Coordinated energy management for a cluster of buildings through deep reinforcement learning

Coordinated Energy Management for a cluster of buildings through Deep Reinforcement Learning

Giuseppe Pinto[a], Marco Savino Piscitelli[a], José Ramón Vázquez-Canteli [b], Zoltán Nagy [b], Alfonso Capozzoli[a*]

[a] *Politecnico di Torino, Department of Energy, BAEDA Lab, Corso Duca degli Abruzzi 24, 10129 Torino, Italy*

[b] *Intelligent Environment Laboratory, Department of Civil, Architectural and Environmental Engineering, The University of Texas, Austin, TX 78712, USA*

* Corresponding author: Tel: +39-011-090-4413, fax: +39-011-090-4499, e-mail: alfonso.capozzoli@polito.it

Abstract

Advanced control strategies can enable energy flexibility in buildings by enhancing on-site renewable energy exploitation and storage operation, significantly reducing both energy costs and emissions. However, when the energy management is faced shifting from a single building to a cluster of buildings, uncoordinated strategies may have negative effects on the grid reliability, causing undesirable new peaks.

To overcome these limitations, the paper explores the opportunity to enhance energy flexibility of a cluster of buildings, taking advantage from the mutual collaboration between single buildings by pursuing a coordinated approach in energy management.

This is achieved using Deep Reinforcement Learning (DRL), an adaptive model-free control algorithm, employed to manage the thermal storages of a cluster of four buildings equipped with different energy systems. The controller was designed to flatten the cluster load profile while optimizing energy consumption of each buildingThe coordinated energy management controller is tested and compared against a manually optimized rule-based one.

Results shows a reduction of operational costs of about 4%, together with a decrease of peak demand up to 12%. Furthermore, the control strategy allows to reduce, the average daily peak and average peak-to-average ratio by 10 and 6% respectively, highlighting the benefits of a coordinated approach.

Keywords: Coordinated energy management, Deep Reinforcement Learning, energy flexibility, grid stability

1. Introduction

The current energy transition is deeply changing the way energy is used and generated. The need of a further decarbonisation of the building sector [1], together with the rapid growth of urban areas, has fostered the use of distributed renewable energy resources. Nonetheless, the rapid penetration of renewable energy sources, characterised by their stochastic behaviour, represents the main cause of an intermittent injection of electricity into the power grid, which can jeopardize grid stability [2]. A recent solution lies in a new paradigm of energy management, which shifts from the supply- to building demand-side control. The latter exploits the novel concept of building energy flexibility, that represents the ability of adapting energy consumption and storage operation without compromising technical and comfort constraints, to increase on-site renewable energy consumption, reduce costs and provide services to the grid (i.e. load shifting, peak shaving) [3]. Among the different strategies aimed at increasing grid stability arises demand response (DR). DR programs are designed to control power demand through different mechanisms that can be classified as i) price-based mechanisms, which aim to encourage consumption in specific periods of the day by reducing tariffs, and ii) event-based mechanisms, such as load curtailment, which are used to preserve network reliability. However, the adoption of price-based programs in some circumstances could be a double-edged sword, causing new undesirable peaks of demand during periods with low electricity prices [4].

In this framework, building energy management should leverage automated algorithms capable to adapt to a changing environment and to learn from user's behaviour and historical building-related data to optimise, coordinate and control the different actors of the smart grids (e.g., producers, service providers, consumers) [5].

A novel approach that aims to exploit the benefits of DR programs while avoiding peak rebounds is represented by the *coordinated building energy management* [6]. This concept arises from the necessity to manage the aggregated power demand of cluster of buildings with the aim of optimising its energy demand shape while considering at single building level user's needs, renewable energy production and the diversity of consumption patterns and energy systems.

1.1.Related works and contributions

Coordinated energy management of buildings can be addressed with a centralised or decentralised approach to the control task, where a centralised controller is assumed to have all the information about the current state of the entire cluster of buildings considered, while decentralised controllers act at single building level [7]. In [8] the authors discussed the advantages of competition, coordination and peer-to-peer transaction to achieve global objectives (e.g., cost minimisation, peak shaving), demonstrating the greater impact of such approach with respect to individual management. In the literature, most of the papers assessing the effectiveness of a coordinated approach in the energy environment, were mainly devoted to electric vehicle charging strategies for providing DR services [9] or shaving peaks [10], and to schedulable appliances for load shifting [11]. Few efforts have been devoted to the coordination at building cluster level of heating ventilation and air conditioning systems (HVACs), which typically represent one of the largest energy end-uses in buildings [12]. This mainly because, with respect to other applications, the management of HVAC systems (including storage systems) is highly influenced by weather conditions, occupant behaviour, comfort requirements, and building features, that highly increase the complexity of the control problem.

Advanced optimal control of HVAC systems at single building level has been widely analysed in the literature, with predictive based control [13],[14]. Among these techniques, model predictive control (MPC) stands out for its ability to optimise complex systems, exploiting a dynamic model to predict building behaviour. However, it requires a detailed model, which enable its application especially at single building scale and rarely at cluster level [15], neglecting the potential benefits of a coordinated approach.

Moreover, model-based approaches are not always effective for real-life implementation at large scale, due to the evolution of the environment (e.g., retrofits, PV integration), and the computational cost associated to the modeling of a cluster of buildings, that can constrain the control scalability.

In this context the many researchers are investigating the use of reinforcement learning (RL) as a valuable control approach in buildings. RL in spite of its adaptive and potentially model-free nature, fits the needs for an effective implementation of energy management in cluster of buildings [16]. RL agent directly learns an optimal control policy through a trial-and-error interaction with the environment, with adaptability potential in case of changes in the environment [17] such as retrofits [18] or demand response programs [19].

In [20] the potential of RL demand response on the market was assessed while in [21] an incentive-based demand response based on RL was proposed. RL control approach has been used in residential demand response [22] or applied to control the operation of different systems, including heat pumps [23], domestic hot water (DHW) [24] and electric water heater [25]. To provide non-intrusive demand response programs, a lot of attention was devoted to the exploitation of control strategy, for indoor temperature setting [26] and thermal storage operation for heating [27] and DHW energy management [28]. Recently, few studies have started to put emphasis on cooperative or competitive coordination mechanisms [29] to account for demand peak shifting when multiple agents take the same control decisions [30].

The presented literature review shows that most studies in the past years have implemented RL for single-agent systems which act greedily and independently of each other, neglecting the opportunity provided by a coordinated control to flatten the peaks on the grid rather than shifting them.

However, it is not surprising that in the past years the need for multi-agent coordination in DR applications was not fully adopted, as the lack of it does not always lead to shifts in the peak demand or "rebound" effects in the daily load profiles. Indeed, in urban settings where the amount of energy storage capacity is not very high, building agents can enable DR without coordination and still be successful in reducing the peaks of electric demand. However, due to the trend towards a wide adoption of electric vehicles and other storage devices such as batteries, this is subject to change in the near future [31,32]. As energy storage devices become more abundant and electrical demand more volatile due to the presence of more renewable energy resources and EV charging stations, properly coordinating all these energy systems in an adaptive manner can be critical without a centralised control or multi-agent cooperation. Nevertheless some pioneering studies have already demonstrated the advantages of a coordinated approach in HVAC systems using advanced control strategies in simulated environment, including heuristic control [33] and reinforcement learning control [34].

This paper explores the opportunity of enhancing demand flexibility for a cluster of buildings by implementing a coordinated energy management, using Deep Reinforcement Learning (DRL) to manage the thermal storages of a cluster of four buildings equipped with different energy systems. The controller was designed with the objective to flatten the total load profile of the cluster while optimizing energy consumption of each building. For benchmarking purposes, the coordinated energy management is then tested and compared against a manually optimized rule-based controller.

On the basis of the literature review the main novelty of the paper can be summarised as follows:

- The paper exploits a single-agent RL centralised controller with a strategy explicitly designed to consider the benefits at multiple levels (i.e., single building, cluster and grid level), against a most common rule-based control strategy that optimise single buildings.
- The paper makes use of a novel simulation environment, CityLearn [35], an OpenAI Gym environment specifically designed to allow RL implementations for the built environment, enhanced to consider a variable electricity price.
- The DRL controller used in this work exploits a state-of-the-art continuous control algorithm i.e., soft actor critic (SAC). The control performances of the agent were deeply analysed to highlights the benefits provided by coordinated energy management.

The paper is organised as follows: Section 2 introduces the methods adopted for developing and testing the controllers, including algorithms and simulation environment. Then, Section 3 describes the methodological framework at the basis of the analysis. Section 4 introduces the case study, explaining in detail the energy modelling of the system and the controllers design and training. Section 5 presents the results of the training and deployment phase, while discussion of results is given in Section 6. Eventually, conclusions and future works are presented in Section 6.

2. Methods

Reinforcement learning is a branch of machine learning mainly aimed at solving control problems. It combines the advantages of dynamic programming, with a trial-and-error approach. RL uses an agent-based control, where the agent learns through the interaction with the controlled environment.

Reinforcement learning can be formalized using a Markov decision process (MDP), a discrete-time stochastic control process [36]. MDP provides a mathematical framework for modelling decision making in situations where outcomes are partly random and partly under the control of an agent.

Markov Decision Process are represented using a 4-tuple (S, A, P, R) made up of:

1. State space (S)

   The state describes the environment completely. Here, must be noticed that state is a term used to represent the environment, while the information seen by the agent, that are a mathematical description of the environment, relevant and informative to the decision to be made are called observations. Often, the agent can see only a part of the state, dealing with the so-called Partially Observable Markov Decision Process (POMDP). In this paper, observation space and state space are considered equal.

