Applied Energy 310 (2022) 118497

Contents lists available at ScienceDirect

Applied Energy

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

Enhancing energy management in grid-interactive buildings: A comparison among cooperative and coordinated architectures

Giuseppe Pinto [a], Anjukan Kathirgamanathan [b,c], Eleni Mangina [c,d], Donal P. Finn [b,c], Alfonso Capozzoli [a,*]

[a] Department of Energy, TEBE Research Group, BAEDA lab, Politecnico di Torino, Italy
[b] School of Mechanical and Materials Engineering, University College Dublin, Ireland
[c] UCD Energy Institute, O'Brien Centre for Science, University College Dublin, Ireland
[d] School of Computer Science, University College Dublin, Ireland

ARTICLE INFO

Keywords:
Deep Reinforcement Learning (DRL)
Building energy flexibility
Soft Actor Critic (SAC)
Multi Agent Reinforcement Learning (MARL)
Grid-interactive buildings

ABSTRACT

The increasing penetration of renewable energy sources has the potential to contribute towards the decarbonisation of the building energy sector. However, this transition brings its own challenges including that of energy integration and potential grid instability issues arising due the stochastic nature of variable renewable energy sources. One potential approach to address these issues is demand side management, which is increasingly seen as a promising solution to improve grid stability. This is achieved by exploiting demand flexibility and shifting peak demand towards periods of peak renewable energy generation. However, the energy flexibility of a single building needs to be coordinated with other buildings to be used in a flexibility market. In this context, multi-agent systems represent a promising tool for improving the energy management of buildings at the district and grid scale. The present research formulates the energy management of four buildings equipped with thermal energy storage and PV systems as a multi-agent problem. Two multi-agent reinforcement learning methods are explored: a centralised (coordinated) controller and a decentralised (cooperative) controller, which are benchmarked against a rule-based controller. The two controllers were tested for three different climates, outperforming the rule-based controller by 3% and 7% respectively for cost, and 10% and 14% respectively for peak demand. The study shows that the multi-agent cooperative approach may be more suitable for districts with heterogeneous objectives within the individual buildings.

1. Introduction

As stated in the European Green Deal, the European Commission has set net-zero carbon emission ambitions for 2050 in response to the emerging climate challenge [1]. Significant progress has been made in decarbonising the electricity sector in recent years, with solar photovoltaic (PV), onshore and offshore wind showing evidence of being promising contributors towards a fully decarbonised energy system [2]. However, renewable solar and wind energy sources are intrinsically variable by nature and this has the potential to create stability issues for the electricity grid with the fluctuating supply needing to be balanced with demand [3]. Villar et al. [4] summarise some of the challenges faced by the new power system paradigm, that is transitioning from a centralised power production, to a decentralised production, thus requiring the need for new flexibility products and markets. The flexibility to manage supply–demand mismatches can come from the supply side (through the use of dedicated standby conventional power plants or storage), through reinforcing interconnections between neighbouring countries or electrical grids [2] or from the demand side [3,5]. Analysing the latter, Demand Side Management (DSM) can be defined as a set of actions that influence the quantity, patterns of use or the primary source of energy consumed by end-users [6]. Demand Response (DR) is one promising pillar of DSM, where consumers curtail or shift their electricity usage in response to financial or other incentives [7].

As buildings represent about 40% of the total primary energy consumption in Europe [8], they are very relevant to participation in DR. A significant portion of building energy demand is towards conditioning the interior spaces for human thermal comfort through the use of Heating, Ventilation and Air Conditioning (HVAC) systems [9]. These loads can often be shifted through the use of active thermal energy storage such as water tanks, and passive thermal mass of the building [10], thus playing an expanding role in the future smart grid [11,12].

[*] Corresponding author.
E-mail address: alfonso.capozzoli@polito.it (A. Capozzoli).

https://doi.org/10.1016/j.apenergy.2021.118497
Received 30 September 2021; Received in revised form 3 December 2021; Accepted 28 December 2021
Available online 28 January 2022

Nomenclature

Acronym

ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
COP	Coefficient of Performance
DNN	Deep Neural Network
DRL	Deep Reinforcement Learning
DR	Demand Response
DSM	Demand Side Management
HVAC	Heating Ventilation and Air Conditioning
KPI	Key Performance Indicator
MDP	Markov Decision Process
MARL	Multi-Agent Reinforcement Learning
MAS	Multi-Agent System
PAR	Peak-to-Average Ratio
PV	Photovoltaic
RL	Reinforcement Learning
RBC	Rule Based Controller
SHW	Sanitary Hot Water
SAC	Soft Actor Critic
SOC	State Of Charge
TES	Thermal Energy Storage

Greek Symbols

γ	Discount rate
π	RL policy
τ	SAC target smoothing coefficient
α	Temperature parameter of SAC

Roman Symbols

A	A set of actions
S	A set of states
C_{peak}	Cost of monthly peak consumption (€)
e_i	Electrical energy consumption of building i (kWh)
$\mathscr{H}$	Entropy
c_{sell}	Grid feed-in tariff (€/kWh)
$P_{Monthly,peak}$	Monthly peak electricity load (kW)
c_{peak}	Monthly peak load electricity tariff (€/kW)
J	Optimisation objective
Pr	Probability
R	Reward function
c_{el}	Time-varying electricity tariff (€/kWh)
P	Transition probabilities between states

However, a building has to be able to meet a minimum required reduction in energy consumption before it can participate in such DR programs, needing to be aggregated or appropriately coordinated to access flexibility markets. While a significant body of literature has investigated the DSM potential of individual buildings, in reality, all entities in a micro-grid setting are interconnected and conventional DSM strategies may result in detrimental effects on the grid reliability (e.g., shifting the peak load to other periods rather than minimising it), limiting the economic benefits for both utilities and consumers [13,14].

In this context, multi-agent systems (MAS) represent a viable alternative to enhance the DSM of multiple entities. Multi-agent systems find their natural use in micro-grid applications, where they are mainly used in power market scenarios [15,16] and micro-grid management [17,18]. MAS leverage several methods, including mathematical methods [19], meta-heuristic methods [20], and heuristic methods, that can be further divided into game-theory based [21] and reinforcement learning based [22–25]. In micro-grid applications, MAS often considers the entire demand as aggregated by a cognitive agent, as done in [26], in which the cognitive agent represents the entire micro-grid demand and coordinates its operation with the generation agents (reactive) to optimise several objective functions including cost, emissions and grid stability. To fully exploit the flexibility associated with buildings, the scale of analysis should be between single buildings and aggregated demand, in the so-called neighbourhood, communities, districts or integrated micro-grid. In this perspective, Labeodan et al. [27] analysed the role of MAS in smart-grid integration, while [28] reviewed the different kinds of MAS applications for smart homes, highlighting the role of MAS architectures, which are briefly described below.

MAS architectures can be classified according to two main categories. The most common ones are the coordinated architecture (centralised) and the cooperative architecture (distributed). Note that these can also be combined to create hierarchical architectures. A brief description of the two main architectures applied to energy management is provided as follows:

- Coordinated energy management exploits a centralised architecture called cognitive-reactive, in which a cognitive agent uses as inputs the observations of all the buildings (reactive agents), that do not have decision-making capabilities, but respond as actuators to the decision taken by the cognitive agent. As a result, coordinated energy management is referred to as centralised training with centralised execution. Hu et al. [13] define coordination in their review of neighbourhood-level coordination and negotiation techniques for DSM as an arrangement of group efforts to harmonise individual efforts in pursuit of common goals. The limitations of this control strategy are the following: (i) the exponential growth of the state and action spaces with the number of reactive agents may limit real-world implementation; (ii) the centralised control may result in sub-optimal solutions for specific buildings; and (iii) private information collection (and their possible sharing) may discourage user participation in a real-world setting.
- Cooperative energy management exploits a distributed architecture, in which each building is represented by an agent that learns the optimal policy according to the specific objective function. As a result, cooperative energy management is referred to as decentralised training with decentralised execution. Hu et al. [13] defines cooperation as voluntary efforts of individuals to work together with the intention of helping each other. The limitations of this approach are the following: (i) the interaction between multiple control strategies can lead to a non-stationary environment thus challenging the learning process; (ii) while the number of agents grows, a large number of models need to be tuned and trained, requiring considerable effort for the definition of reward functions.

