.

References

[1] Villar J, Bessa R, Matos M. Flexibility products and markets: literature review. Electr Power Syst Res 2018;154:329–40. doi: https://doi.org/10.1016/j.epsr.2017.09.005.

[2] I. E. Agency. Renewables 2022; 2022. Available from: https://www.iea.org/world/renewables. [Accessed 26 September 2023].

[3] Li S, Lian J, Conejo AJ, Zhang W. Transactive energy systems: the market-based coordination of distributed energy resources. IEEE Control Syst 2020;40(4):26–52. doi: https://doi.org/10.1109/MCS.2020.2990514.

[4] Hargroves K, James B, Lane J, Newman P. The role of distributed energy resources and associated business models in the decentralised energy transition: a review. Energies 2023;16(10):4231. doi: https://doi.org/10.3390/en16104231.

[5] Brahmane AV, Deshmukh SR. Artificial intelligence-based energy management system for renewable energy sources. In: 2023 4th international conference on electronics and sustainable communication systems, ICESC 2023 - Proceedings. Institute of Electrical and Electronics Engineers Inc.; 2023. p. 727–31. doi: https://doi.org/10.1109/ICESC57686.2023.10193623.

[6] Luo T, Ault GW, Galloway SJ. Demand side management in a highly decentralized energy future. In: 45th international universities power engineering conference UPEC2010; 2010. p. 1–6. https://api.semanticscholar.org/CorpusID:21162136.

[7] Nebey AH. Recent advancement in demand side energy management system for optimal energy utilization. Energy Rep 2024;11:5422–35. doi: https://doi.org/10.1016/j.egyr.2024.05.028.

[8] Totare SK, Eaton. Management demand side by regulating charging and discharging of energy storage system and utilizing renewable energy; 2021. https://api.semanticscholar.org/CorpusID:247572442.

[9] Kaspar K, Ouf M, Eicker U. A critical review of control schemes for demand-side energy management of building clusters. Energy Build 2022;257. doi: https://doi.org/10.1016/j.enbuild.2021.111731.

[10] Hu M, Xiao F, Wang S. Neighborhood-level coordination and negotiation techniques for managing demand-side flexibility in residential microgrids; 2021. doi: https://doi.org/10.1016/j.rser.2020.110248.

[11] Panda S, Mohanty S, Rout PK, Sahu BK, Parida SM, Kotb H, et al. An insight into the integration of distributed energy resources and energy storage systems with smart distribution networks using demand-side management. Appl Sci 2022;12(17):8914. doi: https://doi.org/10.3390/app12178914.

[12] Sigrin B, Mooney M, Gleason M, Preus R. Distributed generation market demand characterization. Technical report NREL/TP-6A20-75528. Golden, CO: National Renewable Energy Laboratory (NREL); 2020. https://www.nrel.gov/docs/fy20osti/75528.pdf.

[13] Pedram O, Asadi E, Chenari B, Moura P, da Silva MG. A review of methodologies for managing energy flexibility resources in buildings; 2023. doi: https://doi.org/10.3390/en16176111.

[14] Goldman CV, Zilberstein S. Decentralized control of cooperative systems: categorization and complexity analysis; 2004.

[15] Panait L, Luke S. Cooperative multi-agent learning: the state of the art. Auton Agent Multi-Agent Syst 11:387–434.

[16] Rahman MS, Mahmud MA, OO AM, Pota HR, Hossain MJ. Agent-based reactive power management of power distribution networks with distributed energy generation. Energy Convers Manag 2016;120:120–34. doi: https://doi.org/10.1016/j.enconman.2016.04.091.

[17] Rahman MS, OO AM. Distributed multi-agent based coordinated power management and control strategy for microgrids with distributed energy resources. Energy Convers Manag 2017;139:20–32. doi: https://doi.org/10.1016/j.enconman.2017.02.021.

[18] IEEE Staff. 2011 IEEE 9th international conference on power electronics and drive systems. IEEE; 2011.

[19] Buşoniu L, Babuška R, Schutter BD. A comprehensive survey of multiagent reinforcement learning; 2008. doi: https://doi.org/10.1109/TSMCC.2007.913919.

[20] Charbonnier F, Peng B, Vienne J, Stai E, Morstyn T, McCulloch M. Centralised rehearsal of decentralised cooperation: multi-agent reinforcement learning for the scalable coordination of residential energy flexibility. Appl Energy 2025;377:124406. doi: https://doi.org/10.1016/j.apenergy.2024.124406).

[21] Charbonnier F, Morstyn T, McCulloch MD. Scalable multi-agent reinforcement learning for distributed control of residential energy flexibility. Appl Energy 2022;314. doi: https://doi.org/10.1016/j.apenergy.2022.118825.

[22] Harrold DJ, Cao J, Fan Z. Renewable energy integration and microgrid energy trading using multi-agent deep reinforcement learning. Appl Energy 2022;318. doi: https://doi.org/10.1016/j.apenergy.2022.119151.

