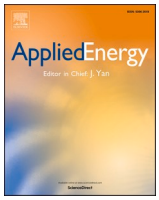
Applied Energy 342 (2023) 121162

Contents lists available at ScienceDirect

Applied Energy

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



Multi-Agent attention-based deep reinforcement learning for demand response in grid-responsive buildings☆

Jiahan Xie [a], Akshay Ajagekar [b], Fengqi You [b,c,*]

[a] Department of Computer Science, Cornell University, Ithaca, NY 14853, USA
[b] Systems Engineering, Cornell University, Ithaca, NY 14853, USA
[c] Robert Frederick Smith School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, NY 14853, USA

HIGHLIGHTS

- Novel multi-agent AI-based technique for demand response in buildings is proposed.
- A deep reinforcement learning strategy is developed for grid-responsive buildings.
- Attention mechanism is leveraged for information sharing and building coordination.
- Proposed demand response strategy exhibits efficient electrical demand reduction.
- Adaptive learning capabilities are also demonstrated with presented case studies.

ARTICLE INFO

Keywords:
Demand response
Deep reinforcement learning
Multi-agent
Buildings

ABSTRACT

Integrating renewable energy resources and deploying energy management devices offer great opportunities to develop autonomous energy management systems in grid-responsive buildings. Demand response can promote enhancing demand flexibility and energy efficiency while reducing consumer costs. In this work, we propose a novel multi-agent deep reinforcement learning (MADRL) based approach with an agent assigned to individual buildings to facilitate demand response programs with diverse loads, including space heating/cooling and electrical equipment. Achieving real-time autonomous demand response in networks of buildings is challenging due to uncertain system parameters, the dynamic market price, and complex coupled operational constraints. To develop a scalable approach for automated demand response in networks of interconnected buildings, coordination between buildings is necessary to ensure demand flexibility and the grid's stability. We propose a MADRL technique that utilizes an actor-critic algorithm incorporating shared attention mechanism to enable effective and scalable real-time coordinated demand response in grid-responsive buildings. The presented case studies demonstrate the ability of the proposed approach to obtain decentralized cooperative policies for electricity costs minimization and efficient load shaping without knowledge of building energy systems. The viability of the proposed control approach is also demonstrated by a reduction of over 6% net load demand compared to standard reinforcement learning approaches, deep deterministic policy gradient, and soft actor-critic algorithm, as well as a tailored MADRL approach for demand response.

1. Introduction

In the United States, buildings account for a significant portion of electricity demand (74%) and primary energy consumption (40%), as well as associated greenhouse gas emissions [1]. In order to achieve demand flexibility, energy efficiency, and resiliency, the U.S. Department of Energy developed the Grid-interactive Efficient Buildings (GEB) program [2], which intends to incorporate technologies for grid-responsive buildings. One way to approach the goal and lessen buildings' reliance on the electrical grid is to use on-site renewable energy sources (RESs) and storage devices. However, to deal with the uncertainties associated with RESs to prevent instability and ensure

☆ The short version of the paper was presented at CUE2022. This paper is a substantial extension of the short version of the conference paper.

* Corresponding author at: Systems Engineering, Cornell University, Ithaca, NY 14853, USA.
E-mail address: fengqi.you@cornell.edu (F. You).

https://doi.org/10.1016/j.apenergy.2023.121162
Received 5 February 2023; Received in revised form 24 March 2023; Accepted 15 April 2023
Available online 26 April 2023

resource availability, their integration into the current grid infrastructure must be done with care. Building control is further complicated due to the need for load shifting as an adaptive response to grid signals [3]. With the development of smart energy management, demand response has been acknowledged as a viable approach to encourage less electricity usage when wholesale market prices are high or when the grid's reliability is jeopardized [4]. Therefore, participating in demand response programs provides solutions for networks of buildings to enhance load controllability and promote economic efficiency while alleviating the stochastic power inputs from RESs [5]. Within a network of grid-responsive buildings, there is an additional obstacle caused by the unknown, intricate energy management strategies used by other buildings. To overcome the challenges mentioned above, there is an urgent need to coordinate multi-building systems to participate in demand response programs with complicated system structures and dynamic electricity prices.

Several techniques have been proposed to optimize the energy consumption of grid-responsive buildings in order to match the power grid's demand, with model-based approaches being the most extensively studied among them. For instance, model predictive control (MPC) has significantly contributed to energy management and demand response programs [6]. MPC is an intricate approach and has received substantial research attention owing to its ability to optimize specific control objectives while adhering to environmental constraints. MPC-based techniques have been demonstrated to optimally control cold storage to maintain indoor thermal comfort while participating in demand response [7]. High penetration of RESs in power networks that are subject to demand response constraints can also be addressed with MPC approaches [8]. Another model-based approach involves modeling the demand response problem as a scheduling problem formulated as mixed-integer linear programming (MILP) problem, which requires knowledge of system dynamics for various appliances utilized in energy management systems [9]. Tailored optimization strategies for demand response modeled as MILP problems have also been investigated to improve the performance of solution algorithms [10]. However, such model-based techniques require accurate modeling of the system dynamics governed by complex physical processes. System parameters necessary for accurate modeling are often hard to obtain in practice due to concerns like privacy and the degradation of energy systems over time [11]. In addition to the reliance of model-based approaches for demand response on extensive information like an accurate forecast of uncertain parameters in the system [12], they are also prone to modeling errors and may not necessarily be adaptive [13]. Due to challenges associated with time-varying system parameters and constructing energy models unique to each building [14], adopting model-based approaches for demand response may not be viable as the problem size increases.

Compared to the model-based techniques, reinforcement learning (RL) has demonstrated significant promise for demand response modeled as a sequential decision-making problem [15]. RL has the advantage of requiring no prior knowledge of the system dynamics and can be adopted in a model-free manner allowing for easier implementation in a practical setting compared to conventional optimization approaches. Moreover, deep reinforcement learning (DRL), which combines the function approximation abilities of deep learning with RL, has been successfully employed in developing several demand response programs [16–19]. DRL can help realize multiple control objectives, which can be exploited to perform joint operations like maintaining thermal comfort in buildings while reducing power consumption with demand response [20]. Simultaneous consideration of objectives like revenue management, user satisfaction, and peak load reduction can also be performed with DRL-based techniques [21]. In a residential setting wherein varying energy resources and operating models are employed by each building, autonomous demand response can be performed with DRL-based load management without the need for underlying knowledge [22]. Uncertainties associated with various factors like load demand, renewable generation, weather, and dynamic electricity pricing can also be addressed with DRL-based demand response algorithms [22,23]. In the cooperative setting of grid-responsive buildings, conventional single-agent DRL techniques may exhibit limitations with respect to high dimensionality and choosing appropriate reward functions. To avoid disparate and inconsistent individual objectives, multi-agent deep reinforcement learning (MADRL) provides a flexible and robust approach for incorporating multiple energy systems [24]. A multi-agent setting can also help speed up learning and improve the performance of individual DRL agents by information sharing and concurrent learning [25].

MADRL is increasingly being used in energy management systems to deal with challenging cooperative learning tasks [26–28]. Autonomous demand response for systems with multiple energy components has also been explored with multi-agent variants of state-of-the-art DRL algorithms [24,29–31]. There are several research challenges associated with developing a MADRL-based technique for demand response in grid-responsive buildings. The first challenge lies in deriving efficient energy management policies that promote cooperation to reduce the overall load demand. Load shaping in grid-responsive buildings has seen significant performance improvement with tailored MADRL approaches in comparison with their single-agent counterparts [32]. However, tailored techniques may inhibit applications to general real-world buildings, while the adoption of conventional DRL algorithms in a multi-agent setting may lead to high peak loads caused by uncoordinated policies. Moreover, a shift in peak loads may also occur instead of flattening as the dynamic electricity pricing is dependent on the incurred loads [15]. Another difficulty resides in addressing the non-stationarity problem associated with the multi-agent setting caused by information sharing among agents and their interaction with the power grid. It is important to ensure that the stability of individual agents is not affected by concurrent learning and dynamic environmental factors. Ensuring the scalability of the MADRL approach with large amounts of information sharing among agents caused by an increasing number of buildings is crucial. Finally, managing the computational resource utilization during the learning and deployment of the multi-agent approach is an additional research challenge.

In this work, we propose a novel MADRL-based control framework that utilizes centralized training to enhance cooperative policies for economical demand response programs and load shaping in grid-responsive buildings. A compact representation of observed information is utilized by individual building agents to prevent scalability issues and reduce the amount of computation. We employ the attention mechanism in the multi-agent setting to improve coordination and facilitate stationarity among the agents. Centralized training of the critic, which estimates cumulative discounted rewards by assigning relevance among the shared information along with each agent utilizing local information, has been demonstrated to achieve efficient performance in cooperative settings [33,34]. In order to adjust to uncertain disturbances by learning from historical trajectories, we use a centralized learning process that adapts dynamically to time-varying factors and interactively learns to select which agents to pay attention to at each time step. The applicability and efficiency of the proposed attention-based MADRL approach are demonstrated through two case studies on both simulated and real-world grid-responsive buildings. The simulated case study comprises of prototype buildings situated in a hot and humid environment in New Orleans and is conducted to validate the learning capabilities of the proposed controller. An extensive analysis of the demand response techniques applied to real-world buildings subject to high penetration of RESs located at Cornell University's Ithaca campus is also performed to substantiate their performance efficacy in terms of load reduction.