1.1. Multi-Agent Reinforcement Learning as a grid-interactive building control framework for districts

Among the different approaches to MAS, Multi-Agent Reinforcement Learning (MARL) has recently attracted growing interest, due to its ability to gradually learn optimal control policies from experiences acquired from the interactions with the environment. For a comprehensive introduction to the broader field of Reinforcement Learning (RL), the reader is referred to standard textbooks [29], while a short literature review of RL applications in the built environment is provided below.

The application of RL in buildings dates back to the 2000s, with the first studies employing it for thermal storage [30–32] and HVAC [33] control, with limited application due to the curse of dimensionality. However, thanks to the introduction of Deep Reinforcement Learning (DRL), the number of applications of RL in buildings has increased, as reviewed by Vázquez-Canteli and Nagy [14] analysing the different kinds of algorithms and modelling techniques of RL for DR. RL has been utilised to control a diverse range of energy systems and one of the advantages lies in its ability to take into account consumer discomfort and integrate human feedback into the control loop. Mason and Grijalva [34] provide a comprehensive review of RL with a focus on autonomous building energy management and Azuatalam et al. [35] review the role of RL for whole-building HVAC control and successfully implement a DR-aware RL controller, able to achieve a maximum weekly energy reduction of 22% compared to a baseline controller. In their review of RL for building controls, Wang and Hong [36] highlight the growing research interest of RL and the potential of MARL to address some of the limitations of other advanced control strategies such as model predictive control. However, the application of MARL for building energy management is relatively new. While some pioneering studies proved the effectiveness of MARL [37–39], further studies need to be performed to explicitly address the advantages deriving from the combination of the different algorithms and architectures in buildings.

The majority of RL works in buildings focused the attention on a single-agent, with the aim to identify effective algorithms for the energy management problem. Indeed, a further challenge is that the energy management problem features continuous state and action spaces (e.g., cooling setpoints, thermal storage state of charge), while many of the typical RL problems face discrete and low-dimensional action spaces, addressed using common algorithm such as Deep Q Networks, which are unlikely to be suitable for such continuous action/state space environments [40].

Among the most recent RL algorithms, the Soft Actor Critic (SAC) algorithm [41] emerges for its ability to handle a continuous action space and it has gained significant interest since its first publication. The effectiveness of the SAC algorithm has been proven in the energy environment [42] and for the energy management of single buildings [43]. Biemann et al. [44] compared SAC algorithm with other three actor-critic algorithms to control the HVAC of a data centre, finding that SAC algorithm showed substantial improvement in both performance and sample efficiency. Pinto et al. [45] used a centralised SAC architecture to optimally control four buildings using the Learn environment [46], while Deltetto et al. [47] exploited the latter to assess the potentialities of a centralised SAC controller for incentive-based DR in a small district of commercial buildings. Although both works provided insights on the electricity cost of the district, the information of the costs associated to each building was not investigated. Since the centralised algorithm does not ensure a bottom-up optimisation, some buildings may face an increase of costs, that should be carefully assessed. In this framework, it may be useful to analyse the effectiveness of different RL architectures. An initial attempt to compare multiple SAC architectures for buildings control was performed by Dhamankar et al. [48]. The authors provided an empirical comparison of independent learners (distributed architecture), centralised critics with decentralised execution (centralised architecture) and value factorisation learners (hybrid architecture). The main limitation of that work is related to the comparison of an average metric, that does not allow to understand the strength and weakness of each approach, especially shifting the attention from the district to single buildings.

1.2. Contributions and structure of this work

The literature review presented in the previous section revealed the following research gaps: the application of advanced control strategies for DSM to date has largely been confined to single buildings. Among the few studies that analysed multiple grid-interactive buildings, the focus has been on micro-grid applications with appliance scheduling or electric vehicles, requiring further analysis on the role of thermostatically controlled loads and thermal storage for grid-interaction.

Moreover, there is a lack of studies aimed at comparing different control architectures when dealing with heterogeneous energy systems. Indeed, individual buildings may have their own independent objectives and the way by which such individual objectives, when part of a district, influence control design problems, needs to be further investigated.

Lastly, considering the multi-objective nature of the grid-interactive DSM problem, a detailed analysis of the advantages and disadvantages of each architecture/algorithm is required.

With these research gaps in mind, this work provides the following contributions and novelty by:

1. Comparing the performance of a coordinated (centralised) and cooperative (decentralised) MARL architecture for the provision of DSM in a district of heterogeneous buildings.
2. Analysis of a DRL controlled grid-interactive district at different scales and time. Assessment of advantages and limitations of the proposed architectures for specific buildings and the entire district.
3. Studying the application of a multi-agent SAC RL algorithm to a district DSM problem with heterogeneous buildings, testing their robustness in different conditions and assessing the versatility of different controller architectures.

The presented paper deals with the energy management of four buildings, equipped with thermal energy storage and PV systems and formulating the problem as a reinforcement learning based one. Two SAC-MARL algorithms are explored: a centralised (coordinated) controller and a decentralised (cooperative) controller, which are benchmarked against a rule-based controller (RBC) that aims at exploiting electricity tariffs to minimise the cost.

The paper is organised as follows: Section 2 presents a detailed description of the proposed methods, together with essential background on the concept and formulation of the RL problem and SAC algorithm. Section 3 provides a description of the case study and control problem, followed by the baseline reference controller and KPIs used for comparison. Section 4 presents the design process of the two DRL controller architectures. Section 5 provides the results of the key findings with focus on both the comparison of the various MARL architectures against the baseline controller and the robustness of the agents under different climate types. Furthermore, Section 6 provides a critical discussion on these results. Section 7 gives the conclusions and summarises potential future research directions to enable and enhance the further use of the SAC MARL technique for real-life energy flexibility applications.

2. Methods

In this section, the methodological framework for the development and assessment of the performances of the two RL architectures (coordinated and cooperative) is presented. In particular, the methodology is structured in three steps, as represented in Fig. 1 and described below in further detail.

Control Problem Definition: The first step describes the environment used for MARL (Section 2.1) and the case study district (Section 3.1). The latter firstly describes the analysed buildings, with a focus on the controllable energy systems and the uncoordinated RBC, which is used as baseline. Lastly, it provides a description of weather data used to test the robustness of the proposed control strategies. Section 3.3 describes the control problem and outlines the electricity tariffs which support the more flexible operation of the energy systems and the reference baseline controller. To quantify controller performance and

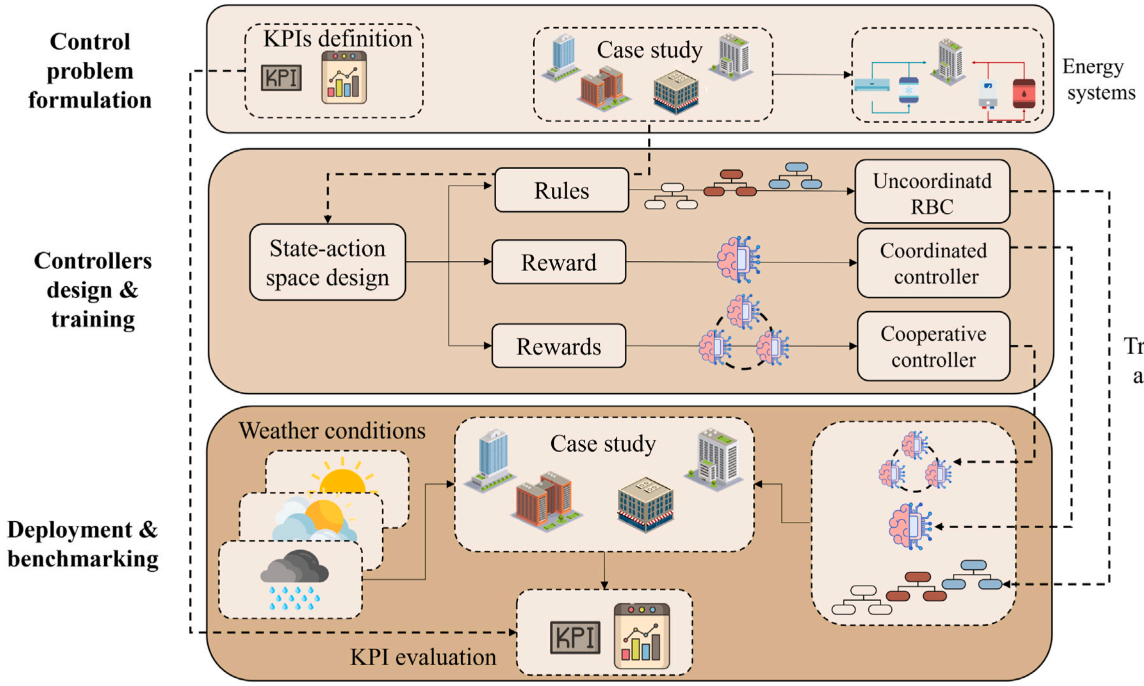

Fig. 1. Methodological framework overview.

allow comparisons to be made between the controllers, Section 3.4 introduces the set of specific KPIs used in this study.