[23] Lai BC, Chiu WY, Tsai YP. Multiagent reinforcement learning for community energy management to mitigate peak rebounds under renewable energy uncertainty. IEEE Trans Emerging Top Comput Intel 2022;6(3):568–79. doi: https://doi.org/10.1109/TETCI.2022.3157026.

[24] Pinto G, Deltetto D, Capozzoli A. Data-driven district energy management with surrogate models and deep reinforcement learning. Appl Energy 2021;304. doi: https://doi.org/10.1016/j.apenergy.2021.117642.

[25] Pinto G, Kathirgamanathan A, Mangina E, Finn DP, Capozzoli A. Enhancing energy management in grid-interactive buildings: a comparison among cooperative and coordinated architectures. Appl Energy 2022;310. doi: https://doi.org/10.1016/j.apenergy.2021.118497.

[26] Chen G. A new framework for multi-agent reinforcement learning – centralized training and exploration with decentralized execution via policy distillation; 2019. http://arxiv.org/abs/1910.09152.

[27] Zhou Y, Liu S, Qing Y, Chen K, Zheng T, Huang Y, et al. Is centralized training with decentralized execution framework centralized enough for MARL? 2023. http://arxiv.org/abs/2305.17352.

[28] Jiang J, Lu Z. Learning attentional communication for multi-agent cooperation; 2018. http://arxiv.org/abs/1805.07733.

[29] Peng B, Rashid T, de Witt CAS, Kamienny P-A, Torr PHS, Böhmer W, et al. FACMAC: factored multi-agent centralised policy gradients; 2020. http://arxiv.org/abs/2003.06709.

[30] Rashid T, Samvelyan M, de Witt CS, Farquhar G, Foerster J, Whiteson S. Qmix: monotonic value function factorisation for deep multi-agent reinforcement learning; 2018. http://arxiv.org/abs/1803.11485.

[31] Zhang J, Zhang Y, Zhang XS, Zang Y, Cheng J. Intrinsic action tendency consistency for cooperative multi-agent reinforcement learning; 2024. www.aaai.org.

[32] Luo S, Li Y, Li J, Kuang K, Liu F, Shao Y, et al. S2rl: do we really need to perceive all states in deep multi-agent reinforcement learning? In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery; 2022. p. 1183–91. doi: https://doi.org/10.1145/3534678.3539481.

[33] Bramlage L, Cortese A. Generalized attention-weighted reinforcement learning. Neural Netw 2022;145:10–21. doi: https://doi.org/10.1016/j.neunet.2021.09.023.

[34] Hu K, Xu K, Xia Q, Li M, Song Z, Song L, et al. An overview: attention mechanisms in multi-agent reinforcement learning; 2024. doi: https://doi.org/10.1016/j.neucom.2024.128015.

[35] Zhang B, Hu W, Ghias AM, Xu X, Chen Z. Multi-agent deep reinforcement learning based distributed control architecture for interconnected multi-energy microgrid energy management and optimization. Energy Convers Manag 2023;277. doi: https://doi.org/10.1016/j.enconman.2022.116647.

[36] Chu Y, Wei Z, Fang X, Chen S, Zhou Y. A multiagent federated reinforcement learning approach for plug-in electric vehicle fleet charging coordination in a residential community. IEEE Access 2022;10:98535–48. doi: https://doi.org/10.1109/ACCESS.2022.3206020.

[37] Zhang G, Hu W, Cao D, Zhang Z, Huang Q, Chen Z, et al. A multi-agent deep reinforcement learning approach enabled distributed energy management schedule for the coordinate control of multi-energy hub with gas, electricity, and freshwater. Energy Convers Manag 2022;255. doi: https://doi.org/10.1016/j.enconman.2022.115340.

[38] Hu D, Ye Z, Gao Y, Ye Z, Peng Y, Yu N. Multi-agent deep reinforcement learning for voltage control with coordinated active and reactive power optimization. IEEE Trans Smart Grid 2022;13(6):4873–86. doi: https://doi.org/10.1109/TSG.2022.3185975.

[39] Zhang G, Hu W, Cao D, Zhang Z, Huang Q, Chen Z, et al. A multi-agent deep reinforcement learning approach enabled distributed energy management schedule for the coordinate control of multi-energy hub with gas, electricity, and freshwater. Energy Convers Manag 2022;255. doi: https://doi.org/10.1016/j.enconman.2022.115340.

[40] Bai Y, Chen S, Zhang J, Xu J, Gao T, Wang X, et al. An adaptive active power rolling dispatch strategy for high proportion of renewable energy based on distributed deep reinforcement learning. Appl Energy 2023;330. doi: https://doi.org/10.1016/j.apenergy.2022.120294.