The main contributions of this work are as follows:

- A novel multi-agent approach for demand response in grid-responsive buildings that employ an actor-critic based deep

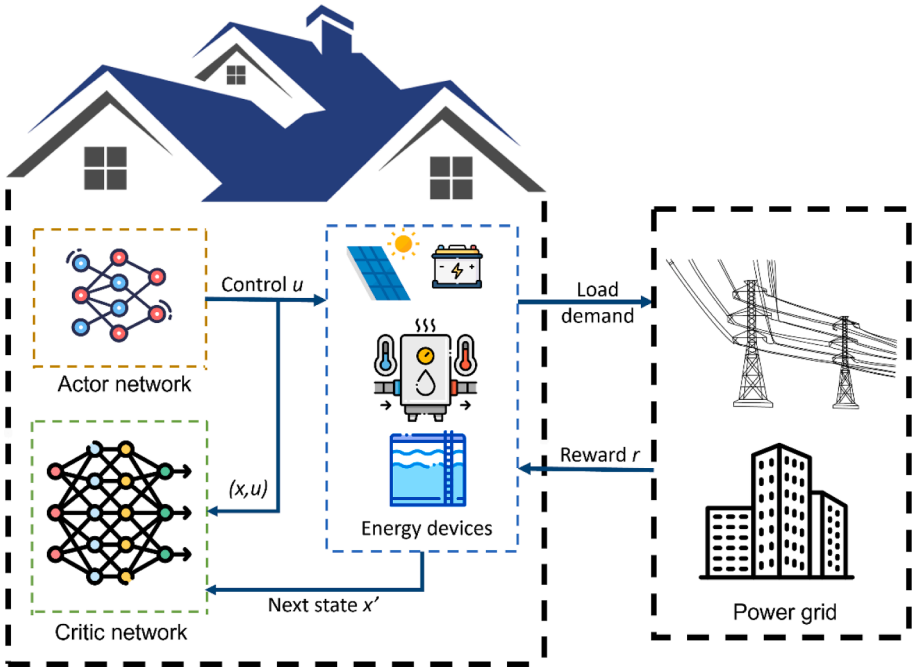


Fig. 1. An overview of the individual agent or building interacting with the power grid considered for the proposed MADRL control framework for demand response in grid-responsive buildings. An agent capable of making real-time demand response decisions is assigned to each building and includes an actor for producing decisions and a critic for estimating the cumulative value of the building state and applied controls at each timestep.

reinforcement learning technique to produce efficient energy management policies resulting in reduced overall electrical demand.

- Computationally efficient multi-agent architecture and learning strategy which utilizes a novel integration of extracting compact representation from local observations with the attention mechanism to promote coordination among buildings.
- Extensive computational experiments with the proposed multi-agent deep reinforcement learning strategy applied to buildings in warmer climates that utilize pre-simulated data obtained with EnergyPlus in addition to real-world buildings that use historically recorded data under harsh climate conditions. A comparative study with state-of-the-art baseline methods for demand response in grid-responsive buildings is also presented.

The remainder of this paper is structured as follows. Problem formulation for autonomous demand response as a sequential decision-making problem and the proposed attention-based MADRL control technique are described in Section 2. Two case studies with various demand response techniques applied to grid-responsive buildings are presented in Section 3. Finally, conclusions are drawn in Section 4. A brief background on DRL and MADRL is also presented in the Appendix, along with additional information on the case studies presented in this work.

2. Multi-agent attention-based method for demand response

In our networks of grid-responsive buildings, demand response programs incorporate incentive payments to encourage using less electricity during high wholesale market prices or when the grid's reliability is jeopardized [4]. To satisfy the residents' thermal comfort and diverse energy loads, energy devices are stored and released by the building agents preemptively. Under this setting, we propose a novel MADRL control technique that leverages the attention mechanism to enhance coordination between buildings for demand response. This MADRL control method is hereafter referred to as the multi-agent attention-based controller (MAAC). In this framework, we assign an agent to individual buildings and embed an actor and a critic within each agent for all grid-responsive buildings. To encourage cooperation between agents, we adopt the centralized training framework to help each building capture relevant information from the grid. During the execution phase, each agent's actor will use the trained deep neural networks (DNNs) to select actions based on only local information to decrease the computation time and achieve real-time, autonomous demand response. Utilizing online learning algorithms, the agents learn the optimal policies in a short period via interactions with the environment. Fig. 1 shows an overview of the individual agents or buildings interacting with the grid. The design and implementation details of the MAAC controller are described in the following subsections.

2.1. Multi-agent Markov decision process formulation

We formulate the demand response problem in grid-responsive buildings as a finite multi-agent Markov decision process (MMDP) with discrete time steps. The grid-responsive building operation cast as MMDP is constructed with CityLearn [35] which facilitates the evaluation of controllers to reshape aggregated electricity demand by controlling the storage of energy in each building. The electric grid supplies power following a dynamic pricing structure to N buildings. Each building has a pre-simulated cooling load, heating load, and non-shiftable load consumed by appliances. Moreover, every building is equipped with photovoltaic (PV) arrays, the energy supply subsystem, including heat pumps and electric heaters, and the energy storage subsystem, which consists of batteries, chilled water tanks, and domestic heat water (DHW) tanks. Each building receives thermal energy supply from storage tanks and energy supply units, such as heat pumps and electric heaters, at every time step. In CityLearn, the energy supply devices are sized to meet the energy demand of the building at any given time throughout the simulation to satisfy the presumption that the building temperature setpoints are always satisfied and enable pre-simulated energy loads of the buildings [35]. Individual agents corresponding to the grid-responsive buildings assume no knowledge of the energy model attributes or the system transition dynamics. Although historical data is used to simulate the operation of the grid-responsive buildings, we ensure no data leakage to maintain the stochasticity of

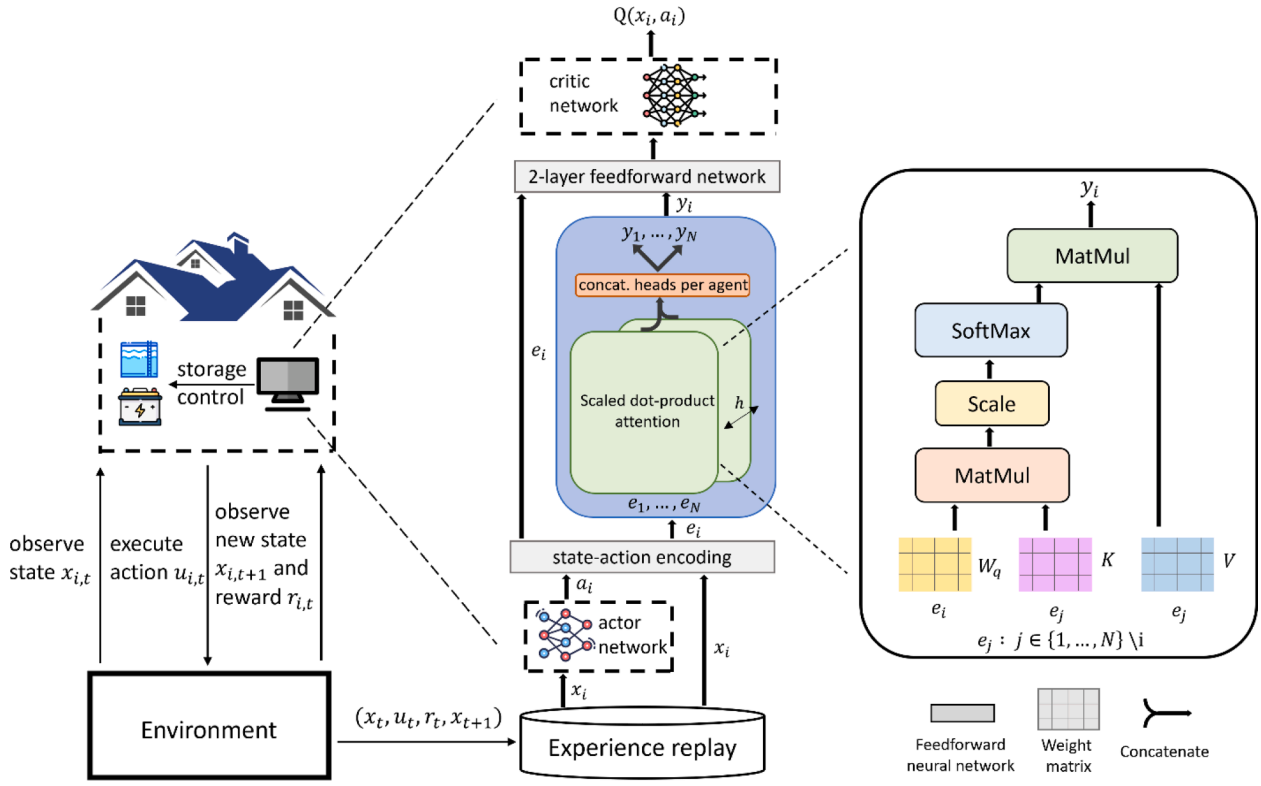


Fig. 2. Systematic workflow for computing the state-action value for each agent in the MAAC-based control algorithm.

the system dynamics. At any time step, the agents observe the current environment state and execute joint actions accordingly. Individual actions obtained with each agent are incorporated by the corresponding building, followed by obtaining reward signals in response to the agent's actions. The system then transitions to the next state, with rewards returned to the agents that learn by updating value functions and policies to maximize the discounted cumulative reward. To ensure that our simulation environment accurately captures the complex interplay between buildings and the electric grid, we model the power and information flows between grid-responsive buildings in an interactive way using CityLearn. By integrating real-time prices and energy system information from other buildings, each building in the simulation can make informed decisions that help manage demand across the district. This coordinated effort leads to a more stable grid and enables the seamless integration of renewable energy sources and efficient demand response strategies. By capturing these grid-interactive features, our simulation environment ensures a thorough and accurate representation of the complex interplay between buildings and the grid.