Controller Design & Training: The second step of the methodology analyses the main components of the RL problem. In particular, Section 4.1 presents the design of the action-space to analyse the possible control actions that can be taken by the agents. Section 4.2 similarly presents the state-space design which is the information provided by the environment to the RL agents. Section 4.3 formulates the reward functions for each architecture analysed to quantify controller performance with respect to control objectives.

Deployment & Benchmark: The last step focuses on the deployment and benchmarking of the trained agents. In particular, to test their robustness, the controllers are deployed in several climates, previously described in Section 3.1 and their performance is evaluated through several KPIs (Section 3.4).

2.1. CityLearn environment

The work exploits a modified version of the CityLearn environment [46] to train and deploy the various MARL controllers. CityLearn is a simulation environment based on OpenAI Gym, specifically built to enable training and evaluation of RL models at the building and district levels. The simulation environment allows control, with an hourly timestep, of a district of buildings managing cooling and sanitary hot water (SHW) storage with different architectures (centralised or distributed). The environment structure, with a detailed description of the metadata and input related to its functioning are provided by Vázquez-Canteli et al. [46]. The original environment was modified by focusing on a district of only four buildings (described further in Section 3.1). Further, the heat pump size and its COP evolution are defined according to [45], taking into account the reduced capacity during low external temperature periods and the effect of external temperature and partial load ratio.

2.2. Reinforcement Learning

RL is an aspect of artificial intelligence, where an agent learns to take the optimal set of actions through interaction in a dynamic environment (such as a building subject to changing weather conditions and varying grid requirements), with the goal of maximising a certain reward quantity [34].

The traditional RL problem can be formalised as a Markov Decision Process (MDP) containing four elements:

1. S, a set of states (e.g., thermal energy storage state of charge, outdoor relative humidity)
2. A, a set of actions (e.g., thermal energy storage discharge rate)
3. $r : S \times A$, a function describing the reward as a result of taking a specific action
4. $P : S \times A \times S' \in [0, 1]$, transition probabilities between the states

Given a state, for every action that the agent takes, this leads to a new state in the environment and based on this, the agent is either rewarded or penalised for taking that particular action. The reward is a feedback mechanism to the agent to indicate how well it is performing at each time step. For the state S_t to satisfy the Markov property, the future state must only be dependent on the current state and current actions, i.e., the future state is independent of the past state, given the present state [29]. The learning problem is further complicated when considering real-world applications featuring many agents. In this case, the learning process is more challenging, as each agent sees a non-stationary environment that is also changing due to the actions of the other agents [14]. In fact, the Markov property may become invalid due to this non-stationarity and hence MDPs do not provide the same theoretical grounding for such MARL problems [49]. According to the employed MARL architecture, the problem can be classified as an MDP characterised by a joint action space with a single reward in a centralised setting, or as a Markov Game characterised by multiple action spaces and rewards in a decentralised setting. Wong et al. [50] provide a high-level overview of the multi-agent learning problem, detailing some of the above issues.

2.2.1. Soft actor critic deep reinforcement learning

The soft actor-critic (SAC) algorithm, an off-policy maximum entropy actor-critic algorithm, as first proposed by Haarnoja et al. [41] is used in this research. An actor-critic method has been selected for its ability to combine advantages of both value-based and policy-based methods.

The SAC differs from traditional actor–critics insofar as the SAC maximises the information entropy of state apart from the conventional cumulative rewards. Standard RL maximises the expected sum of rewards:

$$J(\pi) = \sum_t \mathbb{E}_{(s_t,a_t)\ \rho_\pi}[r(s_t,a_t)] \tag{1}$$

SAC, however, favours stochastic policies and it does this by modifying the objective function with an additional term of the expected entropy ($\mathscr{H}$) of the policy:

$$J(\pi) = \sum_t \mathbb{E}_{(s_t,a_t)\ \rho_\pi}[r(s_t,a_t) + \alpha\mathscr{H}(\pi(\cdot|s_t))] \tag{2}$$

Here α and $\mathscr{H}(\pi(\cdot|s_t))$ is the trade-off between entropy and reward. The advantage of entropy maximisation is that it can lead to policies that can explore more and are able to capture multiple modes of near-optimal strategies [41]. Increasing entropy can also prevent the policy from prematurely converging to a bad local optimum. The control problem can be extended to an infinite horizon by introducing the discount factor γ, the value of which varies between 0 (that prioritises policies with high immediate rewards) and 1 (which considers future rewards as important as current ones). At test time, stochasticity is removed and the mean action is used instead of a sample from the distribution. The SAC algorithm is summarised in Algorithm 1 with the full details available in [41].

Algorithm 1: SAC algorithm adapted from [41]

```
Input: Policy (actor) and soft-Q (critic) DNNs
Initialise target network weights
Initialise experience replay buffer with random policy samples
for each episode do
    for each step do
        sample actions from policy
        sample transition from the environment
        store the transition in the replay buffer
    end
    for each gradient update step do
        update the soft-Q DNN weights
        update the policy DNN weights
        update the target DNN weights
    end
end
Output: Optimised actor and critic DNNs
```

The SAC DRL agent was developed in Python using the PyTorch library [51]. The version of SAC implemented in this paper assumes a constant entropy regularisation coefficient (α) over the course of training.

3. Case study district & control problem

This section provides a description of the case study. In Section 3.1, the energy systems and weather climates used for the analysis are outlined. Next, the control problem is analysed in Section 3.3 and lastly, the KPIs used for the analysis are presented in Section 3.4.

3.1. District

The district includes four buildings: a restaurant, and three multi-family buildings, which can be further demarcated as prosumers which do not export electricity (Building 2 and Building 4) and prosumers which export electricity (Building 1 and Building 3). Each building is equipped with PV panels, a reversible heat pump, an electric heater and two storage devices (chilled water and SHW). The control problem focuses on the energy management of the two storage devices per building, with the aim of optimising costs, profile shape and self-consumption. To quantify the effects of the control strategy, several KPIs, described in the next subsection, have been used.

Table 1
Climate zones (per ASHRAE definitions) considered in this study.

Climate zones	Location	T_{min} [°C]	T_{mean} [°C]	T_{max} [°C]	T_{σ} [°C]
2A	Houston, TX	20.0	27.5	35.5	3.0
3A	Atlanta, GA	16.0	25.5	36.0	4.0
5A	Chicago, IL	8.5	22.0	35.0	4.5

The district electrical load is mainly influenced by the building cooling loads and, as a result, the analysis focuses only on the summer period (defined as the 1st June to 31st August), which represents the simulation period used in this study.

Moreover, as weather conditions influence the cooling load and control strategy, the effects of weather variation on the behaviour and robustness of the controllers was analysed. Whilst studies have investigated the ability of DRL to adapt to different operating conditions (e.g., weather conditions [52], occupancy and set point changes [53]), there is a necessity to study how multiple agents address these changes for each of the cooperative and coordinated environments, which may lead to a non-stationary problem. On the grounds of this, each agent is trained on one climate (2A) and further deployed in the other two climates (3A and 5A), as summarised in Table 1. The climate zones considered are diverse in nature and are as per the ASHRAE standard definitions. This analysis aims to evaluate and compare the ability of the two controllers to adapt to different environmental conditions.

3.2. Energy systems at building level

Fig. 2 shows a schematic of the control architecture with details of the energy systems for a representative building of the district, while a comprehensive formulation of the mathematical problem can be found in [54]. In particular, the scheme highlights the controlled systems (chilled water and SHW storage) and their interaction with other energy systems. The heat pump can either charge the chilled water storage and satisfy the heating and cooling energy demand of the building, although the current analysis only focuses on the summer period. The electric heater is used to charge the SHW storage and to meet the SHW demand, while non-shiftable loads can be satisfied using electricity from PV or imported from the grid. Furthermore, Table 2 reports in detail the geometrical features of the buildings, together with the capacity of the two controlled systems (storage) and the PV size.