2. Action space (A)

   The action is the decision made by the agent on how to control the environment

3. Transition probability (P)

   The transition probability $P(s_{t+1} = s'|s_t = s, a_t = a) = P: S \times A \times S'$ is the probability that, starting in s and performing action a at the time t, the next state will be s'. MDP satisfies the Markov Property, which states the memoryless of the stochastic process, represented as $P(s_{t+1} = s'|s_t) = P(s_{t+1} = s'|s_1, s_2, \dots, s_t)$

4. Reward function (R)

The reward function is used to map the immediate reward r with the tuple $S \times A \times S'$

The main goal of the agent is to find the optimal control Policy (π). A control policy is a mapping between states and actions $\pi: S \rightarrow A$, and it has the aim to maximize the cumulative reward over a time horizon. This concept is summarized introducing the expected return G, that represent the cumulative sum of the reward $G = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$. Where $\gamma \in [0,1]$ is the *discount factor* for future rewards. An agent employing γ equal to 1 considers future rewards as important as current ones, while an agent with γ equal to 0 assign higher values to states that lead to high immediate rewards.

For sake of clarity, an example with an energy system is provided in Figure 1 where the controller (*agent*) is connected to a heat pump and a thermal storage to satisfy the building cooling demand over the summer season. The controller has the role of minimising electricity cost (*reward function*) charging and discharging (*actions*) the storage to satisfy the building demand. The reward function can be minimized charging the storage during night hours, when efficiency is higher and electricity price is low (*states*). The exploitation of these information allows the controller to find the optimal policy.

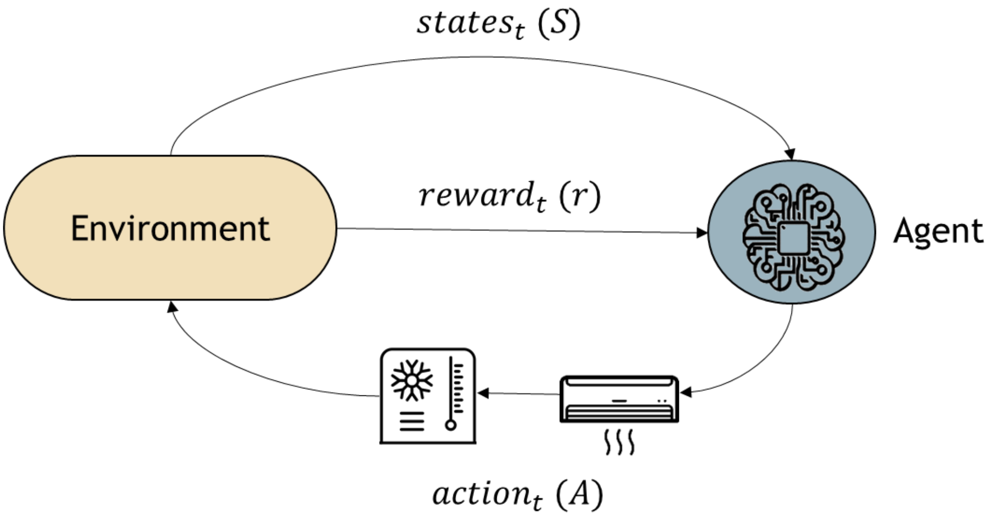

Figure 1: Reinforcement learning control framework

The control problem can be defined by two closely related value function, namely the state-value function $v_\pi(s)$ and action-value function $q_\pi(s, a)$, shown below:

$$v_\pi(s) = \mathbb{E}[r_{t+1} + \gamma v_\pi(s')|S_t = s, S_{t+1} = s'] \tag{1}$$

$$q_\pi(s, a) = \mathbb{E}[r_{t+1} + \gamma q_\pi(s', a')|S_t = s, A_t = a] \tag{2}$$

These functions are used to show the expected return of a control policy π at a state or a *{state,action}* tuple. One of the advantages of the RL is that in its model-free version, the values of v_π and q_π are directly learned from experience, without explicitly calculating the transition probabilities. RL algorithms can improve their policy in two different ways: i) on-policy methods, which attempt to evaluate the policy that is used to make decisions and ii) off-policy methods, which evaluate a policy different from that used to generate the data. Among RL algorithms, the most used one, due to its simplicity, is Q-learning [37]. In Q-learning transitions can be represented with a tabular approach that stores the state-action values (Q-values) that are updated as follows:

$$Q(s,a) \leftarrow Q(s,a) + \mu[Q(s,a) + \gamma \underset{a}{max}\,(Q(s',a') - Q(s,a)] \tag{3}$$

Where s is the next state and μ ∈ [0,1] is the learning rate, which determines to what degree new knowledge overrides old knowledge. When μ is equal to 1 new knowledge completely substitutes old knowledge, while for μ set equal 0 no learning happens. One of the advantages of the off-policy learning lies in their ability to learn from other agent, opening the door to the experience replay, in which previous information is re-used to enhance the current policy. Despite these advantages, a tabular representation of real-world problem may be unfeasible, due to large state and action spaces that needs to be stored.

2.1.Soft actor critic

The combination of RL and high-capacity function approximators such as Deep Neural Networks (DNN) demonstrated to overcome computational problem renewing the interest for the RL topic and promoting its extension to complex problems [38]. Among Deep Reinforcement Learning (DRL) algorithms, an actor-critic method was selected in this paper for its ability to combine advantages of both value-based and policy-based methods. The main idea behind actor-critic is to split the problem using two deep artificial neural networks. The *actor* maps the current state to the action that it estimates to be optimal (policy-based), while the *critic* evaluates the actions by computing the value function (value-based).

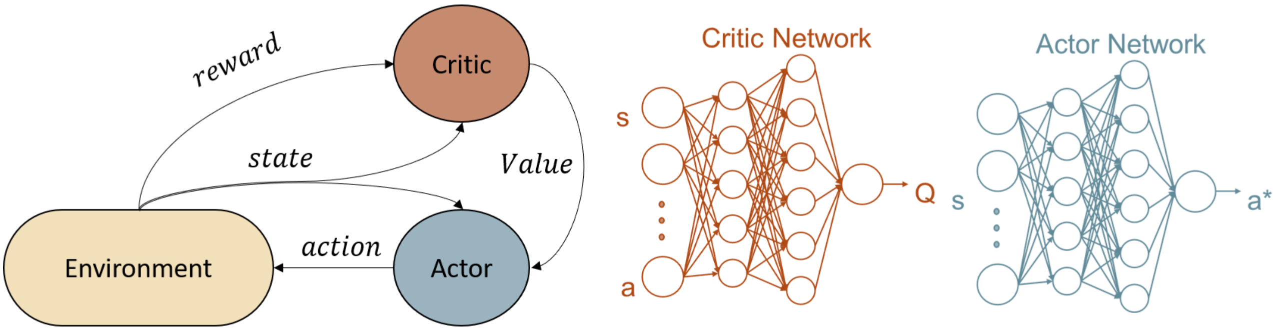

Figure 2: Actor-Critic Environment interaction and neural networks in DRL

The key components of soft actor-critic [39] are:

- An actor-critic architecture, used to map policy and value function with different networks;
- The off-policy formulation, that allows reusing previously collected data, stored in a replay buffer (D) to increase data efficiency;
- The entropy maximization formulation, that helps stabilize the algorithm and the exploration

SAC learns three different functions: (i) the actor (mapped through the policy function with parameters ϕ), (ii) the critic (mapped with the soft Q-function with parameters θ) and (iii) the value function v, defined as:

$$V(\mathrm{s_t}) = \mathbb{E}_{\mathrm{a_t}\sim\pi}[Q(\mathrm{s_t},\mathrm{a_t}) - \ \alpha \log \pi(\mathrm{a_t}|\mathrm{s_t})] \tag{4}$$

$$= \mathbb{E}_{a_t\sim\pi}[Q(s_t,a_t)] + \alpha\mathbb{E}_{a_t\sim\pi}[log\,\pi(a_t|s_t)]$$

$$= \mathbb{E}_{\mathrm{a_t}\sim\pi}[Q(\mathrm{s_t},\mathrm{a_t})] + \alpha H$$

Differently from standard RL algorithm, maximum entropy reinforcement learning optimizes policies to maximize both the expected return and the expected entropy of the policy as follows:

$$\pi^* = \arg\max_{\pi_\phi} \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}\left[r(s_t,a_t) + \ \alpha H\left(\pi_\phi(\cdot\,|s_t)\right)\right] \tag{5}$$

Where $(s_t,a_t)_{\sim\rho_\pi}$is a state-action pair sampled from the agent's policy, and $r(s_t,a_t)$ is the reward for a given state-action pair. Due to the entropy term, H, the agent attempts to maximize the returns while behaving as randomly as possible. The final policy used in the evaluation of the algorithm can be made deterministic by selecting the expected value of the policy as the final action.

The parameters of the critic networks are updated by minimizing the expected error J_Q, which is given by:

$$J_Q(\theta) = \mathbb{E}_{(S_t,a_t)\sim D}\left[\frac{1}{2}\left(Q_\theta(s_t,a_t) - \left(r(s_t,a_t) + \gamma\mathbb{E}_{\mathrm{s_{t+1}}\sim p}[V_{\bar{\theta}}(s_{t+1})]\right)\right)^2\right] \tag{6}$$

Where the value function is implicitly parameterized through the soft Q-function parameters in Equation 6. On the other hand, the temperature parameter α determines the relative importance of the entropy term against the reward, and thus controls the stochasticity of the optimal policy. A high value of the temperature parameters may lead to a uniform behaviour, while a low value of the temperature parameter will only maximize the reward. SAC is highly influenced by the temperature parameter, that needs to be properly tuned to achieve good performances. Unfortunately, tuning this hyper-parameter is very hard, since entropy can vary both across actions and over time, as the policy becomes better. To overcome these problems, in this study a recent version of the SAC that employs alpha automatic optimization [40] was used. To ease the comprehension, the main algorithm logics are summarised in Table 1.