The observable state x_t^i of building i at time step t consists of its local state, the shared global state, and the relevant state information from other agents. All buildings share outdoor weather data as well as their future forecasts at various intervals. Specifically, outdoor temperature, outdoor relative humidity, and direct and diffuse solar radiation are considered. Moreover, we assume carbon intensity emitted during electricity production at the current time step as a global state variable. Each building also receives dynamic electricity prices C_t from the electric grid. In terms of local state space, we record information on the thermal data, which comprises of indoor temperature and its setpoint and indoor relative humidity for each building. The electricity consumption levels, including cooling and heating consumption and the non-shiftable electricity consumption by appliances, are also observed at each time step. Power generated by the PV panels also constitutes the local building state. Lastly, the local state space also includes the state of charge of each building's storage device. Within each building, storing domestic hot water and chilled water for sensible heating and cooling are provisioned. Storage of energy sourced from the electric grid and renewables is also facilitated in the buildings with electrical batteries. Each agent can store or utilize energy associated with these storage devices to address the energy demand. The i'th agent's control action then comprises $u_t^i \equiv \left(u_{t,cooling}^i, u_{t,heating}^i, u_{t,battery}^i\right)$ at time step t. We further constrain the energy transfer levels in $[-1, 1]$ using the known capacity levels of all storage devices.

The reward function is defined as $r_t^i = -sign(e_{i,t}) \cdot C_t \cdot e_{i,t}^2 \cdot \min(0, E_t)$, where $e_{i,t}$ and C_t denote the net electricity demand of building i and cost of electricity, respectively. This reward function not only considers electricity costs incurred by the building but also promotes all the grid-responsive buildings to minimize the overall load $E_t = \sum_{i=1}^{N} e_{i,t}$. Utilizing both local and global components in the reward function incentivizes the building to minimize individual energy costs while coordinating with each other to achieve a global optimum. Squaring each individual $e_{i,t}$ forces the building to shape the load curve and prevent drastic change due to the dynamic pricing. It is worth noting that the reward is positive if the building is self-sufficient. Furthermore, the reward is negative if the building draws power from the district while the district is also consuming electricity from the main grid.

2.2. Attention-based value function estimation

To encourage information sharing and demand response cooperation, it can be helpful to concatenate all information into a single vector and pass it to the fully observable centralized critic. However, it only works well when the number of agents in the communication network is small [36]. When more agents are included, the vector can become large, and the communication overhead is long. Moreover, since not all information is relevant for every agent at each time step, computational resources can be wasted to perform irrelevant calculations. With the help of the attention mechanism, the centralized critic can dynamically select which relevant agents to pay attention to at each time step instead of considering all of them at all time steps. When multiple buildings

Algorithm 1 Learning algorithm for the MAAC controller

Input: Weather data, electricity price, cooling/heating load, non-shiftable load, attributes of energy device, attributes of PV panels
Initialize the environment with N agents
Initialize policy networks π_{θ_i}, critic networks $Q^i_{\phi_k}$ for $k = 1, 2$, and the empty replay buffer $\mathcal{D}$
Set target parameters equal to local parameters $\bar{\theta}_i \leftarrow \theta_i, \bar{\psi}_1 \leftarrow \psi_1, \bar{\psi}_2 \leftarrow \psi_2$
for t in episode length **do**
  Observe state x^i_t for each agent i
  Each agent i executes action $u^i_t \sim \pi_{\theta_i}(x^i_t)$
  Observe new state x^i_{t+1}, reward r^i_t
  Store $(x^i_t, u^i_t, r^i_t, x^i_{t+1})$ in replay buffer $\mathcal{D}$
  if $\mathcal{D}_{size} \geq B_{size}$ and $t \geq T_{update}$ **then**
    for j in N_{update} **do**
      Randomly sample a batch of transitions, $B = (x, u, r, x')$ from $\mathcal{D}$
      Compute targets for all the critic networks:

$$y_i = r_i + \gamma \mathbb{E}_{u' \sim \bar{\pi}_{\theta_i}(x')}[\bar{Q}^i_{\phi_k}(x', u')]$$

      Update all critic networks by one step of gradient descent on the MSBE loss function:

$$L_Q = \sum_{i=1}^{N} \mathbb{E}_{(x,u,r,x') \sim \mathcal{D}}[(Q^i_{\phi_k}(x, u) - y_i)^2] \qquad \text{where } k = 1, 2$$

      Update policy networks by one step gradient ascent on:

$$J(\theta_i) = \mathbb{E}_{x \sim \mathcal{D}}[min_{k=1,2}\{Q^i_{\phi_k}(x_i, \pi_{\theta_i}(x_i))\}]$$

      Update target networks with:

$$\bar{\mu} \leftarrow \rho\mu + (1 - \rho)\bar{\mu}$$

    end for
  end if
end for

Fig. 3. Learning algorithm utilized for the online training phase of the proposed MAAC-based control strategies for demand response in networks of interconnected buildings.

participate in demand response, the demand response environment is nonstationary on the demand side since all buildings are simultaneously influencing the energy demand with individual actions [15]. Additionally, rebound effects may occur when all buildings are trying to delay the energy demand during peak hours. Specifically, the uncoordinated uniform delay will shift the peak demand to the period when the electricity price is lower in expectation, leading to even higher peaks and thus causing vulnerability issues [37]. The improvement in information sharing and coordination by incorporating the attention mechanism can prevent such effects and ensure grid stability.

We parameterize the policy and state-action value functions for each agent with neural networks. A deterministic policy is approximated with an actor network denoted by π_{θ_i} with the parameters θ_i and computation of controls at any time step $u^i = \pi_{\theta_i}(x^i)$ with the local observable state x^i_t. Similarly, the critic network Q_{ϕ_i} is parameterized by ϕ_i. The attention mechanism is utilized for each agent i to query other agents for information about their observations and actions and include that information in calculating its state-action value. Therefore, the contributions from other agents are calculated dynamically at each time step, enabling the critic of each agent to learn the most important information by selectively paying attention to other agents. A detailed workflow for computing the state-action values for each agent is shown in Fig. 2. Concretely, the Q-value for agent i is computed as follows:

$$Q^i_\phi(x, u) = f_i(e_i, y_i) \tag{1}$$

where f_i is a two-layer feedforward neural network while e_i denotes the embedding computed as $e_i = g_i(x_i, u_i)$ by encoding the local state-action pair with a single-layered feedforward neural network g_i. On the other hand, y_i denotes the total contribution from other agents defined in Eq. (2) with $v_j = \text{LeakyReLU}(W_v e_j)$ and leaky rectified linear unit as the nonlinear activation function.

$$y_i = \sum_{j, j \neq i} \alpha_{ij} v_j \tag{2}$$

The attention weight α_j compares the embeddings using the query-key system [38] and applying SoftMax to compute the similarity value between the embedding pairs,

$$\alpha_{ij} = \frac{\exp\left((W_k e_j)^T (W_q e_i)\right)}{\sum_{j=1}^{N} \exp\left((W_k e_j)^T (W_q e_i)\right)} \tag{3}$$

where W_q transforms e_i into a "query" and W_k transforms e_j into a "key." It should be noted that the state-action value for each critic is a function of agent i's observations and actions as well as contributions from other agents. To this end, the parameters associated with the query, key, and values, as W_q, W_k, and W_v, respectively, are shared among all N agents.

2.3. Implementation of the MAAC algorithm for demand response

The learning algorithm employed for demand response in grid-responsive buildings utilizes the architecture for actor and critic networks described above and is presented in Fig. 3. To reduce the variance of value estimates [39], we use two critic networks for each building denoted by $Q^i_{\phi_k}$, where $k = 1, 2$. This prevents overestimation of the state-action value estimates and inhibits suboptimal policies. Moreover, copies of all policy and critic networks are also utilized as target networks to stabilize the training process. For each agent i, the target actor

and critic networks are represented with $\overline{Q}^{i}_{\phi_k}$ and $\overline{\pi}_{\theta_i}$, respectively, for which the parameters are initialized to be same as that of the local networks. Over the duration of the demand response simulation, the transition tuples $(x^i_t, u^i_t, r^i_t, x^i_{t+1}),\ \forall i \in [1, N]$ are recorded and stored in the experience replay buffer $\mathscr{D}$. An experience replay is used to facilitate the temporal decorrelation of experience samples that are produced successively [40]. During each update step, a batch of experiences of size B_{size} is randomly sampled from the replay buffer. The sampled batch of transitions are then used to update the policy and critic networks. Due to the sharing of the set of parameters associated with query, key, and value, all critics are updated to minimize a regression loss function with combined information as shown in Eq (4) with the targets y_i defined in Eq. (5). Likewise, the individual policies are updated to maximize the parametric reward function for each agent's policy in Eq. (6) by performing a gradient ascent step.

$$L_Q = \sum_{i=1}^{N} \mathbb{E}_{(\mathrm{x,u,r,x'})\sim \mathrm{D}} \left[\left(\mathrm{Q}^{\mathrm{i}}_{\phi_k}(\mathrm{x,u}) - \mathrm{y_i} \right)^2 \right] \tag{4}$$

$$y_i = r_i + \gamma \mathbb{E}_{\mathrm{u'}\sim\overline{\pi}_{\theta_i}(\mathrm{x'})} \left[\overline{\mathrm{Q}}^{\mathrm{i}}_{\phi_k}(\mathrm{x'}, \mathrm{u'}) \right] \tag{5}$$

$$J(\theta_i) = \mathbb{E}_{\mathrm{x}\sim\mathrm{D}} \left[\min_{\mathrm{k}=1,2} \left\{ \mathrm{Q}^{\mathrm{i}}_{\phi_k}(\mathrm{x_i}, \pi_{\theta_i}(\mathrm{x_i})) \right\} \right] \tag{6}$$

The critic networks are updated by minimizing the mean squared Bellman error, and it is noteworthy that the local critic networks take in (x_i, u_i), whereas the next-state action being passed to the target is computed by the target policy instead of the local policy. It should also be noted that both the critic networks $Q^i_{\phi_k}, k = 1, 2$ are updated with separate targets and corresponding loss functions as shown in Eq. (4) and (5). Lastly, the target networks' parameters $\mu \equiv (\theta, \phi)$ are updated by Polyak averaging [41],

$$\overline{\mu} = \rho\mu + (1-\rho)\overline{\mu} \tag{7}$$

where μ is the network weights of policy or critic networks, $\overline{\mu}$ refers to the parameters of target policy or critic networks, and ρ is a hyperparameter between 0 and 1.