It can be noticed that, despite having the same floor area, the three multi-family buildings are characterised by different cooling, heating and appliances loads. Indeed, to represent user stochasticity, probabilistic regression models were trained from different open source datasets [54] to create realistic instances of indoor temperature set point, SHW consumption and appliances schedules. Accordingly, the two storage devices are sized to satisfy three times the maximum hourly demand, of cooling and SHW loads respectively, while the heat pump and electric heater are sized to always ensure the meeting of building loads [54]. Based on this information, an optimal control strategy should leverage PV electricity to partially offset non-shiftable, cooling and SHW loads, or even charging thermal storage during renewable overproduction periods, exploiting the energy multi-carrier nature of the control problem.

Lastly, to analyse the contribution of renewable electricity to the building load, Fig. 3 displays PV self-consumption and export for each building, together with their net load for the first three days of the simulation period for climate 2A. As highlighted earlier, Building 1 and Building 3 are prosumers, exporting a certain quantity of energy. On the other hand, Building 2 and Building 4 self-consume renewable energy. It is crucial to notice that the building electrical demand, affected by climatic conditions, directly determines the ability of a prosumer to

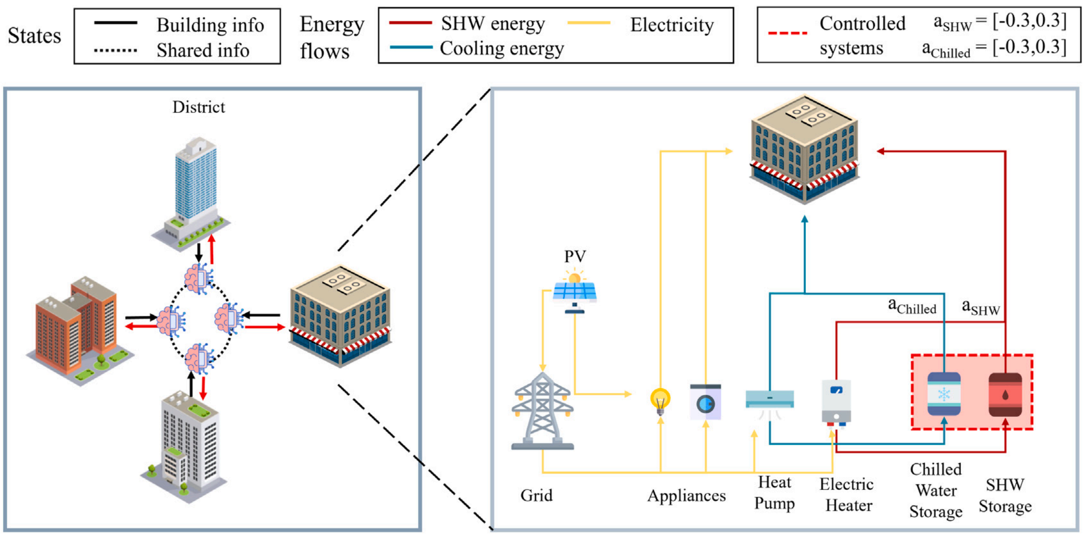

Fig. 2. Building energy management control scheme.

Table 2
Summary of building geometrical features and energy systems in district.

	Type	Floor area [m^2]	Volume [m^3]	TES capacity [kWh]	SHW storage capacity [kWh]	PV capacity [kW]
Building 1	Restaurant	230	710	235	50	50
Building 2	Multi-family	3130	9550	150	75	20
Building 3	Multi-family	3130	9550	200	70	60
Building 4	Multi-family	3130	9550	185	105	20

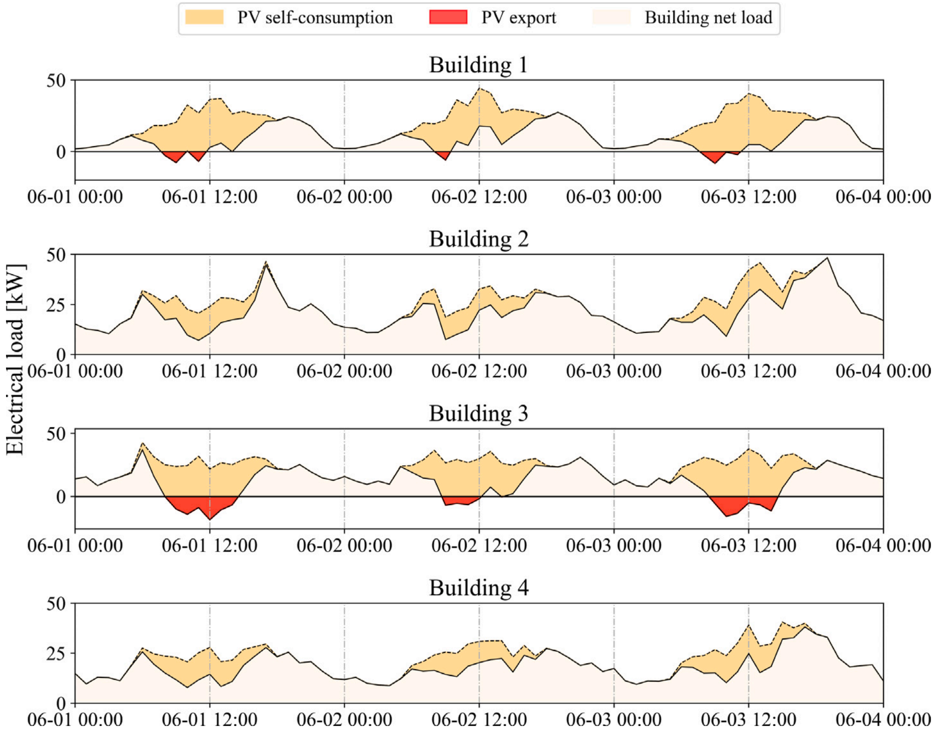

Fig. 3. Electrical load profile for each building in the district for Climate 2A.

export electricity or not. The energy systems are managed with two controller architectures (Section 4), which aim to reduce operational costs and to flatten the aggregated load profile, exploiting the existing sources of flexibility.

3.3. *Definition of the control problem*

The controllers were designed to manage the charging and discharging of chilled water and SHW storage systems for the district of buildings, with the aim to minimise electricity costs, reduce cluster demand peaks and maximise self-consumption. The electricity price

Table 3
Electricity tariff including energy terms and peak terms [55].

	On-peak [$/kWh]	Off-peak [$/kWh]	Sold [$/kWh]	Peak [$/kW]
Price	0.0549	0.0189	0.0100	11.02

Table 4
KPIs Used in MARL controller comparisons.

KPI	Formula	Units
Cost	$\sum_{i}^{n} e_i * c_i$	[$]
Peak	$\max \sum_{i}^{n} \frac{e_i}{\Delta t}$	[kW]
Daily-Peak	$\frac{\sum_i^{n_{day}} Peak_{day}}{n_{day}}$	[kW]
Peak-to-average ratio (PAR)	$\frac{Peak}{\sum_i^n e_i / n_{day}}$	[-]
Daily-PAR	$\frac{\sum_i^{n_{day}} PAR_{day}}{n_{day}}$	[-]
Self-sufficiency (SF)	$\frac{\sum_i^n \sum_{j=1}^T \min(PV_{i,j}, e_{i,j})}{\sum_i^n e_i}$	[%]
Flexibility factor (FF)	$\frac{\sum_i^n e_{i,off-peak}}{\sum_i^n (e_{i,off-peak} + e_{i,on-peak})}$	[-]

tariffs and the PV production are the main drivers of the district cost. In particular, the electricity price (c_{el}), chosen according to [55], varies from $c_{el,off-peak}$ = 0.01891 $/kWh during off-peak hours (21.00–12.00) to $c_{el,on-peak}$ = 0.05491 $/kWh during on-peak hours (12.00–21.00). Moreover, a cost related to the monthly peak load of the district was considered and defined below:

$$C_{Peak} = c_{Peak} * P_{Monthly,Peak} \tag{3}$$

Where c_{Peak}= 11.02 [$/kW] is the tariff related to the monthly peak load $P_{Monthly,Peak}$ [kW], evaluated as the maximum district load for each month. In the context of coordinated energy management, if the cluster of buildings is managed by the same aggregator, it would face a cost related to the district monthly peak demand, that the controller should try to minimise, since it could represent a not negligible part of the total cost faced by the district. Furthermore, any electricity overproduction can be sold to the grid according to the following tariff: c_{sell}= 0.01 $/kWh. The electricity tariffs are summarised in Table 3.