Table 1: soft actor-critic algorithm

```
Input: $\theta_1, \theta_2, \phi$                                                              Initial parameters
  $\bar{\theta}_1 \leftarrow \theta_1, \bar{\theta}_2 \leftarrow \theta_2$                      Initialize target network weights
  $\mathcal{D} \leftarrow 0$                                                                    Initialize an empty replay buffer
  for each iteration do
    for each environment step do
    $a_t \sim \pi_\phi(a_t|s_t)$                                                                Sample action from the policy
    $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$                                                          Sample transition from the environment
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t, r(s_t, a_t), s_{t+1})\}$               Store the transition in the replay buffer
    end for
    for each gradient step do
      $\theta_i \leftarrow \theta_i - \lambda_Q \nabla_{\theta_i} J_Q(\theta_i)$ for $i \in \{1,2\}$    Update of the Q-function parameters
      $\phi_i \leftarrow \phi_i - \lambda_\pi \nabla_\phi J_\pi(\phi)$                          Update policy parameters
      $\alpha \leftarrow \alpha - \lambda \nabla_\alpha J(\alpha)$                              Adjust temperature
      $\bar{\theta}_i \leftarrow \tau \bar{\theta}_i + (1-\tau)\bar{\theta}_i$ for $i \in \{1,2\}$    Soft update of the target network weight
    end for
  end for
Output: $\theta_1, \theta_2, \phi$                                                             Optimized parameters
```

2.2.The CityLearn simulation environment

In order to train, implement and test the developed RL controller, a new simulation environment, CityLearn [35,41], was used. The environment is specifically built to enable training and evaluation of reinforcement learning models for demand response in smart cities through energy simulations with an hourly control timestep. Moreover, the simulation environment makes it possible to control heterogeneous cluster of buildings and offers the possibility to easily implement RL-algorithm to manage cooling and domestic hot water storages with centralised or distributed controllers. The aim is to facilitate and standardize the evaluation of RL agents to enable the comparison of different algorithms. CityLearn is well suited for the easy implementation of both centralized and decentralized multi-agent RL control systems, as well as for the implementation of single-agent independent RL controllers.

The environment allows to control multiple thermal energy storage devices within the building, including water tanks of domestic hot water, cooling and heating systems. The energy demand for space cooling is satisfied by air-to-water heat pumps, and the heating energy is supplied by electric heaters. Furthermore, it also accounts for photovoltaic generation and allows the user to select different performance metrics related to load-shaping. Additionally, users can select up to 28 different state variables which include current weather conditions and forecasts, or the state of charge of the different energy storage devices.

3. Methodological Framework

The section reports the methodological framework adopted in the present paper, with the aim of describing each stage of the process, including the development, training and deployment of DRL control agent. The framework unfolds over four different stages, as shown in Figure 3.

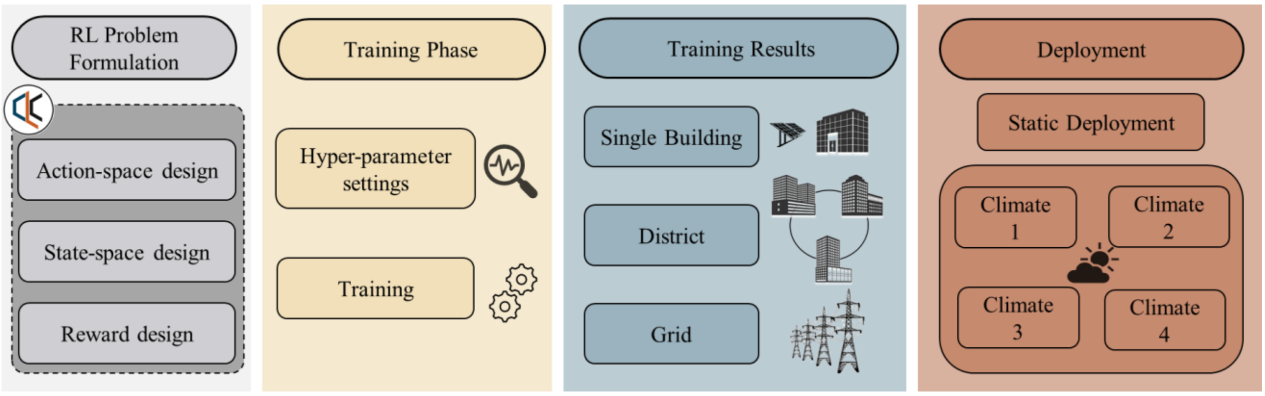

Figure 3: Framework of the application of DRL control

RL Problem formulation: the first stage of the framework was aimed at defining the main components of the reinforcement learning control problem. The *action-space* includes all the possible control actions that can be taken by the control agent. Considering that the aim of the paper is to coordinate multiple buildings, the action space includes multiple actions, 2 for each building. The *state-space* is a set of variables related to the controlled environment which are fed to the agent to learn the optimal control policy which maximizes the reward function. Eventually, a *reward function* was formulated to describe the performance of the control agent with respect to control objectives.

Training phase: in the second stage of the process the DRL agent was trained. As previously introduced in Section 2, DRL agent is characterised by many hyperparameters which require appropriate tuning. In order to enhance the reproducibility of the work, a description about the setting of hyperparameters was provided. The training process was implemented in an off-line fashion using the same training episode (i.e. a time period representative of the specific control problem) multiple times in order to refine agent's control policy.

Training Results: the agent was firstly tested with the same climate used for training, with the aim to specifically analyse the effect of the learned policy on multiple levels, including single buildings, cluster and then on the grid. The performances of the DRL controller were analysed against an RBC controller, by evaluating various key performances indicators (KPI), specifically tailored for each scale of analysis (i.e., single building level or cluster of buildings level).

Deployment: to evaluate the robustness of the trained agent, the algorithm was deployed in four different climates, which also lead to different building thermal-related loads. The agent was tested through a static deployment in one episode and compared with the RBC also during this stage.

4. Case Study

The DRL algorithm described in Section 2.1 was used to control a complex environment that consists of a cluster of 4 commercial buildings, whose load profiles have been assessed through dynamic simulations in EnergyPlus. Each building is equipped with a heat pump, to satisfy cooling demand, an electric heater to meet DHW demand and both cooling and DHW storages. For each building, the heat pump is sized to always match cooling demand, considering a safety factor to account for reduced capacity in case of low external temperature. On the other hand, storages capacity is three times the maximum hourly demand for both cooling and DHW loads [42]. Moreover, two out of four buildings are equipped with photovoltaic systems. Table 2 reports for each building their geometrical features and the main design details of the energy systems. The energy systems are managed by a single-agent centralised DRL controller, which aims to reduce costs and to flatten the aggregated load profile of the cluster reducing peaks.

Table 2: Building and energy systems properties

	Type	Surface [m^2]	Volume [m^3]	Cold Storage Capacity [kWh]	Electric Heater Capacity [kW]	Hot Storage Capacity [kWh]	PV Capacity [kW]
Building 1	Office	5000	13700	235	17	50	0
Building 2	Restaurant	230	710	150	25	75	25
Building 3	Retail	2300	14000	200	23	70	20
Building 4	Retail	2100	10800	185	35	105	0

4.1. Description of the cluster of buildings

The aggregated load pattern of the cluster can result from heterogeneous single building profiles, characterised by both very different intensities and shape. Figure 4 shows the electrical consumption patterns for the first three days analysed. On the left part, it is displayed the load profile for each of the 4 buildings included in the cluster analysed, while on the right part it is showed the total profile. In particular, Building 1 and 4 are characterised by homogeneous daily load profiles, with a peak in the morning, while Building 2 has sudden peak during the evening and Building 3 may have more than a peak per day. As a result, considering that the load profile of the cluster is highly influenced by the single building energy behaviour, to achieve an optimal control at cluster level a coordination at both low and high level is needed. Moreover, the right part of Figure 4 shows at cluster level also the breakdown of electrical demand for cooling, DHW and appliances and the PV production in Buildings 2 and 3.

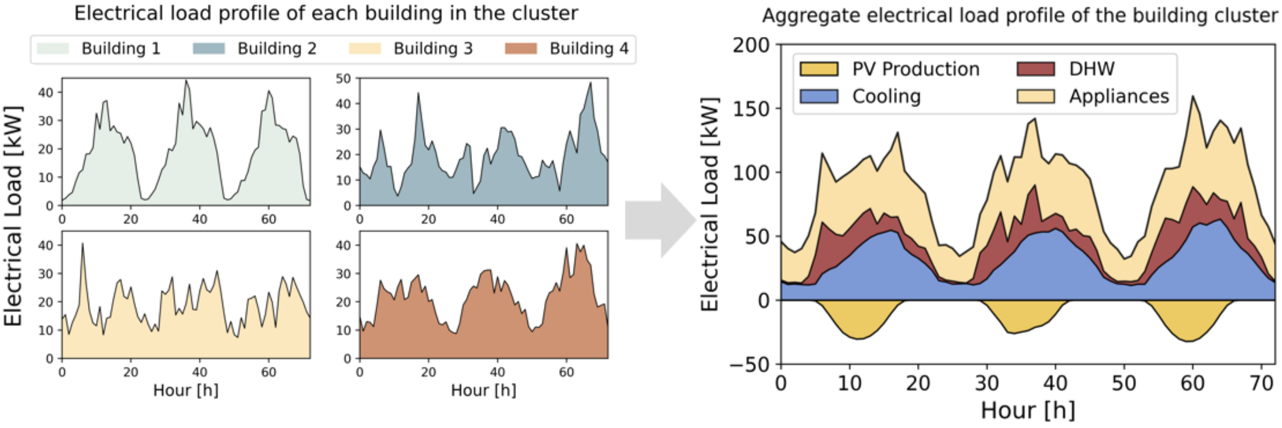

Figure 4: Load Profile for each building (left) and cluster profile electricity and PV production (right)

This representation is useful to underline the electrical demand for cooling and DHW, on which the RL controller can act to enhance cluster flexibility. In fact, since electrical cooling demand represents a large part of the cluster load, the analysis was focus only to the summer period (1st June to 31st August).