3. Performance evaluation of demand response operations in grid-responsive buildings

We demonstrate the viability of the proposed attention-based multi-agent DRL approach for demand response in grid-responsive buildings by performing computational experiments with two case studies. The presented case studies vary in terms of the location of buildings, outdoor weather conditions, indoor climate setpoints, as well as the dynamic pricing offered by the power grid. It should also be noted that the selected grid-responsive buildings also exhibit varying local attributes, like the electric batteries in different buildings that have different capacity levels. We perform several computational experiments for the case studies to demonstrate the generalization and scalability capabilities of the proposed MAAC control framework for demand response. Modeling energy systems in buildings like batteries for applying model-based approaches like MPC for demand response requires acquiring the technical parameters associated with system components, which is typically impractical because of privacy concerns and system aging [42]. Due to the intimate interaction between such energy systems and human behavior, it is not always viable to model the energy systems owing to the variability of changing consumption patterns [32]. Even with all the system information acquired, it is a challenging and labor-intensive process to build energy models for all the buildings since each building is unique, and the model designed for one building would typically not fit another directly [14]. Considering the potential equipment faults, the seasonal and annual variation of RESs, and the changing consumption patterns in a four-year period, we compare our controllers with other DRL-based techniques, which are also model-free and do not require knowledge of the system. The performance of the MAAC-based control algorithm is then compared against the rule-based controller (RBC) and three state-of-the-art DRL techniques, namely, deep deterministic policy gradient (DDPG) [43] and soft actor-critic (SAC) [44] to demonstrate its efficacy. Additionally, a tailored multi-agent DRL approach for demand response in grid-responsive buildings termed MARLISA [32] is also implemented for a fair comparative analysis. Due to their versatility across applications, the DDPG and the SAC controllers are implemented as decentralized controllers without explicit coordination algorithms.

The DDPG and the SAC controllers are implemented as decentralized controllers without explicit coordination algorithms. In the MARLISA setting, each agent shares cumulative predicted electricity consumption of itself and the agents that selected actions before it in a leader–follower schema [32]. In terms of experimental settings, we adopt consistent configuration across all baseline methods and the proposed MAAC technique for demand response in grid-responsive buildings. Additionally, all agents within the MAAC controller utilize consistent architectures and hyperparameter settings with randomly initialized neural networks. Although RBC can be utilized during the exploration phase of the DRL-based controllers, it may affect the overall long-term performance of the controllers [32]. So, all baselines and the proposed MAAC controller perform random exploration during the initial exploration phase. All the actor networks have (400, 300) as hidden dimensions with ReLU activation between them. The critic networks, apart from MAAC, also have hidden dimensions (400, 300) with Leaky ReLU activation. In contrast, the MAAC critic has a four-head, 256 hidden units network with Leaky ReLU activation. Moreover, all the DRL-based techniques utilize a discount factor $\gamma = 0.99$ and the Polyak averaging coefficient $\rho = 0.005$. For the experience replay incorporated for learning with the DRL techniques, we construct a replay buffer $\mathscr{D}$ of size 10^5. During each learning step, a transitions batch of size $B_{size} = 200$ is sampled uniformly from the replay buffer. In addition, all optimization steps required to update the policy and critic networks are performed using the Adam optimizer.

3.1. Simulated building operation with renewable energy systems in the warm climate of New Orleans

The proposed MAAC control technique is evaluated in a hot and humid environment in New Orleans. The grid-interactive district has nine connected buildings from DOE prototyping with pre-simulated energy demand using EnergyPlus [45]. The dataset used for the evaluation of this case study is part of the CityLearn challenge 2021 [46], wherein the grid-responsive buildings comprise of an office building, a fast-food restaurant, a retail building, a strip mall, and five residential buildings [32]. The corresponding chilled water tanks storage capacities, DHW tanks storage capacities, and the PV panels capacities are provided in Table A1 of the Appendix. We conduct simulations using same weather data in the grid-interactive environment for ten epochs or episodes to test the performance and the convergence of different DRL controllers. Each episode spans 365 days, and all the DRL controllers perform random action exploration for the first 250 days during the first episode. After 250 days, the SAC, MARLISA, and MAAC controllers inhibit stochasticity and exploit their policies by obtaining controls from the trained actor networks over the subsequent episodes. The training process is terminated after convergence of the corresponding learning curves with less than 1% improvement of the associated metric over subsequent episodes as the termination criteria. The DDPG controllers' action noises scale to zero for taking deterministic actions. The manually optimized RBCs are tuned to perform deterministic storage actions by storing more energy during the nighttime and releasing it during the daytime.

We first evaluate the MAAC controller's learning capabilities of minimizing the district's net electricity demand over the training

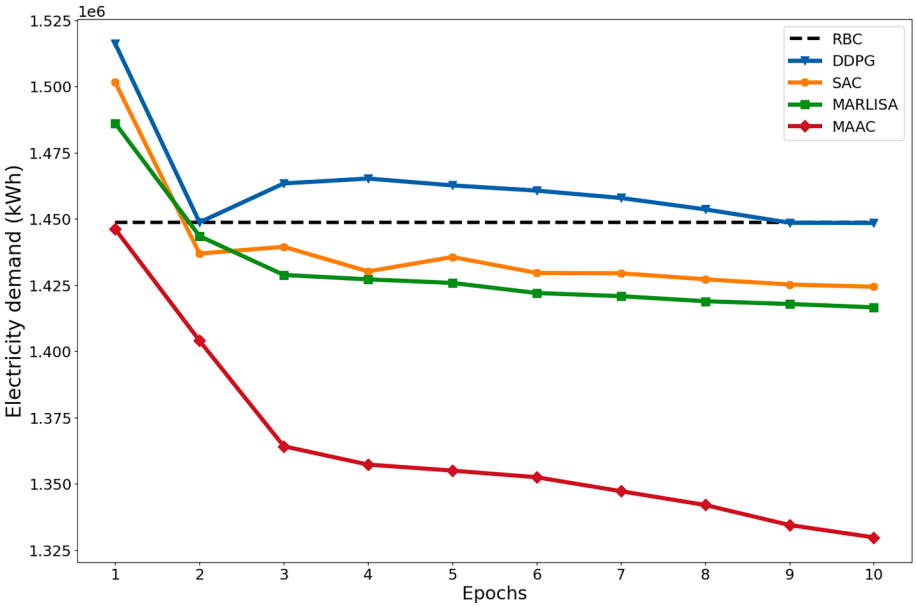


Fig. 4. Electricity demand reduction comparisons between the proposed method and other RL-based methods, using the RBC as the baseline.

Table 1
Various performance metrics for demand response performance evaluation obtained with different DRL-based controllers during last year for the New Orleans case study.

Metrics	DRL-based Controllers			
	DDPG[1]	*SAC*[2]	*MARLISA*[3]	*MAAC*[4]
Annual cumulative electricity costs ($/kWh · kWh)	1,568,617	1,565,856	1,559,726	1,558,475
Ramping	492,803	265,898	266,561	268,261
1 – load factor	0.590	0.516	0.523	0.522
Peak Demand (kW)	541.8	484.6	499.2	482.3
Average Daily Peak (kW)	315.3	267.6	268.3	267.5

[1] Deep deterministic policy gradient (DDPG) is an off-policy policy gradient algorithm.
[2] Soft actor-critic (SAC) is an off-policy actor-critic DRL algorithm.
[3] MARLISA is a tailored multi-agent DRL approach for demand response in buildings.
[4] Multi-agent attention-based controller (MAAC) is the proposed MADRL algorithm.