To benchmark the performance of the two DRL architectures, a RBC was used as baseline. The RBC uses a distributed architecture, with the aim to minimise individual building energy cost. This is achieved by exploiting the electricity tariff, charging chilled water and SHW storage over the night period and discharging uniformly over the day to reduce electricity consumption during on-peak hours. In this configuration, the individual building controller does not share any information with the other buildings. To avoid a sudden shifted peak that could lead to higher cost, both charging and discharging operations are uniform.

3.4. Key performance indicator design

Due to the multi-objective nature of the problem, the optimal control strategy needs to optimise multiple objectives, finding a trade-off between all. Several KPIs [56], shown in Table 4, are used to quantify the performance of the controller, considering: an economic KPI (*Cost*), grid-interaction KPIs (*Peak, Peak-to-average ratio (PAR), Daily Peak and Daily PAR*) and flexibility KPIs (*Flexibility Factor, Self-sufficiency*). To analyse the effects of the proposed control strategies on a daily basis, this study calculates and investigates Peak and PAR during the entire simulation period and at a daily granularity, to emphasise building interaction with the grid. Furthermore, the self-sufficiency indicator, defined as the ratio between self-consumption and total consumption, is introduced to quantify the impact of the control strategy on renewable electricity integration. Lastly, the flexibility factor, defined as the ratio between off-peak imported electricity consumption and total imported electricity consumption, is used to analyse the amount of electricity consumed during each tariff period. The mathematical definition of these KPIs is provided in Table 4.

4. Design of multi-agent reinforcement learning control strategies

This section describes the design of the two DRL architectures, denoted as coordinated (centralised) and cooperative (distributed) approaches. Section 4.1 provides a description of the state-space, Section 4.2 outlines the action-space and finally Section 4.3 details the reward functions utilised by each approach. Together, these characterise the MARL approach utilised.

Fig. 4 shows the framework of the two proposed DRL architectures. The image on the left describes the coordinated architecture. System level information is shared with the control level, which coordinates actions using all the information available for the cluster of buildings, with the aim of finding the optimal coordination. On the other hand, the cooperative management (image on the right) exploits multiple controllers, that share only common information such as weather forecast, grid information or district total electrical load, to find the best policy for each building.

4.1. Design of action-space

The case study considers the problem of optimising a cluster of buildings composed of prosumers, by acting on the charging and discharging processes of the thermal storage in the buildings. More specifically, the control actions are related to chilled water storage, that can be charged with a heat pump and discharged to meet building cooling demand, and a SHW storage, that can be charged by an electric heater. Therefore, each building has two control actions and, depending on the type of architecture considered, the number of controller actions is two (cooperative RL) or eight (coordinated RL).

For each control time step (with a resolution of one hour), the DRL agent selects actions between [−1,1], where −1 represents a complete discharge of the storage system and 1 represents a complete charge. The action-space is then constrained between [-1/3,1/3], to facilitate realistic charging and discharging time, according to [30]. The action-space is represented by a tuple of eight values for the coordinated controller and four tuples of two values for each of the four cooperative controllers.

4.2. Design of state-space

The agents learn the optimal control policy, observing the effects of its actions on the environment states. Therefore, the definition of the state-space, together with the reward function, is crucial to help the learning process of the controller and represents one of the points of differentiation between the two architectures. The variables selected by both architectures are reported in Table 5 with further commentary below.

The variables can be categorised into weather, district and building states. Both architectures use weather and district variables, while the main difference is related to the building variables. In particular, the coordinated architecture has access to information for all buildings, e.g., by collecting the State of Charge (SoC) of the eight storage devices, exploiting the information to optimally control the buildings. On the other hand, the cooperative architecture exploits only the information

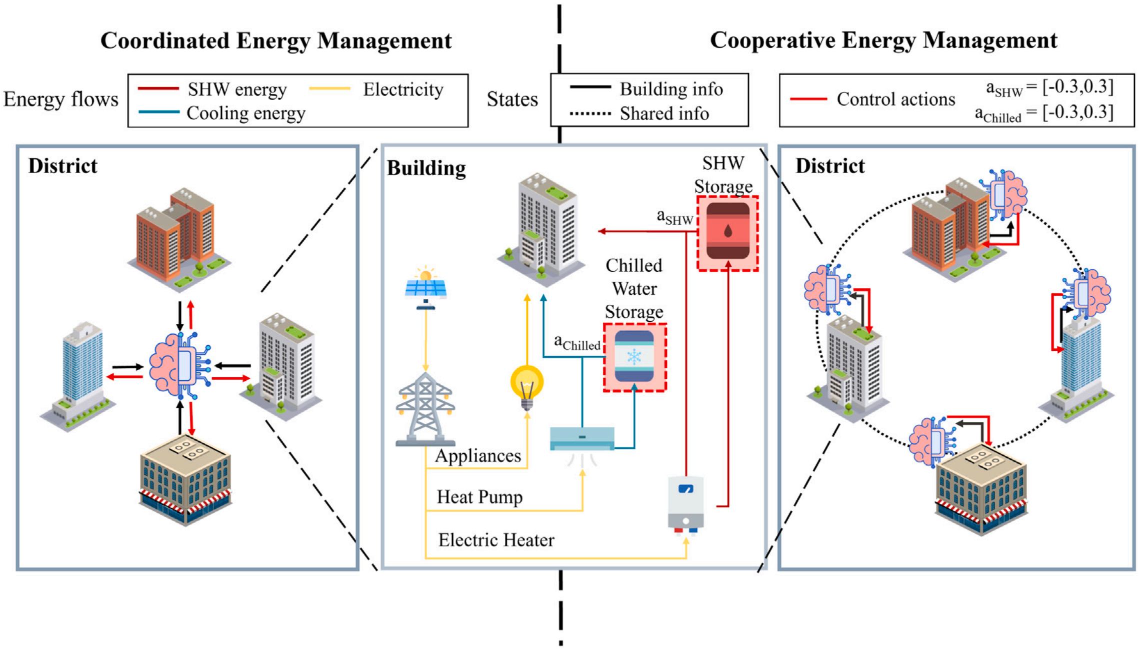

Fig. 4. Coordinated and cooperative control architectures.

Table 5
State-space description for coordinated and cooperative DRL agents.

Variable	Unit
Weather	
Outdoor air temperature	[°C]
Outdoor air temperature forecast (1, 2, 6 hr ahead)	[°C]
Direct solar radiation	[W/m^2]
Direct solar radiation forecast (1, 2, 6 hr ahead)	[W/m^2]
Diffuse solar radiation	[W/m^2]
District	
District total load	[kW]
Electricity price (c_{el})	[\$/kWh]
Electricity price forecast (1, 2 hr ahead)	[\$/kWh]
Hour of day	[h]
Building	
Non-shiftable load	[kW]
Solar generation	[kW]
Chilled water storage SoC (state of charge)	[-]
SHW storage SoC	[-]

related to the controlled building, being unaware of the information from other buildings.

Weather variables, such as outdoor air temperature, direct and diffuse solar radiation, were included to account for their influence on the cooling load. For outdoor air temperature and direct solar radiation, both short (1 and 2 h ahead) and medium (6 h ahead) term forecasts were used to exploit the potential predictive capabilities of the controllers. CityLearn considers the weather as estimated with a generic model with a pre-calculated prediction error. In detail, the prediction error increases with the forecast time-horizon for both temperature and solar radiation. The errors start from 2.5% for 6 h ahead predictions and they increase up to 10% for 24 h ahead. Therefore, the PV generation is evaluated considering a prediction model of the solar radiation with a pre-determined accuracy.

Common variables amongst the buildings were included in district states, such as hour of day, electricity price and electricity price forecasts, with a time horizon of 1 and 2 h ahead, together with the district total electrical load. The states involving information of the specific energy system were categorised as building variables, such as the appliance electrical load (non-shiftable load), the PV electricity production, and the SOC of the cooling and SHW storage devices. As previously explained, for the coordinated architecture, the centralised controller collects these four variables for each building, together with district and weather variables, to find the control strategy. The cooperative architecture, however, exploits only the four states of the controlled building.