4.2. Energy systems and control objectives

The control problem consists in the optimally management of the charging and discharging of the 8 storages to satisfy cooling and DHW demand of the four buildings included in the same cluster. The goal of the control policy is to minimize costs and to avoid peaks at cluster level. The most influencing factor to take into account are the energy cost and the heat pump efficiency. In particular, the energy cost considered in the paper is based on the Austin (Texas) electricity tariffs [43]. In detail, were assumed an off-peak electricity tariff during night-time period 20:00-7:00 (0.03025 $/kWh) and an on-peak electricity tariff during daytime period 7:00-20:00 (0.06605 $/kWh). On the other hand, the efficiency of the heat pump was modified from CityLearn original implementation to consider partial load ratio (PLR) and the effect of external temperature not only on the coefficient of performance (COP), but also on the design capacity. A description of the relation among COP, design capacity (DC) and external temperature was defined according to real data sheet of heat pumps. Moreover, the heat pump operation at part load condition was modelled according to UNI EN 14825 [44].

Eventually, COP was evaluated according to Equation 5. The external temperature rise has a twofold effect, firstly it reduces COP (increasing electricity consumption) and secondly it increases the maximum cooling power deliverable by the heat pump. Moreover, heat pump efficiency is influenced not only by external variables, but also from controller actions affecting the cooling load. Finally, the fraction in Equation 5 accounts for part load ratio and intermitting operation of the heat pump.

$$COP(T,PLR) = COP_T(T) * \left(\frac{\frac{Q_{cooling}(action)}{DC(T_i)}}{0.9*\frac{Q_{cooling}(action)}{DC(T_i)}+0.1} \right) \quad (5)$$

4.3.Baseline rule-based control

The effectiveness of the DRL controller was assessed through a comparison with a manually optimised rule-based controller. In the baseline strategy, both cooling and DHW storages are charged during the night period, when electricity price is lower and heat pumps can take advantage, in terms of efficiency, from lower temperature (i.e., higher values of COP). To limit peak demand, the charging process was spread over the whole night period, while the discharging process is homogeneous throughout the day.

4.4. Design of the deep reinforcement learning controller

The DRL control algorithm described in section 2.1 was trained and tested in the CityLearn environment, including constraints related to the maximum charging and discharging rate of the storages and ensuring that cooling and DHW demands are always met. In the next sub-sections, action space design is presented, together with the description of the reward function design and of the state-space, to properly characterize the DRL control problem.

4.4.1. Action-space design

The analysed case study deals with multiple buildings, each one with two storages that could be controlled. Therefore, the two actions have different targets: the first one is related to the operation of the cold storage, that can be charged by the heat pump to store energy or discharged to meet building cooling load; the second action is related to the operation of the hot storage, that can be charged by an electric heater or discharged to meet DHW demand.

Since each building has different storages and heat pump capacities, the action space makes use of normalized values. In particular, the controller uses actions between -1 and 1, where -1 represents the full storage discharge in the control timestep and 1 represents the full storage charge. However, considering that a full charge/discharge in a single timestep is not feasible, in this work, the action space was constrained into the interval [-0.33,0.33], imposing therefore a complete charge or discharge time of 3 hours according to [42].

In conclusion, at each control time step the agent selects 8 values (one for each storage) to charge or discharge the energy storage devices. This information is used to select the best actions that maximise the reward function.

4.4.2. State-space design

The states represent the environment as it is observed by the control agent. At each control timestep, the agent choses among the available actions according to the values assumed by the states. In particular, states should be easy to measure in real-world implementation, and they should be selected according to the meaningfulness of the information they provide for predicting the reward function. The variables used for representing the state-space are reported in Table 3 and in the following further described.

Table 3: State-space

Variable group	Variable	Unit
Weather	Temperature	°C
	Temperature Forecast (6h)	°C
	Direct Solar Radiation	W/m^2
	Direct Solar radiation Forecast (6h)	W/m^2
District	Total Load	kW
	Electricity Price	€/kWh
	Electricity Price Forecast (1,2,3 h)	€/kWh
	Hour of day	h
Building	Non-shiftable load	kW
	Heat Pump Efficiency	[-]
	Solar generation	W/m^2
	Cooling Storage SOC	[-]
	DHW SOC	[-]

The variables used for representing the state-space can be categorised as weather, district and building states.

Weather states, such as *outdoor temperature* and *direct solar radiation*, were selected because of their strong influence on the magnitude of building loads for space cooling. Additionally, their 6 hours-ahead values were used to provide useful information about temperature and solar radiation changes and enhancing the predictive capabilities of the controller.

District states includes variables that assume the same value for all the buildings over time, such as *hour of day*, *electricity price*, *forecast of the electricity price* and *weather conditions*.

Building states include variables related to the electricity production (photovoltaic system) and consumption of the buildings (*non-shiftable load*). These states are specific of the single building, that can have different energy systems (PV) or trend of consumption. Additionally, *heat pump efficiency, cooling and domestic hot water state of charge* were included.

Figure 5 summarises the states and action space interaction selected in this work. The central controller receives as states high-level information such as weather conditions and electrical demand of the whole cluster of buildings. Moreover, it also receives low-level information from each building such as appliances loads and energy systems information.

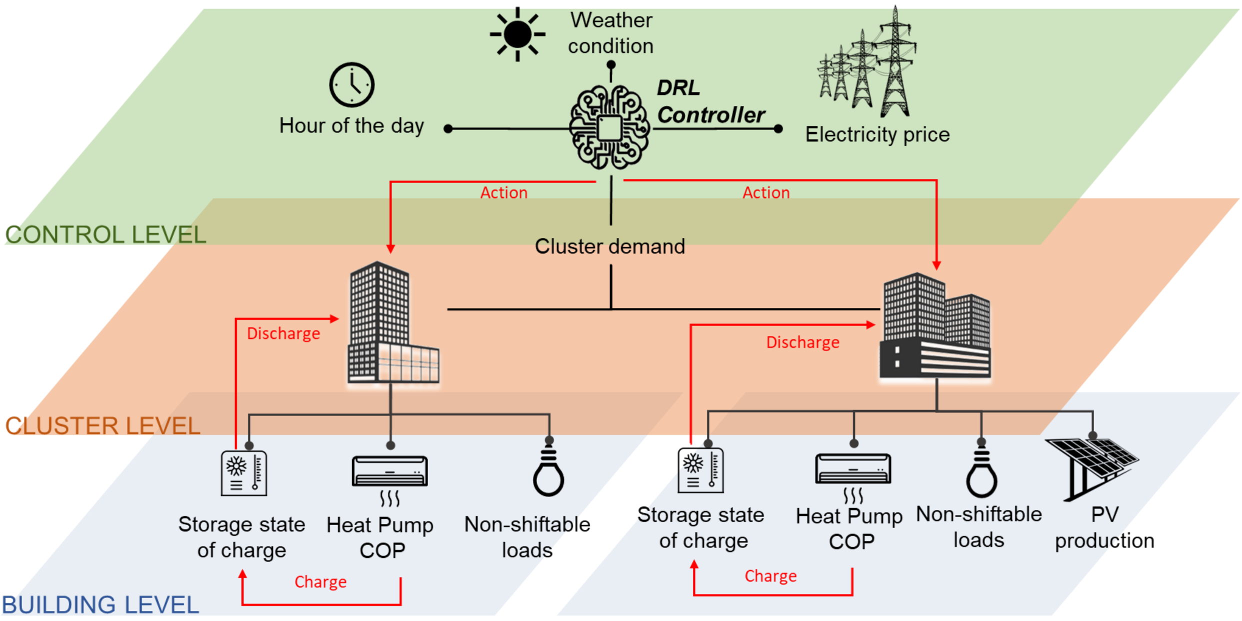

Figure 5: State-action space representation of the DRL controller

4.4.3. Reward design

The reward function plays a key role for defining how the agent assess the quality of the control policy during the learning phase. It was conceived to allow the agent learning a control policy during training period which minimize the energy cost at cluster level and reduce the demand peaks.

In particular, the reward was formulated as follows:

$$R = \sum_{i=1}^{n} e_i^2 * c_{el_i} \tag{6}$$

Where e_i^2 is the squared energy consumption of the i-th building, while c_{el} is the electricity tariff in that time step. To obtain a more uniform load profile at cluster level, the controller tries to minimise the sum of the squared consumption of each building for each time step. This formulation was chosen since the squared minimization approach tries to flatten the profile rather than shifting the consumption to low electricity tariff, avoiding simultaneous charge (and discharge) of storages. In order to consider the economic aspect of the problem, the electricity tariff in the specific timestep was included. Moreover, due to the relation between consumed energy and costs, the controller tries to minimize energy consumption, increasing system efficiency.

The design of the reward function highly influences DRL performances, searching compromises between energy savings and grid stability.

4.5.Training and deployment

The subsection describes the setting of hyperparameters during the training phase. Then, a description of the different climatic conditions analysed for the deployment phase is presented.

4.5.1. Training phase

The DRL framework is characterised by several hyperparameters that strongly affect the behaviour of the control agent. The aim of this subsection is to illustrate the hyperparameters set during the formulation of the control problem. For the sake of reproducibility, Table 4 reports the value of the main hyperparameters. In particular, the two hyperparameters that mostly influence the results are the number of *training episodes* and the *temperature* α . Differently from many other control fields, the number of training episodes is relatively low. This is justified by the problem nature, in which actions are constrained by energy balance, finding the optimal policy quickly. Furthermore, as explained in 2.1 α highly influences the outcome of the policy. While in certain application α could be set a-priori as a constant, in this study a version of SAC algorithm that optimizes the temperature parameter was adopted. As a reference, both starting and final value of temperature and entropy coefficient are provided below.