process. The overall net electricity demand within the district observed after the termination of each episode is plotted in Fig. 4. Conventional DRL approaches like DDPG and SAC are unable to demonstrate any significant performance improvement over the RBC baseline as learning progresses. Moreover, the SAC, MARLISA, and MAAC controllers consume less electricity than the manually tuned RBC controllers, while the DDPG controllers do not exhibit load reduction ability. MARLISA, which is a tailored multi-agent DRL approach for demand response, leads to only 0.17% reduction in net load as compared to the baseline. On the other hand, the proposed MAAC controller is able to lower net demand over subsequent episodes, unlike the other DRL-based demand response techniques. Additionally, the proposed MAAC strategy for demand response results in a significant reduction of net load demand with more than 8% reduction over the RBC baseline after ten episodes. The DRL strategies DDPG and SAC converge quickly to higher load demand levels in contrast to the multi-agent DRL approaches. In comparison, MARLISA and MAAC learn slowly, which allows them to achieve improved local optima. The proposed MAAC technique for demand response in grid-responsive buildings employs sharing local information under a multi-agent setting which allows it to yield a substantial advantage over MARLISA. We also evaluate the final year's electricity prices paid for the New Orleans case study are listed in Table 1. The MAAC-based controllers incur the least electricity costs. Additionally, the lowest peak demand and average daily peak throughout the year is observed with the MAAC controller, demonstrating its ability to shape the load curve and ensure grid stability. Fig. 5 visualizes the net electricity demand under different DRL-based controllers for a week in summer, during which the demands are higher and load shaping is more critical. As the simulation results have demonstrated, all other DRL-based controllers perform better load flattening than the DDPG controller. Among the controllers that use the soft-actor-critic framework, MAAC controllers exhibit superior performance compared to SAC and MARLISA during peak demand periods between Aug 20th to Aug 21st and between Aug 23rd and Aug 24th, with daily peaks significantly lower than controller-free or no storage demand. In contrast, SAC and

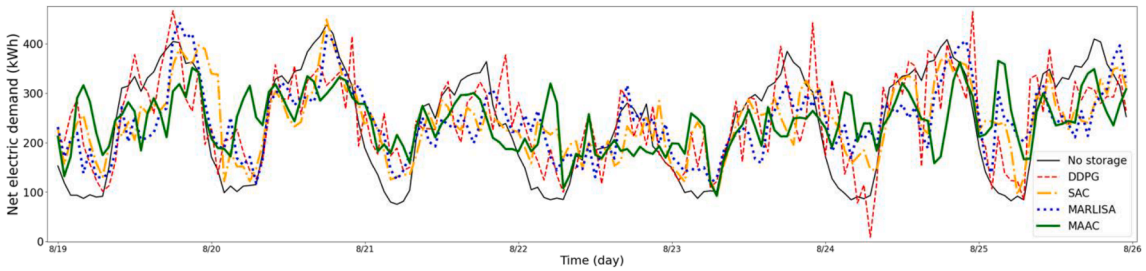


Fig. 5. Load profiles comparison between different DRL-based controllers during a high-demand week in summer. The controller-free demand profile is also shown when no energy storage devices are equipped and utilized.

Table 2
The cumulative price paid for electricity demand ($/kWh·kWh) of different RL-based controllers during the studied 48-hour period in the four seasons along with the incurred annual cumulative costs.

Month	DRL-based Controller			
	DDPG[1]	*SAC*[2]	*MARLISA*[3]	*MAAC*[4]
March	72,650	40,441	52,624	26,610
July	26,076	23,254	26,787	22,872
October	17,570	16,043	17,016	15,097
January	61,882	59,791	66,340	58,524
Annual cumulative electricity costs	2,346,235	2,197,627	2,225,274	2,155,593

[1] Deep deterministic policy gradient (DDPG) is an off-policy policy gradient algorithm.
[2] Soft actor-critic (SAC) is an off-policy actor-critic DRL algorithm.
[3] MARLISA is a tailored multi-agent DRL approach for demand response in buildings.
[4] Multi-agent attention-based controller (MAAC) is the proposed MADRL algorithm.

MARLISA show daily peaks almost as high as controller-free demand during these periods. In addition, MAAC controllers exhibit relatively higher electricity demand during off-peak periods, leading to more stable energy demand for buildings and increased grid reliability.

3.2. Load shaping of real-world buildings in Cornell University campus with high penetration of renewable energy sources

3.2.1. Experiment setup

To further test the robustness of the model, we perform another real-world case study on the Cornell University's Ithaca campus over a four-year horizon (2016–2019) with hourly resolution. In this case study, we consider a chemistry lab building, a building for both lecture halls and laboratories in the Engineering quadrangle, another academic building located in the Agricultural quadrangle, an Engineering library building, and an administration building in the Engineering quadrangle, the applied physics department building in the central part of the campus, university administration building in the central part of the campus, a library in the Agriculture Quad, building with lecture halls located in the Agriculture Quad, and one bioscience research institute adjacent to the Agriculture Quad. The energy consumption data of each building are downloaded from Cornell Energy Portal [47]. The electricity prices are obtained from New York State Electric & Gas service operator. We utilize the weather data from Visual Crossing API service [48], while the weather predictions data are downloaded from the Iowa environmental mesonet [49]. Moreover, the building's chilled water tanks storage capacities, the DHW tanks storage capacities, and the PV panels capacities are listed in Table A2.

All the DRL controllers perform random action exploration for the first 250 days in the first year. Similarly, the SAC, MARLISA, and MAAC controllers perform stochastic policies until the beginning of the fourth year by sampling from the policies distribution with entropy regularization. For the DDPG controllers, they have uncorrelated, mean-zero Gaussian noise for exploration until the start of the fourth year. Starting from the fourth year, the SAC, MARLISA, and MAAC controllers exploit their policies by taking mean action. Furthermore, the DDPG controllers' actions noises are set to 0 to execute deterministic actions. The RBCs are tuned to perform uniform actions to charge 4.2% of their maximum capacities every hour from 12 PM to 8 PM and release 3% of their maximum capacities every hour from 10 AM to 9 PM.

3.2.2. Computational results

We compare the results of different DRL controllers' performances during the evaluation phase. In the fourth year, all DRL-based controllers except the DDPG-based controllers outperform the RBCs. Numerically, the MAAC-based controllers outperform the second-best controllers by reducing over US $40,000 more electricity costs, and they significantly exceed other benchmark controllers in terms of minimizing the total electricity costs. The MAAC-based controllers reduce US $92,158 more electricity costs compared to the RBCs. However, the SAC-based controllers can only reduce US $50,124 more electricity costs compared to the RBCs, and the MARLISA-based controllers can only curtail US $22,477 more electricity costs in comparison with the RBCs. To analyze the demand response behaviors of each DRL-based controller, we investigate four 48-hour periods in spring, summer, fall, and winter, respectively. The numerical results of the cumulative price paid for electricity demand are presented in Table 2. The MAAC-based controllers demonstrate superior performance, especially in winter and spring, when the electricity demand is higher on the Cornell University's Ithaca campus. Notably, the MAAC controllers reduced 34.2% of electricity cost compared to the second-best controllers in the studied interval in spring. To have more intuitive visual comparisons, we plotted the electricity demand of each RL controller and the real-time electricity price in the four examined periods. As illustrated in Fig. 6, the MAAC-based controllers have more promising control strategies in every studied interval. In the studied spring period, the electricity price is at a peak between hours 31 and 45. The SAC, MARLISA, and MAAC controllers all start decreasing their demands in response to the increasing electricity price, whereas the DDPG controllers are not manifesting such an effect significantly. Comparing the SAC, MARLISA, and MAAC controllers, we can see that the proposed method reduces the most electricity demand while still having a steady demand load curve, which helps guarantee the grid's stability. The immediate demand reduction from hour 30 to 31 happen simultaneously with the sudden increase in electricity price, without any lag time at all. This abovementioned effect shows the real-time, instant adjustment ability of the MAAC controllers concerning unforeseen external information. During the 48 h of the summertime, the electricity price is steadier, and we can see the load curve of MAAC is also steadier. Moreover, we can see that during hours 26–35, when the electricity price is higher, the MAAC controllers proactively reduce the demand to cope with the rising price. Although they start to increase the demand after hour 34, the absolute values of the demand are still low. Within the examined period in fall, there are two intervals when the price is higher, hours 16–21 and hours 39–41. The MAAC controllers again demonstrate their ability to take intelligent actions. The electricity demand starts decreasing concurrently with the increase in price, and the demand rises back when the price is lower, during which the controllers can store more energy for the preparation of the next peak price period. Moreover, this pre-storing behavior is more obvious between hours 29–38 when other controllers require less demand. Nevertheless, it is beneficial to confront the rising price from hours 39–41. Additionally, although the SAC controllers are also achieving as low demand as the MAAC controllers do between hours 39–41, the pre-stored energy by the MAAC controllers enables them to prevent the bounce-back demand as the SAC controllers are experiencing, which will undermine the reliability of the grid. Finally, in the studied period in winter, as all controllers are performing distinct control strategies, the MAAC controllers still coped with the increase in price the best, as shown at hours 15 and 36.