4.3. Design of reward functions

The reward function must be representative of the defined control problem and it assesses the effectiveness of the control policy. In this work, comparable reward functions were defined for the coordinated (Section 4.3.1) and cooperative 4.3.2 RL controllers, to benchmark their respective performance. Reward function definition is indirectly related with the previously defined KPIs. The KPIs were defined according to the objective functions that the controller had to achieve. However, the results of the training process are only affected by the cumulative values of the reward function, and not by the evolution of the single KPIs. In fact, KPIs were evaluated to assess the performance of the control policy in a post-processing phase, after the reward (which includes different contrasting objectives) reached convergence.

4.3.1. Reward function for coordinated RL controller

For the coordinated DRL controller, the reward (R) was formulated as a linear combination of three different contributions: the profile flattening term, a cost term and an overproduction term. This is defined as follows:

$$R = \sum_{i=1}^{n} {e_i}^2 \times k_1 + \sum_{i=1}^{n} \left|\min(e_i, 0)\right| \times c_{el} \times k_2 + \sum_{i=1}^{n} \left|\max(e_i, 0)\right| \times c_{sell} \times k_3 \quad (4)$$

The formulation of the flattening term employs a square factor, that leads to a more homogeneous consumption of the cluster [45]. On the other hand, the second and third terms are related to the electricity used/produced from the cluster, with the final goal of reducing operative costs. In particular, e_i is negative if the building imports electricity from the grid and positive if the building sells electricity to the grid. For a specific building, these two terms are mutually exclusive, as the last term assumes electricity overproduction, whereas the second term assumes electricity import from the grid. This formulation is used to reduce electricity costs for the buildings (second term) and to increase

Table 6
Reward function hyperparameter values.

Variable	Coordinated controller	Cooperative controller
k_1	-10^{-5}	-10^{-4}
k_2	−5	−5
k_3	−350	−350

self-consumption in buildings (third term), penalising it when selling electricity rather than trying to exploit renewable overproduction. Considering that in the SAC framework the magnitude of the reward has effects on the results, the terms k1, k2 and k3 were defined to maximise the reward, while balancing the flattening term and the economic results. Therefore, these terms were varied, to achieve optimal trade-off between performance at single building and district scale. The values chosen for the coordinated DRL coefficients (k_1, k_2, k_3) are reported in Table 6.

4.3.2. *Reward function for cooperative RL controller*

To allow a fair comparison among the two architectures, the reward of the cooperative DRL controller was formulated as for the coordinated case, using a linear combination of three terms related to the profile flattening, the imported electricity and the self-consumption. The general formulation of the reward (R_i) for building i is as follows:

$$R_i = \sum_{i=1}^{n} e_i^2 \times k_1 + \left|\min(e_i, 0)\right| \times c_{el} \times k_2 + \left|\max(e_i, 0)\right| \times c_{sell} \times k_3 \qquad (5)$$

The main difference with respect to the previous architecture (coordinated RL controller) is related to the self-consumption and cost terms. While the profile flattening term is similar, the imported electricity and self-consumption terms consider only the controlled building, with the same aim of Eq. (4) previously described. For example, Building 2 and 4 will never experience the overproduction term, due to the lower capacity of PV panels. The values of the three coefficients (k_1, k_2, k_3) are reported in Table 6. It is important to notice that the k_1 term for the two architectures are different. This is due to the fact that in the coordinated approach, the entire electricity consumption of the district is squared, while for the cooperative approach the electricity consumption of each building is first squared and then summed up for all the buildings. Analysing the two quantities previously described (average district power squared and the sum of squared power of each building) a suitable value of k1 for each architecture was set.

As mentioned in Section 2, DRL algorithms are characterised by several hyperparameters, that directly influence controller performance. These parameters need to be tuned according to the specific control problem and they can be further divided into RL hyperparameters and control problem related hyperparameters. To obtain a fair benchmark among the two controllers, RL hyperparameters (decay rate, temperature coefficient, learning rate) were subjected to a hyperparameter optimisation, the results of which are reported in Appendix A to promote the reproducibility of the analysis. To perform hyperparameter optimisation, a grid-search process was used, exploring the search space completely. However, prior to that, domain expertise knowledge and previous experiences were used to constrain the possible search space. Moreover, control problem hyperparameters include the episode length, the starting period of learning and the training episodes, on which a specific analysis was performed. Fig. A.11, reported in Appendix A, shows the evolution of the reward function with the number of episodes. To account for stochasticity the mean and standard deviation of 15 simulations were used. It is possible to notice that after the environment initialisation and around 3 episodes the reward function stabilises. Furthermore, as the number of episodes grows, the standard deviation of the coordinated architecture tends to increase, while the standard deviation of the cooperative architecture remains stable. Therefore, the analysis of the results suggested that a trade-off between simulation period and variance can be found at around 5 episodes, selected for the work. The controllers were then tested over 3 months (an episode) using the three climates described in Table 1.

5. Results

This section describes and analyses the results obtained from the implementation of the two DRL architectures, comparing them with the benchmark RBC strategy. Section 5.1 describes the results of the deployment of both controllers for climate zone 2A (Table 1). More specifically, the financial cost accruing to single users and to the district are described, highlighting how a part of the total cost is related to the district peak, and how the different architectures influences the latter. Following this, the attention is shifted towards the district load, with a focus on storage operation and self-consumption, quantifying the results of cooperative and coordinated approaches at this level. Moreover, the section compares the different use of energy under the control strategies and quantifies the advantages based on several KPIs. Lastly, Section 5.2 presents a summary of the results for deployment for other climates than that outlined in (Table 1) based on the same KPIs.

5.1. *Comparison with baseline RBC*

Fig. 5 shows the energy consumption costs for each building (left) and the energy consumption and peak load (penalty) costs for the district (right). Results are presented for the 3 months simulation period. As it can be seen, both coordinated and cooperative RL result in a lower cost at the district level, namely 3% and 7% savings, respectively. However, the main difference between the two approaches is related to single building costs. For the coordinated approach, Building 2 and 4 experience a cost increase in comparison to RBC strategy of 4% and 3%, compensated by the reduction of the peak term. On the other hand, the cooperative architecture shows a cost reduction for each building, leading to greater overall savings at the district scale.

To analyse further the basis for these results, Fig. 6 shows the district electrical load evolution with the three control strategies for a three day period during the first week of June. This figure highlights the contribution of both PV self-consumption and PV export. Fig. 6 (a) shows how the uncoordinated RBC approach leads to demand peaks during the night due to the charging of the storage devices during these periods, while discharging them during the day, exporting the overproduction of renewable electricity around 12 p.m., June 1. On the other hand, Fig. 6 (b) shows the coordinated approach, tries to exploit PV production as well as flattening the load profile. Lastly, Fig. 6 (c) displays the cooperative approach, in which buildings try to reduce peak consumption, as at around 6 a.m., June 1, and maximise self-consumption, which can be attributed to a reduction of electricity export of Building 1 and Building 3 around 12 p.m., June 1.

To understand the differences between the two proposed control strategy and the baseline, a detailed comparison is provided for Building 1 in Fig. 7 for a three day period. The plotted variables are normalised with respect to their maximum values and include: the state of charge (SOC) for the cooling and SHW storage, the solar radiation, the outdoor temperature and the electricity price. These variables have been selected to highlight the behaviour of an optimal control strategy. Indeed, the best control policy for a prosumer aims at maximising self-consumption, exploiting the lower electricity price and resulting in the minimum district peak demand. To achieve such objectives, both coordinated and cooperative controller shift the charge between the two storage (TES (cooling) and SHW) devices, thereby flattening building electrical load. In particular, they tend to charge the chilled water storage during the night, exploiting the lower ambient temperatures (higher COP) and the SHW TES during the day, to use possible PV over-production. The two storage devices are discharged