[41] Shao Y, Li R, Zhao Z, Zhang H. Graph attention network-based drl for network slicing management in dense cellular networks. In: IEEE wireless communications and networking conference, WCNC; 2021, March. Institute of Electrical and Electronics Engineers Inc.; 2021. doi: https://doi.org/10.1109/WCNC49053.2021.9417321.

[42] Shao Y, Li R, Hu B, Wu Y, Zhao Z, Zhang H. Graph attention network-based multi-agent reinforcement learning for slicing resource management in dense

cellular network. IEEE Trans Veh Technol 2021;70(10):10792–803. doi: https://doi.org/10.1109/TVT.2021.3103416.
[43] Zhu D, Yang B, Liu Y, Wang Z, Ma K, Guan X. Energy management based on multi-agent deep reinforcement learning for a multi-energy industrial park. Appl Energy 2022;311. doi: https://doi.org/10.1016/j.apenergy.2022.118636.
[44] Xie J, Ajagekar A, You F. Multi-agent attention-based deep reinforcement learning for demand response in grid-responsive buildings. Appl Energy 2023;342. doi: https://doi.org/10.1016/j.apenergy.2023.121162.
[45] Ajagekar A, Decardi-Nelson B, You F. Energy management for demand response in networked greenhouses with multi-agent deep reinforcement learning. Appl Energy 2024;355. doi: https://doi.org/10.1016/j.apenergy.2023.122349.
[46] Ye Y, Tang Y, Wang H, Zhang XP, Strbac G. A scalable privacy-preserving multi-agent deep reinforcement learning approach for large-scale peer-to-peer transactive energy trading. IEEE Trans Smart Grid 2021;12(6):5185–200. doi: https://doi.org/10.1109/TSG.2021.3103917.
[47] Iqbal S, Sha F. Actor-attention-critic for multi-agent reinforcement learning; 2018. http://arxiv.org/abs/1810.02912.
[48] Haarnoja T, Zhou A, Hartikainen K, Tucker G, Ha S, Tan J, et al. Soft actor-critic algorithms and applications; 2018. arXiv abs/1812.05905. https://api.semanticscholar.org/CorpusID:55703664.
[49] Vazquez-Canteli JR, Henze G, Nagy Z. Marlisa: multi-agent reinforcement learning with iterative sequential action selection for load shaping of grid-interactive connected buildings. In: BuildSys 2020—Proceedings of the 7th ACM international conference on systems for energy-efficient buildings, cities, and transportation. Association for Computing Machinery, Inc.; 2020. p. 170–9. doi: https://doi.org/10.1145/3408308.3427604.
[50] Almilaify Y, Nweye K, Nagy Z. Scalex: scalability exploration of multi-agent reinforcement learning agents in grid-interactive efficient buildings. In: BuildSys 2023—Proceedings of the 10th ACM international conference on systems for energy-efficient buildings, cities, and transportation. Association for Computing Machinery, Inc.; 2023. p. 261–4. doi: https://doi.org/10.1145/3600100.3623749.
[51] Nweye K, Kaspar K, Buscemi G, Pinto G, Li H, Hong T, et al. A framework for the design of representative neighborhoods for energy flexibility assessment in CityLearn. In: Building simulation conference proceedings, vol. 18. International Building Performance Simulation Association; 2023. p. 1814–21. doi: https://doi.org/10.26868/25222708.2023.1404.
[52] I.E. Lab. CityLearn: OpenAI Gym environment for urban energy management; 2024. Available from: https://github.com/intelligent-environments-lab/CityLearn. [Accessed 9 October 2024].
[53] Vazquez-Canteli J. R., Dey S., Henze G., Nagy Z., CityLearn: Standardizing research in multi-agent reinforcement learning for demand response and urban energy management, 2020. arXiv:2012.10504.
[54] Vaswani A, Brain G, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. Attention is all you need. Adv Neural Inform Process Syst 2017;30:5998–6008.
[55] Ansi/Ashrae addendum a to ansi/ashrae standard 169-2013; 2020. www.ashrae.org.
[56] Nweye K, Kaspar K, Buscemi G, Fonseca T, Pinto G, Ghose D, et al. CityLearn v2: energy-flexible, resilient, occupant-centric, and carbon-aware management of grid-interactive communities. J Build Perform Simul 2025;18:17–38. doi: https://doi.org/10.1080/19401493.2024.2418813.
[57] Chu Y, Wei Z, Fang X, Chen S, Zhou Y. A multiagent federated reinforcement learning approach for plug-in electric vehicle fleet charging coordination in a residential community. IEEE Access 2022;10:98535–48. doi: https://doi.org/10.1109/ACCESS.2022.3206020.
[58] Tang L, Xie H, Wang X, Bie Z. Privacy-preserving knowledge sharing for few-shot building energy prediction: a federated learning approach. Appl Energy 2023;337. doi: https://doi.org/10.1016/j.apenergy.2023.120860.