Table 4: Hyperparameter settings

	Variable	**Value**
1	DNN architecture	2 Layers
2	Neurons per hidden layer	256
3	DNN Optimizer	Adam
4	Batch size	512
5	Learning rate λ	0.003
6	Decay rate τ	0.005
7	Temperature* α	Starting = 1, Final = 0.05
8	Entropy coefficient* H	Starting = 8, Final = 5
9	Target model update	1
10	Episode Length	2208 Control Steps (92 days)
11	Training Episodes	5

As previously stated in Section 4.1, a training episode includes 3 months, from 1st of June to 31st of August (2208 control steps). The weather file used in this work for the training phase is referred to the climatic zone of the USA named 2A, Hot-Humid.

4.5.2. Deployment phase

In the last phase of the process the agent was deployed for the same cluster of buildings but considering four different climates to assess the adaptability capabilities of the learned control policy to different configurations related to the controlled environment. Each agent was deployed for one episode including the period between 1st June and 31st August.

The first climate is 2A Hot-humid: this climate is the same on which the agent was trained on. This scenario is compared to the baseline RBC to assess the effectiveness of the trained agent. Then, the adaptability is tested with the deployment of the agent in warm-humid climate (3A), mixed-humid climate (4A) and cold-humid climate (5A). The thermal related load patterns changed according to climatic conditions.

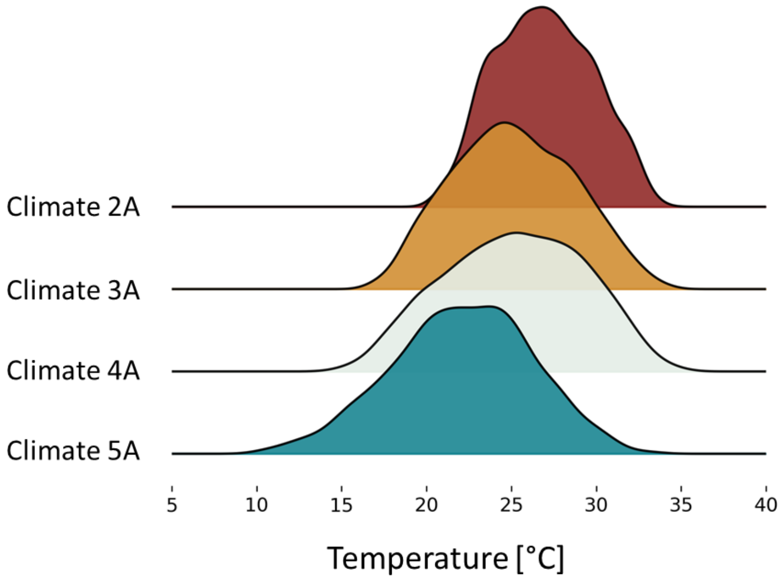

Figure 6: Temperature distribution of the different deployment climates

Figure 6 shows the patterns of outdoor air temperature in the four climates selected, highlighting how the external temperature is strongly different in amplitude and distribution. In particular, the Climate 2A, the one on which the agent is trained on, has a distribution with a narrow amplitude, with a mean temperature of about 27.5 °C. On the other hand, climates 3A and 4A have a different temperature distribution, but the same mean value of about 25.5 °C. Lastly, Climate 5A is the coldest climate considered, with a mean temperature of 22.5 °C and a more uniform distribution with respect to Climate 2A.

5. Results

The section reports the results of the implemented framework. Firstly, a brief evolution of the control policy is presented. Then, a comparison between the two control strategies by analysing the results with a focus at single building scale and cluster level is provided. Eventually, the analysis focuses on the load curve and on the role of storage devices for the grid stability. To this purpose a further comparison is performed computing the load duration curve for the cluster of buildings also considering the case without storages. Furthermore, to summarise the performance of the two control strategies a numeric comparison is provided.

5.1.Training results

The subsection presents the evolution of the DRL control strategy over the training period and compare it with the RBC. In particular, Table 5 reports the evolution of the reward function over the training period, together with the normalized values of cost and peak compared to the RBC (where a value smaller than 1 suggest a better performance of the DRL). The first episode is used to store states and actions and after that it can be observed a quick convergence of both cost and peak term, that stabilize after episode 4. To prove this point, a sensitivity analysis was performed on the number of training episodes, spanning from 5 to 20, which showed little to no improvements with more than 5 episodes.

Table 5: Reward and KPI evolution over training period

	Episode 1	**Episode 2**	**Episode 3**	**Episode 4**	**Episode 5**	**Deployment**
Reward	-343	-337	-297	-271	-271	-265
Cost	1.1	1.1	1.06	0.98	0.97	0.96
Peak	1.07	1.29	0.96	0.96	0.96	0.88

5.1.1. Comparison of controllers at single building level

Figure 7 shows the charging and discharging patterns of both storages determined by the RBC and RL controller. In particular, the figure shows the days related to the maximum peak demand of the RBC, to highlight the difference with the DRL control strategy.

Moreover, the figure shows the relation between the control process and the forcing variables (i.e., external temperature and the electricity price). To allow an easier comparison, all quantities were normalised on maximum values between 0 and 1, where the maximum temperature is 35 °C.

It can be observed that RBC charges both storages mutually at a lower rate, exploiting off-peak tariff and the highest COP of the heat pump. However, to exploit off-peak tariff and avoid sudden peaks, RBC control strategy leads the heat pump to work at part load. Moreover, it has no information about outdoor temperature evolution and so on the efficiency of the heat pump.

On the other hand, the DRL controller learns to charge the two storages as soon as the electricity price and the temperature tend to decrease. However, the main difference is related to the discharge pattern, since DHW is used as soon as needed to reduce electricity demand, while cooling storage is discharged when external temperature is high, avoiding using heat pump when the COP is low.

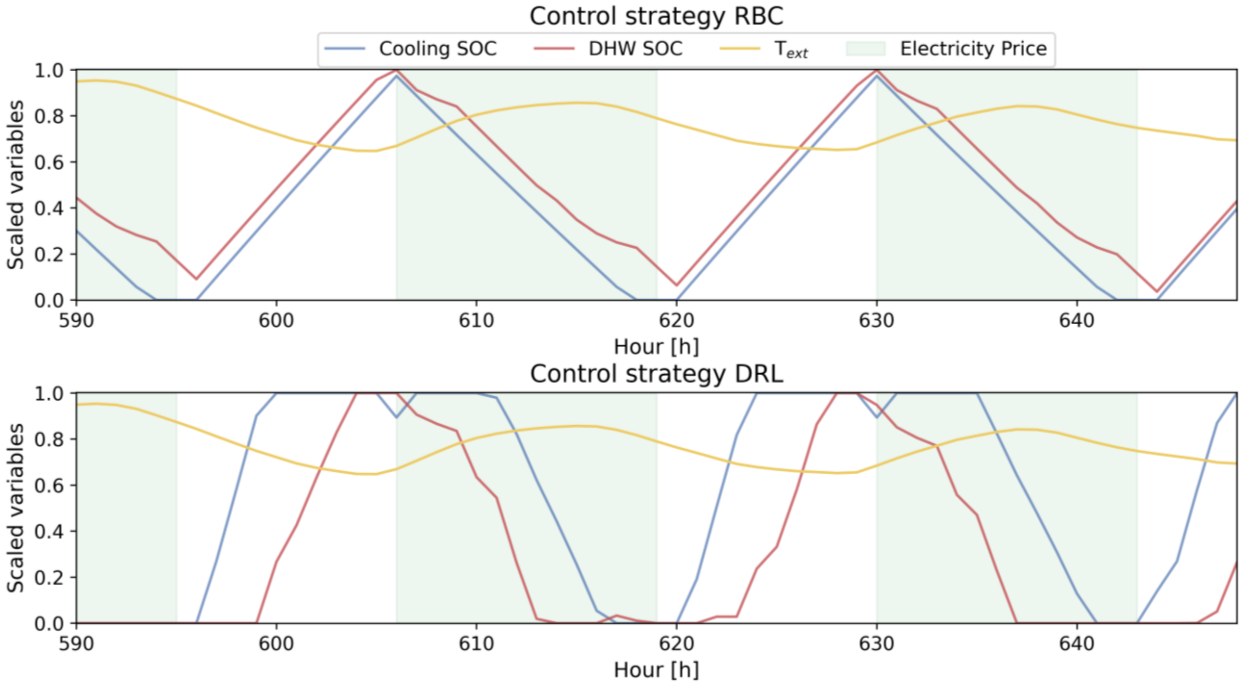

Figure 7: State of charge of storages and forcing variables scaled between 0 and 1

5.1.2. Comparison of controllers at cluster level

Figure 8 shows a comparison between the aggregate electrical load obtained through the implementation of the RBC controller in the simulation environment, in which each building is optimised to minimise its own costs, and the DRL controller, that optimises cluster behaviour. In particular, Figure 8 shows three days during which the RBC determined the occurrence of load peak at cluster level that could cause stress on the grid.