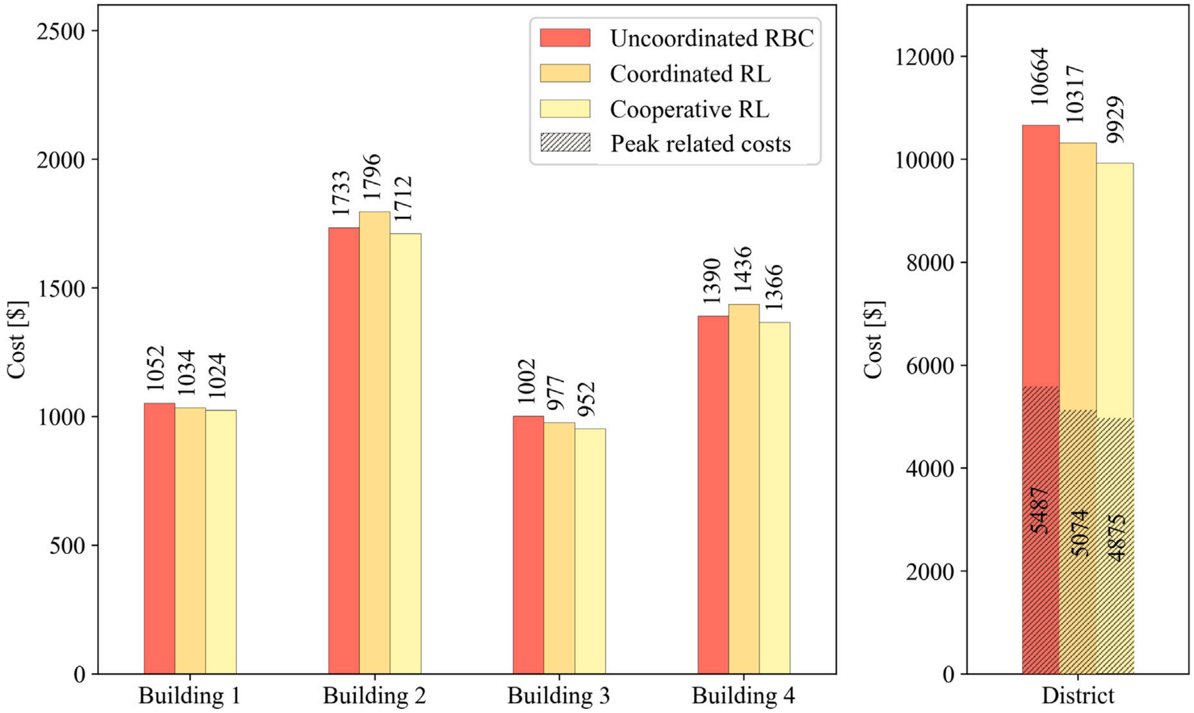

Fig. 5. Cost related to the energy term for each building (left) and total district cost, sum of energy and peak terms (right), for the different control strategies over the entire simulation period.

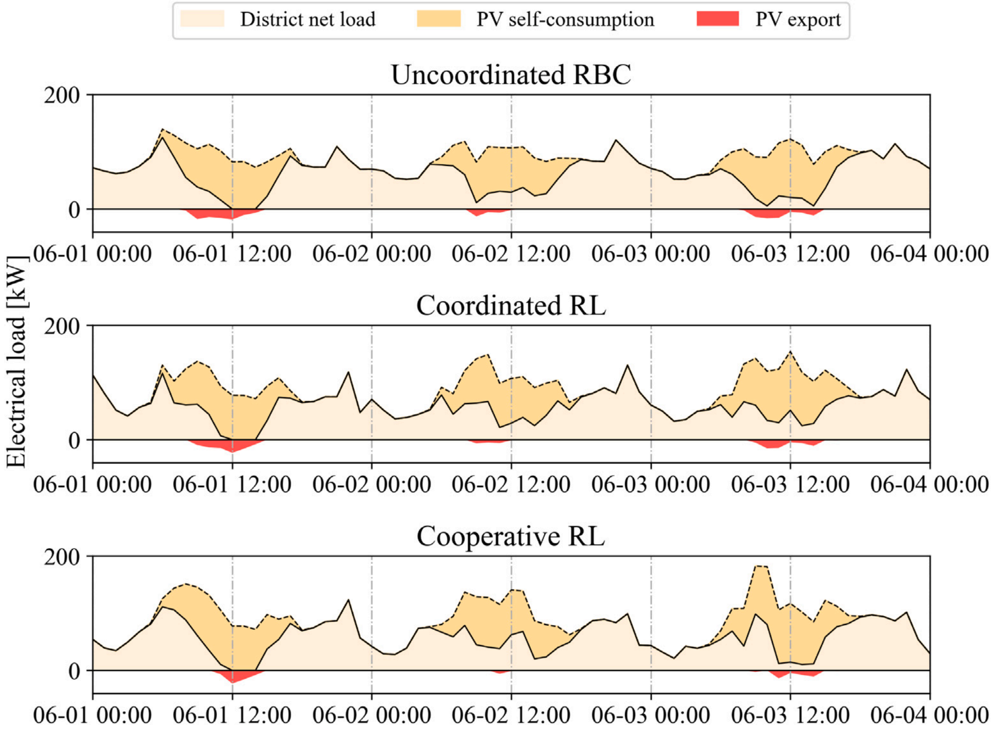

Fig. 6. District electrical load profile for each control strategy during a three-days period.

during high-electricity price periods and low period of PV production, to obtain a flatter profile.

Furthermore, to assess the ability of the controller to adapt to weather conditions and grid requirements, the mean values and standard deviations of SOC for storage devices for a single day period, averaged over the entire season have been showed in Fig. 8. It can be noticed how both RL controllers have higher standard deviation for TES state of charge compared to RBC, explained noticing the variability of SHW and weather conditions, that strongly affect cooling demand. It is important to highlight that the optimal control strategy should not be searched looking at mean values, since the control actions of a specific building depends on: weather conditions, electricity price, building load and grid requirements, in turn influenced by other building control

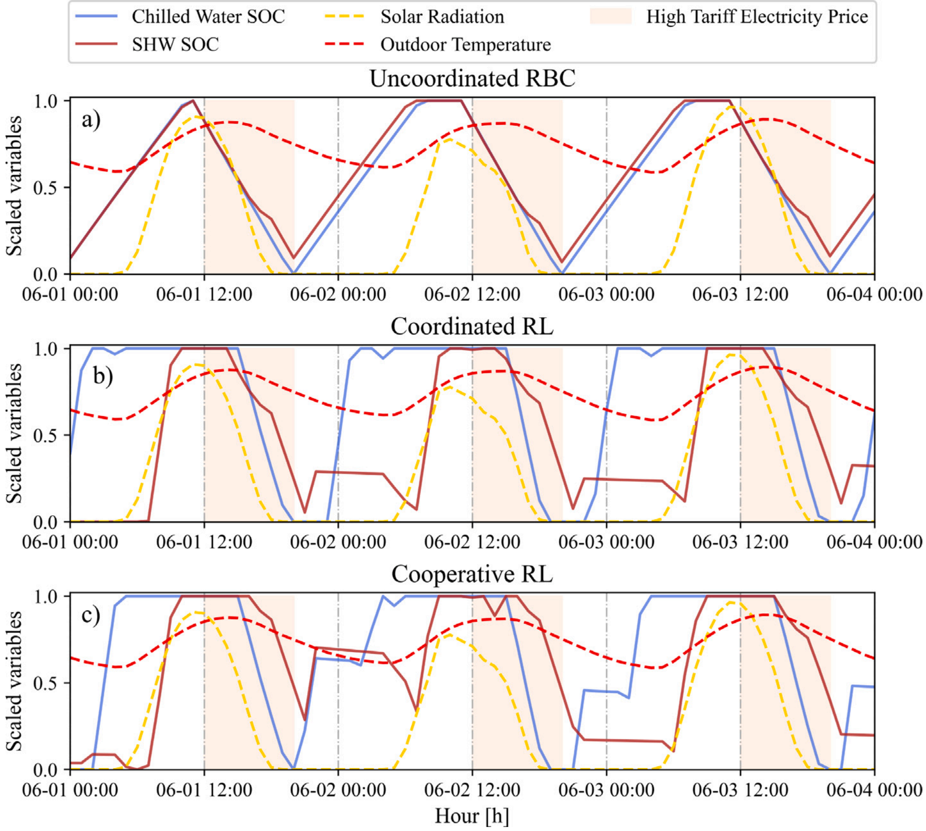

Fig. 7. Comparison of control strategies for Building 1.

actions. However, Fig. 8 can be used to understand how much optimal control actions can be influenced by external factors.

Fig. 9 reports on the evolution of exported electricity at district level for the two controllers and the baseline over the entire simulation period (3 months). Although the absolute exported quantities only represent a small percentage of the total district consumption, their comparison can provide insights into the effectiveness of the control strategies, since minimisation of exported electricity is one of the most effective ways to reduce costs. Given that one of the objectives of the RL controllers is to minimise exported electricity and considering Fig. 9, it can be observed how the uncoordinated RBC is outperformed by the two proposed control strategies. In particular, the cooperative RL reduces the electricity sold to the grid by approximately a quarter compared to the coordinated controller, consequently increasing savings.