Next, we focus on the specific energy control behaviors of the proposed MAAC-based controllers. In Fig. 7, we study the same 48-hour period in spring and summer, when the cooling and heating demand are more balanced than the periods in fall and winter. As shown in Fig. 7, heating energy consumption and non-shiftable electricity consumption for appliances dominate the total energy usage. While the electricity consumption for appliances is fixed, we can see that the MAAC-based controllers are reacting to the dynamic environment by utilizing the DHW storage tanks. Recall from Fig. 6 that the electricity price is the highest from hours 31 to 45 during the studied period in spring, Fig. 7 exhibits that the MAAC-based controllers are using almost entirely the stored heating energy during this interval. This, combined with the MAAC controller utilizing solely electrical heating devices for the

Fig. 6. Net electricity demand curves of different RL-based controllers in response to the electricity price during 48-hour periods in seasons of (a) spring, (b) summer, (c) fall, and (d) winter.

heating supply between hours 46–48 due to low electricity prices, leads to a steep increase in electricity consumption observed during this period. More specifically, Fig. 8 further examines the proposed method's detailed heating energy supply and storage behavior in the same period. Between hours 31 and 33, the MAAC-based controllers are using 100% stored energy to satisfy the building's heating consumption, eliminating any electricity demand for heating, which would have taken up a significant amount due to the cold weather in spring Ithaca. From hours 34 to 38, 94.4% of the heating consumption is supplied by discharging the DHW storage tanks. And this is made possible by prior storage to the DHW tanks between hours 26 and 30, as illustrated in Fig. 8. Since the electricity price from hours 7 to 15 and from hours 46 to 48 is almost zero, the MAAC-based controllers take advantage of these intervals to charge the DHW storage tanks in advance. In the studied 48-hour period during summer, we can also see the superb performance of the proposed method. As reflected in Fig. 7, the studied period in summer has higher cooling consumption during the daytime while having higher heating consumption during nighttime. And the corresponding control strategies are exerted by the proposed controllers intelligently. During the daytime periods, when hours 16 to 25 and hours 38 to 46, the MAAC controllers are releasing a huge amount of cooling storage from the chilled water tanks, as depicted in Fig. 9, reducing 27.4% the cooling demand. From hours 26 to 37, when the electricity price is the highest, we can also see that the heating energy is released from the storage device accordingly. Before ending the discussion of the detailed energy supply and storage behaviors of the proposed method, we analyze the heating energy supply and storage behavior during the studied period in winter. It is of important value because the demand in winter is the highest throughout

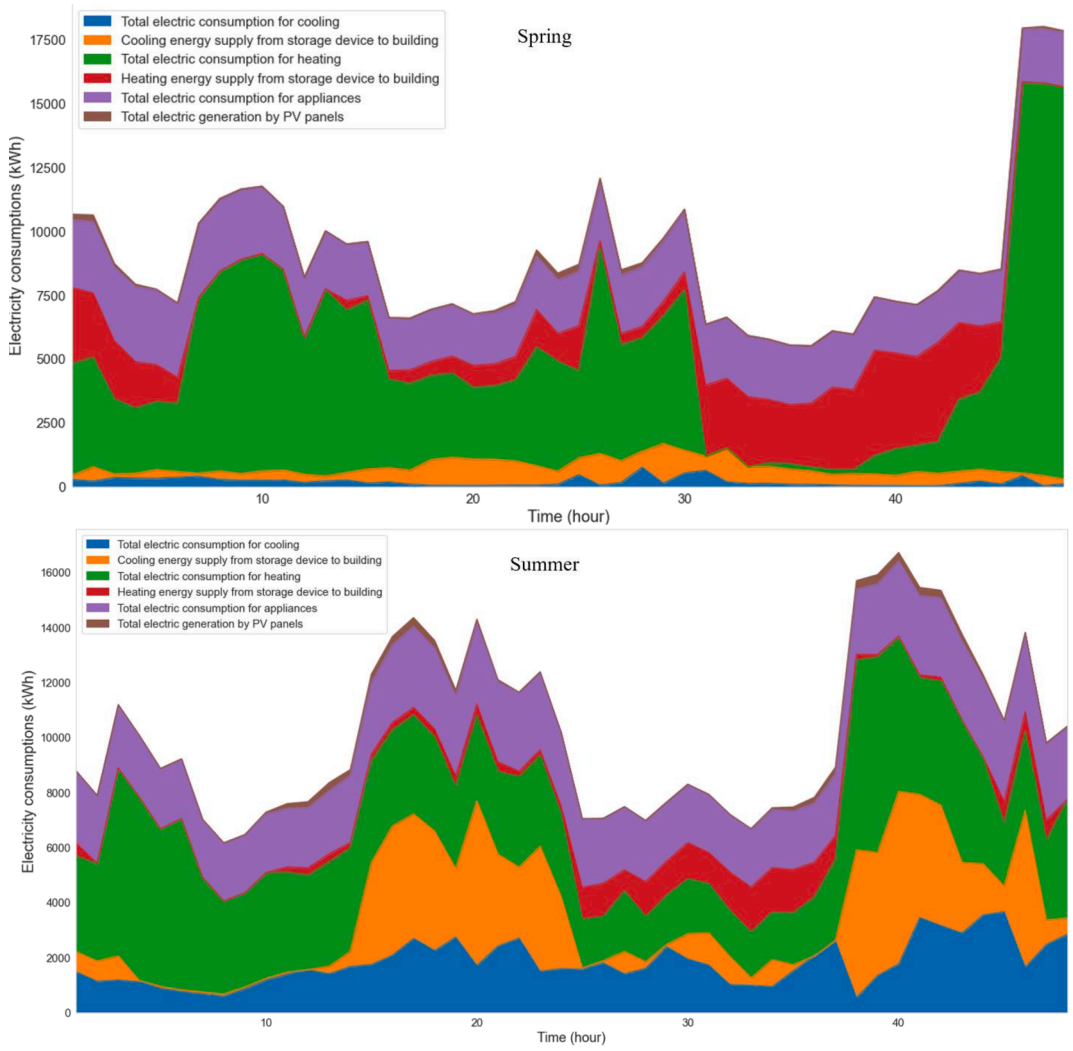


Fig. 7. Overall energy consumption, supply, and generation profile of the MAAC method during 48-hour periods in spring and summer.

the year and is caused by the high heating energy demand. From Fig. 8, we can see that the MAAC controllers are constantly supplying heating energy from the heating device to DHW storage tanks between hours 0 and 10, 24 and 34, 37 to 48. The only intervals where the storage device is only being discharged are between hours 13 to 16 and hours 36, when the price is higher.

4. Conclusion

In this work, we developed a MADRL-based control framework for efficient demand response in grid-responsive buildings, which assigns a DRL agent for energy management in each building. The proposed multi-agent demand response technique extracts compact representations of building states and controls to realize a scalable framework while leveraging the attention mechanism to promote coordination among agents. A real-time autonomous coordinated demand response in grid-responsive buildings was facilitated by utilizing a centralized learning strategy combined with decentralized execution in agents to produce controls in a computationally efficient manner. The applicability and efficiency of the proposed multi-agent attention-based controller were demonstrated with two case studies on grid-responsive buildings under different environmental conditions. Compared to both single-agent and multi-agent DRL approaches, the proposed demand response technique demonstrated better learning performance as well as the highest net load reduction. The adaptive capabilities of the attention-based demand response technique were also substantiated with a case study comprising of buildings in real-world scenarios. The obtained computational results further show the proposed method's superior capability of load shaping and a reduction in net load demand of more than 6% over both conventional and state-of-the-art RL approaches. The MAAC controllers have shown superior automatic demand response capabilities, however, the critic networks of each building's embedded agents can access information from other buildings, which could be a privacy concern. In order to address these concerns, future work should investigate data security and privacy issues associated with this method. One strategy for improving the MAAC controller is to explore federated learning techniques [50], which can securely combine data from all buildings through a central server while maintaining privacy and performance. It would also be valuable to investigate the impact of increasing numbers of agents on the learning capacity of the centralized critic network. Ultimately, ensuring that the MAAC approach is effective and ethically responsible is crucial to address privacy concerns and scalability issues.

Author contribution

JX, AA, and FY developed the models, conducted the simulations and analyzed the results.

JX, AA, and FY wrote the manuscript.

All authors reviewed the final manuscript.

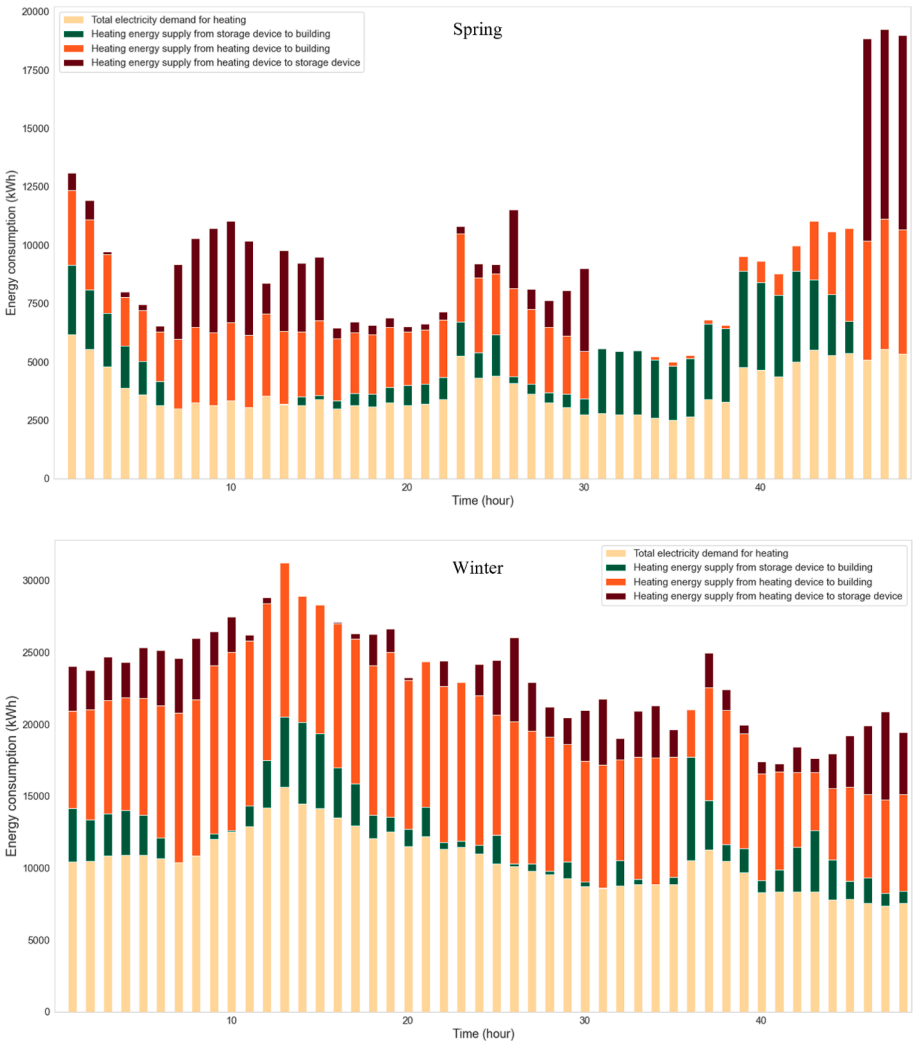


Fig. 8. Heating energy demand and supply distributions of the MAAC method during 48-hour periods in spring and winter.