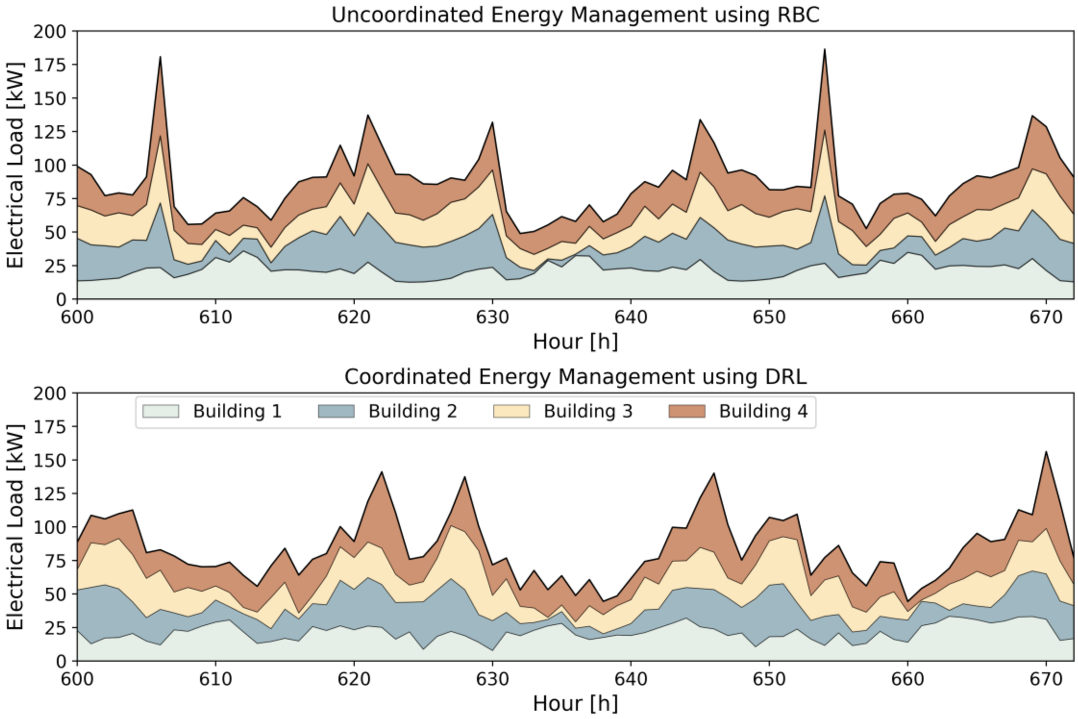

Figure 8: Comparison between uncoordinated and coordinated energy management

As shown in Figure 8, the DRL controller is capable to better flat the aggregate load profile and to diversify the charge time of the storages among the buildings in the cluster. As a result, cluster profile is more homogeneous and, in this particular situation, a great reduction can be observed looking at the two peaks (hour 605 and 655). This result does not represent the average performance of the DRL controller but highlights the potential of buildings coordination in increasing grid stability in specific situations.

To understand how these results have been achieved, Figure 9 shows the average evolution of the state of charge related to the storage device. As can be seen, the cooling and DHW storages are charged during the night homogeneously, as the RBC. The main differences is related to the storage discharges. In particular, as soon as the electricity price increases, the DHW storages start the discharge phase almost simultaneously, since they are only influenced by the electricity price. On the other hand, the agent learned the dependency between external temperature and heat pump COP (the higher the temperature lower the COP). As a result, the optimal policy discharges the cooling storage during the hottest hour, maximising heat pump efficiency.

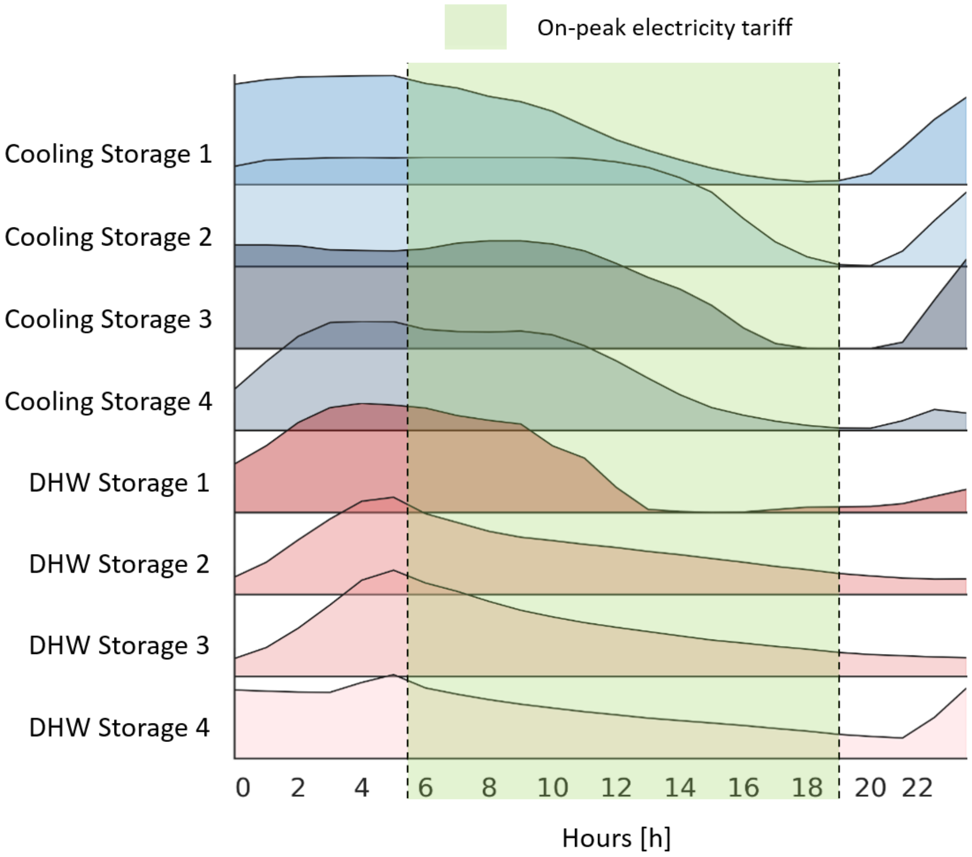

Figure 9: State of charge of storages averaged over a day

5.1.3. Comparison of controllers at grid level

To highlight the flexibility provided by the introduced framework in terms of load profile flattening, the load duration curves resulting from the application of RBC and DRL control and from the case without storages were compared in Figure 10. As can be seen, the baseload increases with both RBC and DRL, underlying the importance of the storages in increasing buildings energy flexibility. However, RBC leads to the creation of new undesirable peaks (as shown inside the “zoom” area) while DRL algorithm, thanks to the coordinated approach, is able to reduce them.

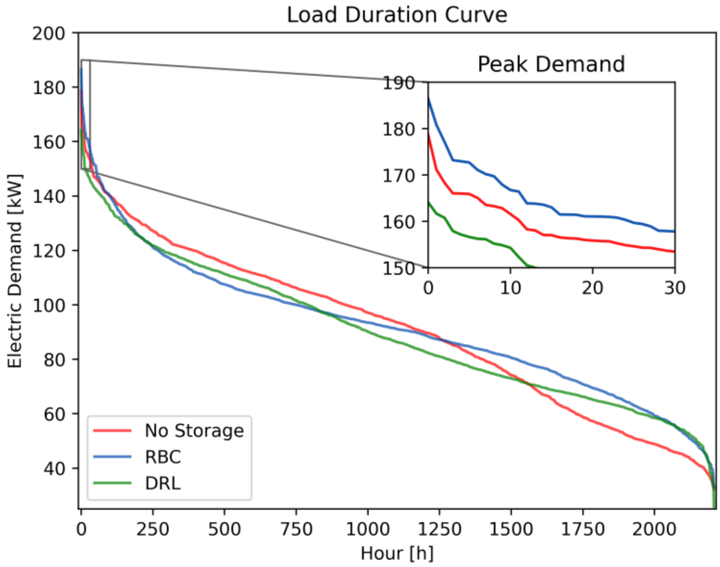

Figure 10: Load duration curve for the base case without energy storages in buildings and the two control strategies

Eventually, to provide a comprehensive analysis of the results different KPIs are introduced to compare the performances of the control strategies. In particular, the KPIs chosen are the *total energy consumption*, the *total energy cost*, *maximum peak*, *average daily peak*, *peak-to-average ratio* (PAR) and *daily peak-to-average ratio*. These KPIs have been chosen to summarise the advantages of DRL control strategies at cluster level (energy consumption, costs, maximum peak and peak-to-average ratio) and the effect on the grid (average daily peak and daily peak-to-average ratio).

Table 6 shows the performance of the two control strategies with respect to the main criteria selected. In order to allow an easier comparison, the values of KPIs are normalised on the RBC values.

Table 6: Comparison between performances of the two control strategies

	Energy Consumption	**Electricity Cost**	**Peak**	**Peak-to-average ratio (PAR)**	**Average daily peak**	**Average daily PAR**
Manually Optimised RBC	1	1	1	1	1	1
DRL	1.03	0.96	0.88	0.96	0.90	0.94

DRL outperforms the manually optimised RBC. In particular, DRL controller exploits the storage charge and discharge to increase heat pump efficiency, while slightly reduces electricity cost. Nevertheless, it must be noticed that the manually optimised RBC already took full advantage from off-peak electricity tariff, therefore the economic improvement of DRL over RBC are closely related to the more efficient use of energy.

On the other hand, the coordinated approach showed good results at cluster level, reducing maximum peak of 12% and average daily peak of 10%. Moreover, the PAR and average daily PAR reduction of 4 and 6% respectively highlight the benefits of building coordination that can be translated into a more homogenous energy consumption.

Furthermore, the advantage provided by the increased grid stability could be translated into reduced electricity tariff, with additional advantages for users.

The DRL approach is able to reduce peaks of 12% with respect to the RBC and 8% with respect to the no storage case, but more importantly the peak demand rapidly decreases, resulting in a more homogeneous profile.

5.2. Deployment of deep reinforcement learning controller in different climatic conditions

The last section analyses the deployment of the agent in the other 3 climates described in 4.5.2. After the training and deployment of the agent in Climate 2A, a simulation of 3 months was run using the trained agent with climate 3A, 4A and 5A. To evaluate the performances of the agent in the new climates, as done before, the DRL controller was compared with the RBC controller, analysing the previously introduced KPIs and normalizing them on the RBC values. Figure 11 summarizes the results of the deployment phase, where 100 represents the RBC performance.

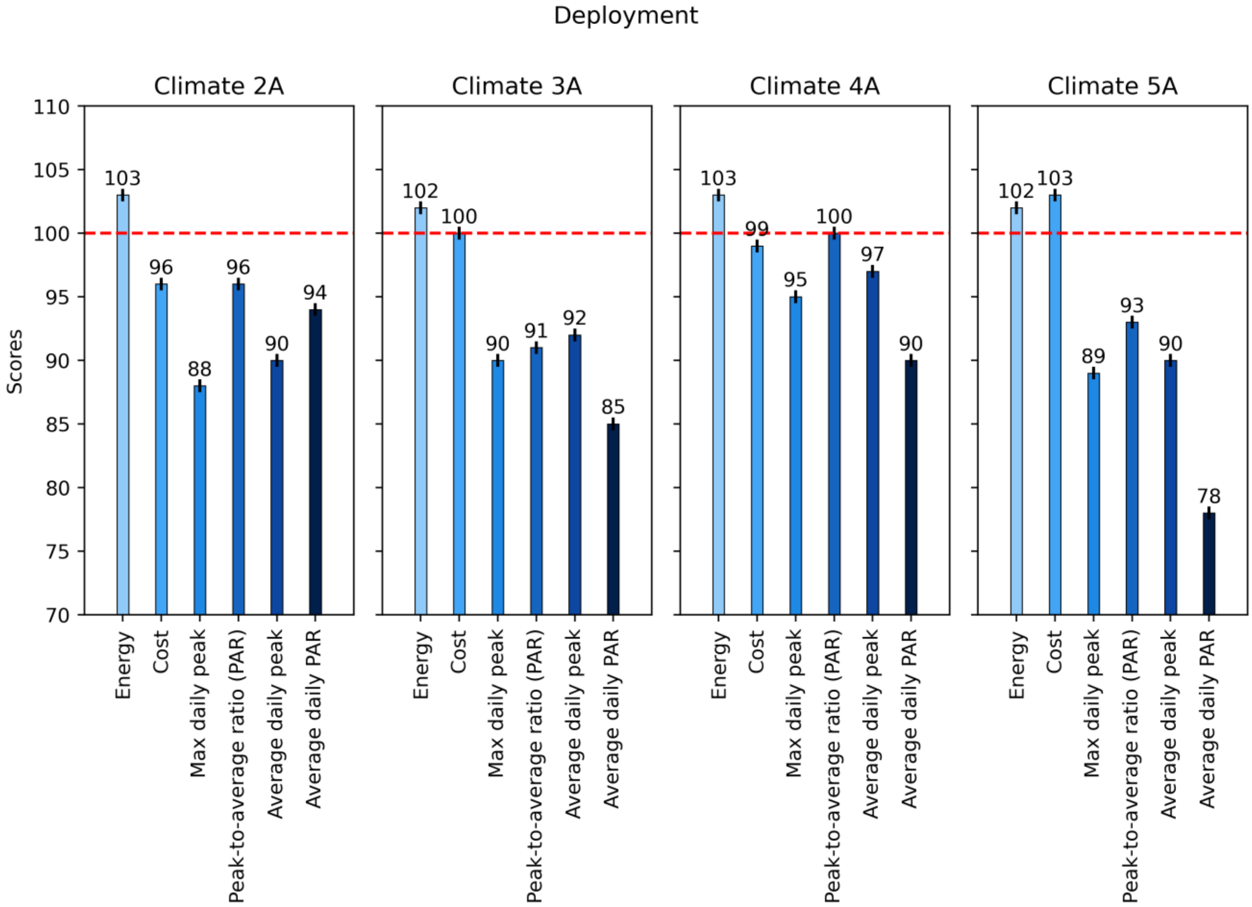

Figure 11: KPI comparison for the four deployment cases

It can be seen how the controller is able, also in Climate 3A, to flatten the load pattern. This is highlighted by the peak and PAR reduction, looking both at maximum and daily values. These results are achieved consuming slightly more electricity with respect to the RBC, but with the same energy cost.

Looking at Climate 4A, it can be noticed a peak reduction of around 5%, but with negligible effect on the PAR. On the other hand, looking at the average daily values, it can be noticed that the daily PAR is 10% lower with respect to the RBC, highlighting the more homogeneous consumption.

Eventually, analysing climate 5A, the coldest one, it can be seen how the cost slightly increases, around 3%, however there are great improvement at district level, with a peak reduction of 11% and an average daily PAR 22% lower with respect to the RBC case.

6. Discussion

The presented paper aims to exploit model-free DRL controller to coordinate the energy management of a cluster of buildings. The analysis is performed with the CityLearn environment, an openAI gym environment where a detailed representation of the heat pump and a variable electricity price have been implemented. The DRL controller was designed to act on the DHW and cold storages of 4 buildings to optimise both energy costs and peak demand at cluster level.

The control problem analysed involved renewable energy sources, variable electricity price and building coordination. To compare the DRL performances and underline the effect of a coordinated energy management versus a single building optimisation, a manually optimised RBC controller baseline was introduced.

Despite the complex environment, the DRL controller found the optimal policy to exploit environment behaviour, consuming energy more efficiently and charging and discharging storages to optimise the cluster load profile. Additionally, due to the problem nature, the solution was found with a very short training period of 5 episodes. The analysis highlights how the real-world implementation could be done with a relatively small amount of data for the training, proving the versatility of the proposed approach. However, to study the interaction among states, actions and rewards it is still necessary a simulation environment when dealing with a district scale.

Looking at the problem formulation, forecast information on electricity price and weather helped to rapidly find an optimal policy, highlighting how important is the proper design of the state-space.

Moreover, the design of the reward function plays a key role for the DRL controller behaviour. It is therefore necessary to find an optimal trade-off between the advantages of single users and cluster that are bounded to the case study. During the work, the adoption of the square minimization was found to be effective at both single building and cluster level, proving to be scalable independently by the number of buildings.

To test its adaptability, the controller was deployed considering four different climatic conditions. The results highlight that the controller flattened the cluster load profile, almost independently from the external conditions, while the economic performances varied with the different cases. Even with the same (or slightly higher) electricity costs, the services provided to the grid, such as peak reduction and load shaping, justify the adoption of the DRL controller with respect to the RBC.

Eventually, the strength of the proposed approach is not only the mere improvement of energy performances, but the opportunity provided by its adaptive nature to account the cluster environment evolution. In fact, a large environment may involve rapid changes, such as consumption pattern modification and demand response programs.

7. Conclusion and future perspectives

The present paper discussed the design and application of a DRL controller with the aim to coordinate multiple buildings in a novel simulation environment. The problem was formulated to provide benefits for users and grid, while a specific analysis to assess how this performance can be achieved was performed.

A fundamental aspect is related to the development of proper state space and reward function. The effectiveness of the reward function was proven by the deployment phase, in which the DRL agent exploited the policy learned during training in different scenarios. In particular, the developed controller has shown a cost reduction of around 4% with peak reduction of 12%. Moreover, daily peaks were reduced on average by 8%, decreasing daily PAR with values that varies from 6 to 22%. The research has shown how the single-agent centralised DRL controller was able to coordinate different buildings, increasing grid stability and reducing energy costs.

Future works will be focused on:

- The implementation and comparison between the proposed centralised controller and a decentralised DRL approach, in which the controllers can cooperate or compete. The opportunity provided by a multi-agent configuration are several. Firstly, specific reward function can be designed and tailored according to renewable electricity production; secondly for each building the relative importance among the objectives as reduce their cost or flatten the profile can be decided, with the aim of facilitating the participation to demand response programs. Moreover, in multi-agent configuration the control policy could be potentially transferred in similar buildings.

- Enhancing the simulation environment for considering variations of the indoor thermo-hygrometric conditions in buildings. In fact, in this study, building loads are evaluated with EnergyPlus software considering a fixed internal temperature. To provide a more realistic implementation, CityLearn will be modified using black-box models to analyse the relation among cooling loads provided to the buildings and the internal temperature. This has the twofold advantages to exploit building thermal mass and consider thermal comfort for users.
- Implementing dynamic electricity price tariff and demand response programs to study the interaction of buildings with the grid. The use of a dynamic electricity price tariff can further increase the benefit provided by a more refined controller with respect to reactive controller. Furthermore, online implementation of adaptive controller could provide an efficient management strategy not only to the DR event, but also to avoid the rebound effect, thanks to the coordinated approach.

A major effort to build upon this research work will be then focused on fully addressing all the mentioned challenges that are behind the next generation of "smart districts" in smart cities.

Nomenclature

Symbols

A = Action space

a = Action

c_{el} = Electricity price

$\mathcal{D}$ = Replay Buffer

e = Energy consumption

G = Return

P = Transition Probabilities

q = Action-value

r = Reward

S = State space

s = State

v = State-value

α = Temperature parameter

γ = Discount factor

θ = Soft-Q network parameters

λ = Learning rate

ϕ = Policy network parameters

τ = Decay rate

H = Shannon Entropy of the policy

π = Policy

π^* = Optimal Policy

Abbreviations

COP = Coefficient of Performance

DC = Design Capacity

DHW = Domestic Hot Water

DNN = Deep Neural Network

DR = Demand Response

DRL = Deep Reinforcement Learning

EMS = Energy Management System

HVAC = Heating, Ventilation and Air Conditioning

KPI = Key Performance Indicator

MDP = Markov Decision Process

MPC = Model Predictive Control

PAR = Peak-to-average ratio

PLR = Partial-load ratio

POMDP = Partially Observable Markov Decision Process

PV = Photovoltaic

RBC = Rule Base Control

RL = Reinforcement Learning

SAC = Soft Actor-Critic

SOC = State-of-Charge

Acknowledgements

References

[1] Leibowicz BD, Lanham CM, Brozynski MT, Vázquez-Canteli JR, Castillo N, Nagy Z. Optimal decarbonization pathways for urban residential building energy services. Appl Energy 2018;230:1311–25. https://doi.org/10.1016/j.apenergy.2018.09.046.

[2] International Energy Agency IEA. IEA statistics: world energy statistics and balances. 2014.