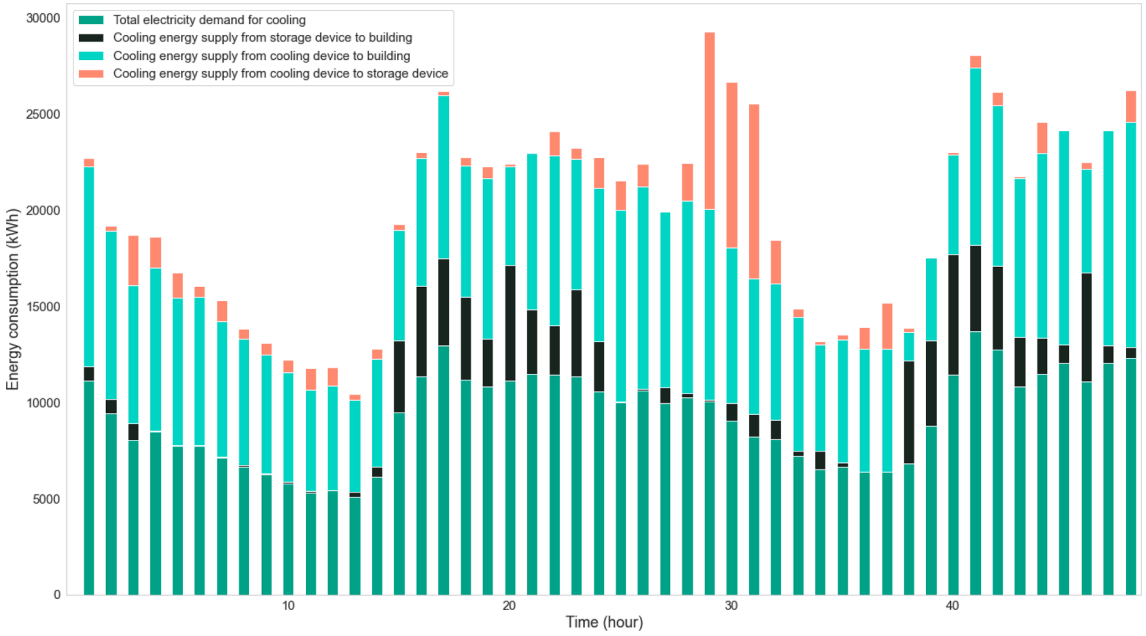


Fig. 9. Cooling energy demand and supply distributions of the MAAC method during a 48-hours period in summer.

Declaration of Competing Interest

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

Data availability

Data will be made available on request.

Appendix

In this section, we provide additional information relevant to the proposed multi-agent attention-based DRL controller for demand response in grid-responsive buildings, including background on MADRL along with details on the buildings used for the presented case studies. Brief preliminaries on DRL and the characteristics associated with learning in a multi-agent setting are provided in Section A1. Properties of the energy components installed in the grid-responsive buildings considered in both case studies on New Orleans and New York are also provided in Section A2.

A1. Background on multi-agent deep reinforcement learning

RL is a machine learning paradigm that deals with sequential decision-making [51]. RL can be utilized to handle problems that can be formalized as a Markov Decision Process (MDP). An MDP consists of a set of states S, a set of actions A, a reward function $r : S \times A \in \mathbb{R}$, discount factor $\gamma \in [0,1)$, and the transition dynamics $P : S \times A \times S^{'} \in [0,1]$ that satisfies the Markov property $p(s_{t+1}|s_t, a_t) = p(s_{t+1}|s_t, a_t, ..., s_1, a_1)$. The RL agent learns a deterministic policy $\pi : S \rightarrow A$ or a stochastic one $\pi : S \times A \in [0,1]$ that maps states to actions to maximize the cumulative reward. The traditional value-based learning algorithm, such as Q-learning, is efficient only when the state space is low-dimensional [52]. To overcome the limit of discrete state and action pairs and to address the issue of high dimensionality, DRL uses DNNs as the nonlinear function approximator to replace the table lookup. Meanwhile, policy-based learning, such as the policy gradient method [53], directly learns a policy without calculating the intermediate value function for action selection. Again, DNNs can be used to parameterize the learned policy. Typical value-based learning methods introduce bias by bootstrapping, while policy-based learning methods tend to have high variance by learning through estimations [51]. To this end, actor-critic methods [54] are widely used to harness the advantages of both methods. The action-value function is typically used for the learned value function defined as $Q(s_t, a_t) = \mathbb{E}\left[\sum_{t'=t}^{\infty} \gamma^{t'-t} r_{t'}\right]$.

In this work, we consider the fully cooperative MADRL setting. Specifically, we consider the framework of Markov games [55], a generalization of MDPs. A Markov game for N agents is a tuple $\langle S, A_1, ..., A_N, T, R_1, ..., R_N \rangle$ where S is a set of states, $A_1, ..., A_N$ is a collection of action sets for N agents, $T : S \times A_1 ... \times A_N \rightarrow P(S)$ is the state transition function, which specifies the probability distribution over possible next states given the current state and actions for each agent, and $R_i : S \times A_i \rightarrow \mathbb{R}$ is the reward function for each agent that is a function of global state and actions of all agents. As real-world settings often feature incomplete observations, each agent i receives its local observation, $x_i \in X_i$, which contains partial information from the global state, $s \in S$. Each agent i learns a policy $\pi_i : X_i \rightarrow P(A_i)$ to maximize the expected discounted return, $J_i = \mathbb{E} \ominus_{a_i \sim \pi_i, s \sim T} \left[\sum_{t=0}^{\infty} \gamma^t r_{it}(s_t, a_{it})\right]$ where $\gamma \in [0,1]$ is the discount factor that can be regarded as encoding an increasing uncertainty about future rewards. Treating individual agent independently so that it views the other agents as a component of the environment was one of the first solutions to the MARL problem. The agents in this approach are commonly referred to as independent learners [36], who run separate learning algorithms such as Q-learning. This method has no scalability concern, and each agent only requires local observations during the training and execution [36] phase. However, it may not work well when each agent is using DNNs for approximating its Q-function because of the need for the experience replay to stabilize the training using DNNs [56], which becomes problematic due to the environment's non-stationarity. To perform better, agents can communicate in the environment by learning the time to transmit relevant messages to intended recipients [36]. An alternative method to resolve the non-stationarity problem is to use a fully observable critic. With the fully observable critic that can incorporate the observations and actions of all agents, the environment is stationary despite the change in other agents' policies [36].

A2. Additional details for experimental setup

In this section, we provide additional information relevant to the experimental setup of the MADRL-based control framework for demand response in grid-interactive buildings for the two case studies. The interaction of grid-responsive buildings with the power grid, including the operation of energy storage devices equipped in these buildings, is modeled with CityLearn v1.1.1. For the energy storage devices, each building receives thermal energy supply from storage tanks and energy supply units, such as heat pumps and electric heaters, at every time step. The energy supply devices are sized to meet the energy demand of the building at any given time throughout the simulation to satisfy the presumption that the building temperature

Table A1
Storage capacities of the chilled water tanks, DHW tanks, and PV panels of each building in the New Orleans case study. Note that the chilled water and DHW tanks' capacities are established as multiples of the buildings' maximum cooling/heating energy consumption.

	Energy storage device		
Building ID	Chilled water tank	DHW tank	PV panel [kW]
1	2	2	120
2	3	3	0
3	2	0	40
4	3.5	1.5	25
5	2	2	0
6	3	3	0
7	3	3	0
8	2	3	0
9	3	2	0

Table A2
Storage capacities of the chilled water tanks, DHW tanks, and PV panels of each building in the Cornell University's Ithaca campus case study. Note that the chilled water and DHW tanks' capacities are established as multiples of the buildings' maximum cooling/heating energy consumption.

	Energy storage device		
Building name	Chilled water tank	DHW tank	PV panel [kW]
Baker Lab	1	1.5	720
Bard Hall	1.5	3	210
Bradfield Hall	1	2	650
Carpenter Hall	2	3	160
Clark Hall	2	1	350
Day Hall	1.5	2	150
Mann Library	1	1	600
Warren Hall	1.5	1.5	180
Weil Hall	2.5	2.5	320

setpoints are always satisfied, which enables us to simulate the energy loads of the buildings [35]. Building attributes for the New Orleans case study are borrowed from the CityLearn challenge 2021 [46] and are reported in Table A1. We further analyze the relationship between individual building attributes and their pre-simulated loads in the CityLearn package to select appropriate attributes for the Ithaca, New York case study. The storage capacities of individual energy storage devices equipped in each building are selected such that they would be able to satisfy the corresponding building loads at any given time and are reported in Table A2.

References

[1] "An assessment of energy technologies and research opportunities," *Quadrennial Technology Review. United States Department of Energy,* pp. 12-19, 2015.
[2] A. Roth and J. Reyna, "Grid-interactive efficient buildings technical report series: Whole-building controls, sensors, modeling, and analytics," USDOE Office of Energy Efficiency and Renewable Energy (EERE), 2019.
[3] Wang Z, Hong T. Reinforcement learning for building controls: The opportunities and challenges. Appl Energy 2020;269:115036.
[4] "Benefits of demand response in electricity markets and recommendations for achieving them," in "US Dept. Energy, Washington, DC, USA, Tech. Rep," 2006.
[5] Yang S, Gao HO, You F. Model predictive control for Demand- and Market-Responsive building energy management by leveraging active latent heat storage. Appl Energy 2022;327:120054.
[6] Mariano-Hernández D, Hernández-Callejo L, Zorita-Lamadrid A, Duque-Pérez O, Santos García F. A review of strategies for building energy management system: Model predictive control, demand side management, optimization, and fault detect & diagnosis. Journal of Building Engineering 2021/01/01/ 2021.;33:101692.
[7] Hu G, You F. Multi-zone building control with thermal comfort constraints under disjunctive uncertainty using data-driven robust model predictive control. Adv Appl Energy 2023;9:100124.
[8] Chen W, You F. Sustainable building climate control with renewable energy sources using nonlinear model predictive control. Renew Sust Energy Rev 2022;168:112830.
[9] Henggeler Antunes C, Alves MJ, Soares I. A comprehensive and modular set of appliance operation MILP models for demand response optimization. Appl Energy 2022;320:119142.
[10] Pang S, et al. Collaborative power tracking method of diversified thermal loads for optimal demand response: A MILP-Based decomposition algorithm. Appl Energy 2022;327:120006.
[11] Sharma P, Dutt Mathur H, Mishra P, Bansal RC. A critical and comparative review of energy management strategies for microgrids. Appl Energy 2022;327:120028.
[12] Yang S, Gao HO, You F. Model predictive control in phase-change-material-wallboard-enhanced building energy management considering electricity price dynamics. Appl Energy 2022;326:120023.
[13] Ceusters G, et al. Model-predictive control and reinforcement learning in multi-energy system case studies. Appl Energy 2021;303:117634.
[14] Ajagekar A, Mattson N, You F. Energy-efficient AI-based control of semi-closed greenhouses leveraging robust optimization in deep reinforcement learning. Adv Appl Energy 2023;9:100119.
[15] Vázquez-Canteli JR, Nagy Z. Reinforcement learning for demand response: A review of algorithms and modeling techniques. Appl Energy 2019/02/01/ 2019.;235:1072–89.
[16] Ajagekar A, You F. Deep reinforcement learning based unit commitment scheduling under load and wind power uncertainty. IEEE Trans Sustain Energy 2023;14:803–12.
[17] Lu R, Hong SH. Incentive-based demand response for smart grid with reinforcement learning and deep neural network. Appl Energy 2019;236:937–49.
[18] Jin R, Zhou Y, Lu C, Song J. Deep reinforcement learning-based strategy for charging station participating in demand response. Appl Energy 2022;328:120140.
[19] Kong X, Kong D, Yao J, Bai L, Xiao J. Online pricing of demand response based on long short-term memory and reinforcement learning. Appl Energy 2020;271:114945.
[20] Azuatalam D, Lee W-L, de Nijs F, Liebman A. Reinforcement learning for whole-building HVAC control and demand response. Energy and AI 2020;2:100020.
[21] Zhong S, et al. Deep reinforcement learning framework for dynamic pricing demand response of regenerative electric heating. Appl Energy 2021;288:116623.
[22] Ye Y, Qiu D, Wang H, Tang Y, Strbac G. Real-Time Autonomous Residential Demand Response Management Based on Twin Delayed Deep Deterministic Policy Gradient Learning. Energies 2021;14(3):531.
[23] Lu R, Hong SH, Zhang X. A Dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach. Appl Energy 2018;220:220–30.
[24] Aladdin S, El-Tantawy S, Fouda MM, Eldien AST. MARLA-SG: Multi-Agent Reinforcement Learning Algorithm for Efficient Demand Response in Smart Grid. IEEE Access 2020;8:210626–39.
[25] Nguyen TT, Nguyen ND, Nahavandi S. Deep Reinforcement Learning for Multiagent Systems: A Review of Challenges, Solutions, and Applications. IEEE Trans Cybern 2020;50(9):3826–39.
[26] Ahrarinouri M, Rastegar M, Seifi AR. Multiagent Reinforcement Learning for Energy Management in Residential Buildings. IEEE Trans Ind Inf 2021;17(1):659–66.
[27] Nagarathinam S, Menon V, Vasan A, Sivasubramaniam A. Marco-multi-agent reinforcement learning based control of building hvac systems. In: in Proceedings of the Eleventh ACM International Conference on Future Energy Systems; 2020. p. 57–67.
[28] Zhang Q, Dehghanpour K, Wang Z, Qiu F, Zhao D. Multi-Agent Safe Policy Learning for Power Management of Networked Microgrids. IEEE Trans Smart Grid 2021;12(2):1048–62.
[29] Lu R, Hong SH, Yu M. Demand Response for Home Energy Management Using Reinforcement Learning and Artificial Neural Network. IEEE Trans Smart Grid 2019;10(6):6629–39.
[30] Lu R, Li Y-C, Li Y, Jiang J, Ding Y. Multi-agent deep reinforcement learning based demand response for discrete manufacturing systems energy management. Appl Energy 2020;276:115473.
[31] Zhang X, Lu R, Jiang J, Hong SH, Song WS. Testbed implementation of reinforcement learning-based demand response energy management system. Appl Energy 2021;297:117131.
[32] J. R. Vazquez-Canteli, G. Henze, and Z. Nagy, "MARLISA: Multi-agent reinforcement learning with iterative sequential action selection for load shaping of grid-interactive connected buildings," in *Proceedings of the 7th ACM international conference on systems for energy-efficient buildings, cities, and transportation*, 2020, pp. 170-179.
[33] S. Iqbal and F. Sha, "Actor-attention-critic for multi-agent reinforcement learning," in *International conference on machine learning*, 2019: PMLR, pp. 2961-2970.
[34] 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;311:118636.
[35] J. Vázquez-Canteli, S. Dey, G. Henze, and Z. Nagy, *CityLearn: Standardizing Research in Multi-Agent Reinforcement Learning for Demand Response and Urban Energy Management.* 2020.
[36] Oroojlooy A, Hajinezhad D. A review of cooperative multi-agent deep reinforcement learning. Appl Intell 2022:1–46.
[37] Gelazanskas L, Gamage KAA. Demand side management in smart grid: A review and proposals for future direction. Sustain Cities Soc 2014/02/01/ 2014.;11:22–30.
[38] Vaswani A, et al. Attention is all you need. Adv Neural Inf Proces Syst 2017;30.

[39] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International conference on machine learning*, 2018: PMLR, pp. 1587-1596.
[40] Mnih V, et al. Human-level control through deep reinforcement learning. Nature 2015;518(7540):529–33.
[41] Polyak BT, Juditsky AB. Acceleration of Stochastic Approximation by Averaging. SIAM J Control Optim 1992;30(4):838–55.
[42] Zeng L, Qiu D, Sun M. Resilience enhancement of multi-agent reinforcement learning-based demand response against adversarial attacks. Appl Energy 2022; 324:119688.
[43] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971,* 2015.
[44] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*, 2018: PMLR, pp. 1861-1870.
[45] Crawley DB, et al. EnergyPlus: creating a new-generation building energy simulation program. Energ Buildings 2001;33(4):319–31.
[46] Z. Nagy, J. R. Vázquez-Canteli, S. Dey, and G. Henze, "The citylearn challenge 2021," in *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 2021, pp. 218-219.
[47] "EMCS Portal - Cornell University." portal.emcs.cornell.edu (accessed.
[48] *Visual Crossing Weather* Visual Crossing Corporation. [Online]. Available: https://www.visualcrossing.com/.
[49] Herzmann D, Arritt R, Todey D. Iowa environmental mesonet. Ames, IA: Iowa State Univ., Dep. of Agron; 2004. Available at mesonet. agron. iastate. edu/request/coop/fe. phtml (verified 27 Sept. 2005).
[50] 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: 120860.
[51] Sutton RS, Barto AG. Reinforcement learning: An introduction. MIT press; 2018.
[52] Arulkumaran K, Deisenroth MP, Brundage M, Bharath AA. Deep reinforcement learning: A brief survey. IEEE Signal Process Mag 2017;34(6):26–38.
[53] Sutton RS, McAllester D, Singh S, Mansour Y. Policy gradient methods for reinforcement learning with function approximation. Adv Neural Inf Proces Syst 1999;12.
[54] Konda V, Tsitsiklis J. Actor-critic algorithms. Adv Neural Inf Proces Syst 1999;12.
[55] Littman ML. "Markov games as a framework for multi-agent reinforcement learning," in Machine learning proceedings. Elsevier 1994;1994:157–63.
[56] J. Foerster *et al.*, "Stabilising experience replay for deep multi-agent reinforcement learning," in *International conference on machine learning*, 2017: PMLR, pp. 1146-1155.