[3] Jensen SØ, Marszal-Pomianowska A, Lollini R, Pasut W, Knotzer A, Engelmann P, et al. IEA EBC Annex 67 Energy Flexible Buildings. Energy Build 2017;155:25–34. https://doi.org/https://doi.org/10.1016/j.enbuild.2017.08.044.

[4] Shen L, Li Z, Sun Y. Performance evaluation of conventional demand response at building-group-level under different electricity pricings. Energy Build 2016;128:143–54. https://doi.org/https://doi.org/10.1016/j.enbuild.2016.06.082.

[5] O'Dwyer E, Pan I, Acha S, Shah N. Smart energy systems for sustainable smart cities: Current developments, trends and future directions. Appl Energy 2019;237:581–97.

https://doi.org/10.1016/j.apenergy.2019.01.024.

[6] Hu M, Xiao F, Wang S. Neighborhood-level coordination and negotiation techniques for managing demand-side flexibility in residential microgrids. Renew Sustain Energy Rev 2021;135:110248. https://doi.org/10.1016/j.rser.2020.110248.

[7] Fiorini L, Aiello M. Energy management for user's thermal and power needs: A survey. Energy Reports 2019;5:1048–76. https://doi.org/10.1016/j.egyr.2019.08.003.

[8] Guerrero J, Gebbran D, Mhanna S, Chapman AC, Verbič G. Towards a transactive energy system for integration of distributed energy resources: Home energy management, distributed optimal power flow, and peer-to-peer energy trading. Renew Sustain Energy Rev 2020;132. https://doi.org/10.1016/j.rser.2020.110000.

[9] Verschae R, Kawashima H, Kato T, Matsuyama T. Coordinated energy management for inter-community imbalance minimization. Renew Energy 2016;87:922–35. https://doi.org/10.1016/j.renene.2015.07.039.

[10] Zhou K, Cheng L, Wen L, Lu X, Ding T. A coordinated charging scheduling method for electric vehicles considering different charging demands. Energy 2020;213:118882. https://doi.org/10.1016/j.energy.2020.118882.

[11] Chang T, Alizadeh M, Scaglione A. Real-Time Power Balancing Via Decentralized Coordinated Home Energy Scheduling. IEEE Trans Smart Grid 2013;4:1490–504. https://doi.org/10.1109/TSG.2013.2250532.

[12] Martinopoulos G, Papakostas KT, Papadopoulos AM. A comparative review of heating systems in EU countries, based on efficiency and fuel cost. Renew Sustain Energy Rev 2018;90:687–99. https://doi.org/10.1016/j.rser.2018.03.060.

[13] Afram A, Janabi-Sharifi F. Theory and applications of HVAC control systems - A review of model predictive control (MPC). Build Environ 2014;72:343–55. https://doi.org/10.1016/j.buildenv.2013.11.016.

[14] Serale G, Fiorentini M, Capozzoli A, Bernardini D, Bemporad A. Model Predictive Control

(MPC) for enhancing building and HVAC system energy efficiency: Problem formulation, applications and opportunities. Energies 2018;11. https://doi.org/10.3390/en11030631.

[15] Gonzato S, Chimento J, O'Dwyer E, Bustos-Turu G, Acha S, Shah N. Hierarchical price coordination of heat pumps in a building network controlled using model predictive control. Energy Build 2019;202:109421. https://doi.org/10.1016/j.enbuild.2019.109421.

[16] Wang Z, Hong T. Reinforcement learning for building controls: The opportunities and challenges. Appl Energy 2020;269:115036. https://doi.org/10.1016/j.apenergy.2020.115036.

[17] Mason K, Grijalva S. A review of reinforcement learning for autonomous building energy management. Comput Electr Eng 2019;78:300–12. https://doi.org/10.1016/j.compeleceng.2019.07.019.

[18] Brandi S, Piscitelli MS, Martellacci M, Capozzoli A. Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings. Energy Build 2020;224:110225. https://doi.org/10.1016/j.enbuild.2020.110225.

[19] Vázquez-canteli JR, Nagy Z. Reinforcement learning for demand response : A review of algorithms and modeling techniques. Appl Energy 2019;235:1072–89. https://doi.org/10.1016/j.apenergy.2018.11.002.

[20] Kazmi H, Mehmood F, Lodeweyckx S, Driesen J. Gigawatt-hour scale savings on a budget of zero : Deep reinforcement learning based optimal control of hot water systems. Energy 2018;144:159–68. https://doi.org/10.1016/j.energy.2017.12.019.

[21] Lu R, Hong SH. Incentive-based demand response for smart grid with reinforcement learning and deep neural network. Appl Energy 2019;236:937–49. https://doi.org/10.1016/j.apenergy.2018.12.061.

[22] Ruelens F, Claessens BJ, Vandael S, De Schutter B, Babuska R, Belmans R. Residential Demand Response of Thermostatically Controlled Loads Using Batch Reinforcement Learning. IEEE Trans Smart Grid 2017;8:2149–59. https://doi.org/10.1109/TSG.2016.2517211.

[23] Vázquez-Canteli J, Kämpf J, Nagy Z. Balancing comfort and energy consumption of a heat pump using batch reinforcement learning with fitted Q-iteration. Energy Procedia 2017;122:415–20. https://doi.org/10.1016/j.egypro.2017.07.429.

[24] Kazmi H, D'Oca S, Delmastro C, Lodeweyckx S, Corgnati SP. Generalizable occupant-driven optimization model for domestic hot water production in NZEB. Appl Energy 2016;175:1–15. https://doi.org/10.1016/j.apenergy.2016.04.108.

[25] Ruelens F, Claessens BJ, Quaiyum S, Schutter B De, Babuška R, Belmans R. Reinforcement Learning Applied to an Electric Water Heater : From Theory to Practice 2018;9:3792–800. https://doi.org/10.1109/TSG.2016.2640184.

[26] Lam E, de Nijs F, Stuckey PJ, Azuatalam D, Liebman A. Large Neighborhood Search for Temperature Control with Demand Response. In: Simonis H, editor. Princ. Pract. Constraint Program., Cham: Springer International Publishing; 2020, p. 603–19.

[27] Vázquez-Canteli JR, Ulyanin S, Kämpf J, Nagy Z. Fusing TensorFlow with Building Energy Simulation for Intelligent Energy Management in Smart Cities. Sustain Cities Soc (under Rev 2018.

[28] De Somer O, Soares A, Kuijpers T, Vossen K, Vanthournout K, Spiessens F. Using Reinforcement Learning for Demand Response of Domestic Hot Water Buffers: a Real-Life Demonstration 2017:1–6.

[29] Kofinas P, Dounis AI, Vouros GA. Fuzzy Q-Learning for multi-agent decentralized energy management in microgrids. Appl Energy 2018;219:53–67. https://doi.org/10.1016/j.apenergy.2018.03.017.

[30] 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: ISBN, editor. BuildSys '20, Yokohama, Japan: Association for Computing Machinery; 2020. https://doi.org/10.1145/3408308.3427604.

[31] Marinescu A, Dusparic I, Taylor A, Canili V, Clarke S. P-MARL: Prediction-based Multi-

Agent Reinforcement Learning for non-stationary environments. Proc Int Jt Conf Auton Agents Multiagent Syst AAMAS 2015;3:1897–8.

[32] Marinescu A, Dusparic I, Clarke S. Prediction-Based Multi-Agent Reinforcement Learning in Inherently Non-Stationary Environments 2017;12. https://doi.org/https://doi.org/10.1145/3070861.

[33] Huang P, Fan C, Zhang X, Wang J. A hierarchical coordinated demand response control for buildings with improved performances at building group. Appl Energy 2019;242:684–94. https://doi.org/10.1016/j.apenergy.2019.03.148.

[34] Vazquez-Canteli J, Detjeen T, Henze G, Kämpf J, Nagy Z. Multi-agent reinforcement learning for adaptive demand response in smart cities. J Phys Conf Ser 2019;1343. https://doi.org/10.1088/1742-6596/1343/1/012058.

[35] Vázquez-Canteli JR, Kämpf J, Henze G, Nagy Z. CityLearn v1.0: An OpenAI Gym Environment for Demand Response with Deep Reinforcement Learning. Proc. 6th ACM Int. Conf. Syst. Energy-Efficient Build. Cities, Transp., New York, NY, USA: Association for Computing Machinery; 2019, p. 356–357. https://doi.org/10.1145/3360322.3360998.

[36] Sutton RS, Barto AG. Reinforcement Learning: An Introduction. MIT Press Cambridge 1998. https://doi.org/10.1016/S0140-6736(51)92942-X.

[37] Watkins CJCH, Dayan P. Q-learning. Mach Learn 1992;8:279–92. https://doi.org/10.1007/BF00992698.

[38] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing Atari with Deep Reinforcement Learning 2013:1–9.

[39] Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. 35th Int Conf Mach Learn ICML 2018 2018;5:2976–89.

[40] Haarnoja T, Zhou A, Hartikainen K, Tucker G, Ha S, Tan J, et al. Soft Actor-Critic Algorithms and Applications 2018.

[41] Vázquez-Canteli JR, Kämpf J, Henze GP, Nagy Z. CityLearn - GitHub repository 2019. https://github.com/intelligent-environments-lab/CityLearn.

[42] Henze GP, Schoenmann J. Evaluation of reinforcement learning control for thermal energy storage systems. HVAC R Res 2003;9:259–75. https://doi.org/10.1080/10789669.2003.10391069.

[43] Austin Energy. Electricity Tariff Pilot Programs n.d. https://austinenergy.com/ae/.

[44] UNI EN 14825:2019 “Condizionatori d’aria, refrigeratori di liquido e pompe di calore, con compressore elettrico, per il riscaldamento e il raffrescamento degli ambienti - Metodi di prova e valutazione a carico parziale e calcolo del rendimento stagionale.” Italy: 2019.