To relate the storage operation with the controller performance benefits, Fig. 10 shows the electricity consumption at a district level for the entire simulation period (3 months) as follows: (i) on-peak periods with direct building consumption; (ii) off-peak periods with direct building consumption; (iii) PV production and associated self-consumption, and; (iv) storage discharge (either from the grid or PV) and used to charge either the cooling and SHW storage. Furthermore, to assess the contribution of storage for the integration of renewable energy sources, the bar plot also includes results considering the absence of storage (No Storage), which results in self-consumption from PV being halved compared to the RBC case. The effectiveness of the RBC strategy is evident by examining the consumption reduction during on-peak periods with respect to the No Storage scenario. Despite slightly increasing the amount of on-peak period electricity consumption, the coordinated controller further increases the advantages with respect to RBC, reducing off-peak electricity optimally using the thermal storage, leading to cost savings. These advantages are even greater for the cooperative controller, which shows a slight increase of self-consumption and the highest storage operation, emphasising the role of storage towards the optimal energy management of the district.

Table 7
Results of the MARL controllers deployed on Climate 2A (performance improvement in brackets).

KPI	Climate 2A		
	RBC	Coordinated	Cooperative
Cost [$]	10663	10311 [-3.3%]	9927 [-6.9%]
Peak [kW]	171	154 [-9.7%]	147 [-13.8%]
Daily-Peak [kW]	123	125 [+2.0%]	109 [-11.2%]
Peak-to-average ratio (PAR) [-]	2.31	2.13 [-7.7%]	2.05 [-11.2%]
Daily-PAR [-]	1.66	1.72 [+4.2%]	1.51 [-8.5%]
Self-sufficiency [%]	0.240	0.243 [+1.6%]	0.248 [+3.5%]
Flexibility Factor (FF) [%]	0.66	0.62 [-5.7%]	0.64 [-2.0%]

Lastly, in order to analyse the performance of the three controllers, Table 7 summarises the values assumed by the KPIs to assess the benefits provided by the two RL architectures for the entire simulation period (3 months). The table also shows different KPIs for the RBC, used as a benchmark, while displaying the same KPIs for the coordinated and cooperative architectures with relative improvement (or worsening) in square brackets. For all but the last two KPIs, a lower value indicates a better control policy. Therefore, is it clear that the cooperative architecture outperforms the coordinated architecture especially for the daily-Peak and daily-PAR, where the coordinated controller performs worse than RBC. The coordinated controller is able to reduce costs and peaks with respect to the RBC of around 3% and 10%, respectively. Examining the flexibility factor, it can be seen how both RL controllers perform worse than the RBC. However, the flexibility factor KPI was lower for the DRL controllers because of the decreasing use of off-peak tariff energy consumption.

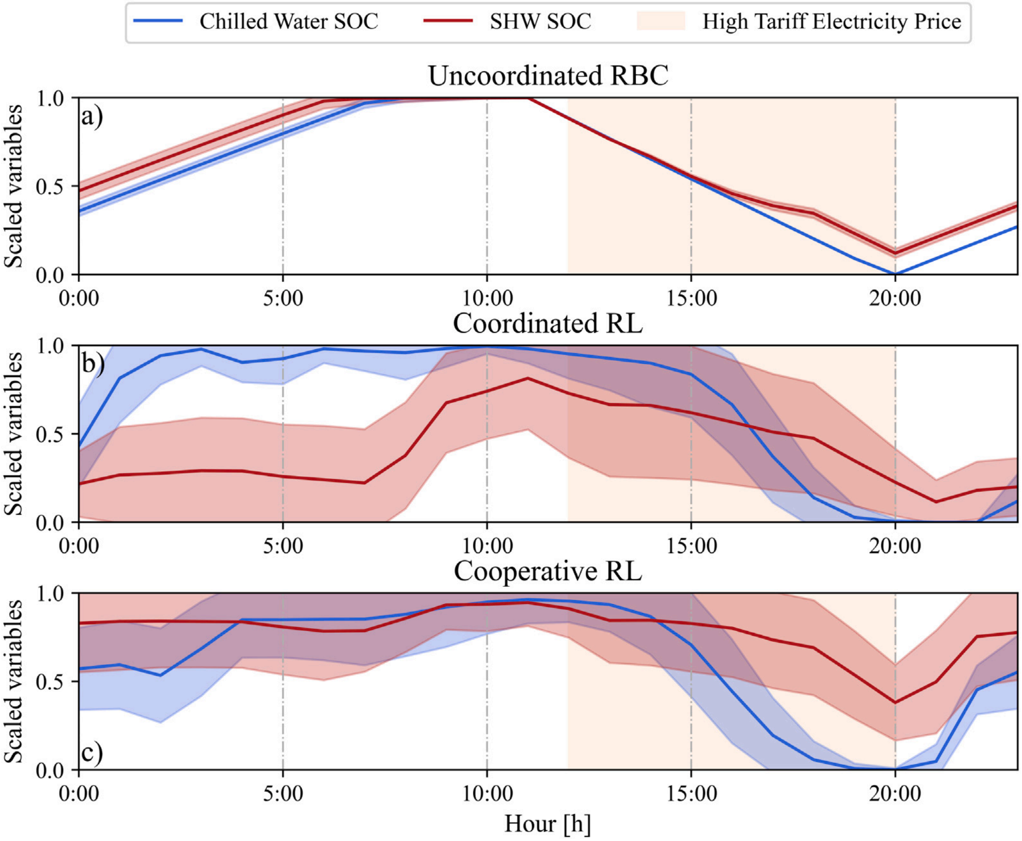

Fig. 8. Daily average hourly scale profiles of SOC with relative standard deviations for the three control strategies in Building 1.

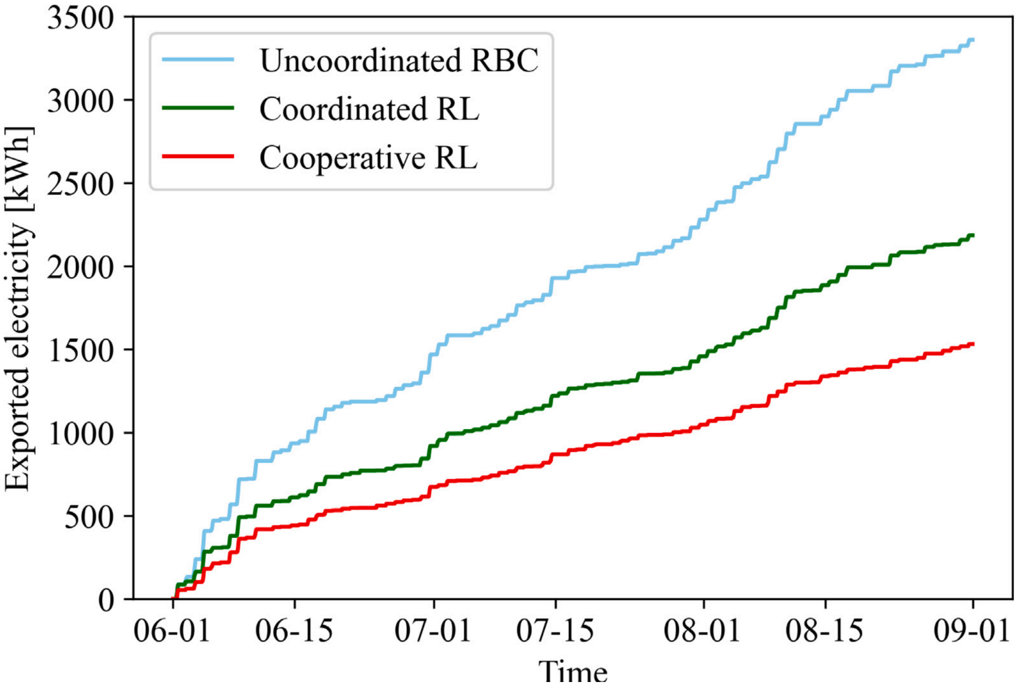

Fig. 9. Comparison of district cumulative exported electricity between control strategies over the entire simulation period (3 months).

5.2. Deployment of RL controllers for different climates

Tables 8 and 9 report the results of the deployment in climates 3A and 5A, comparing the performance of the two RL controllers with respect to RBC. These two climates are characterised by a lower temperature and solar radiation, thus requiring less cooling energy in the summer period, as highlighted by the lower costs. The analysis of the three tables presented has the role to study the robustness of the controllers, here highlighted by the use of KPIs at a different time horizon (Peak, Daily-Peak) as well as examining the adaptability of

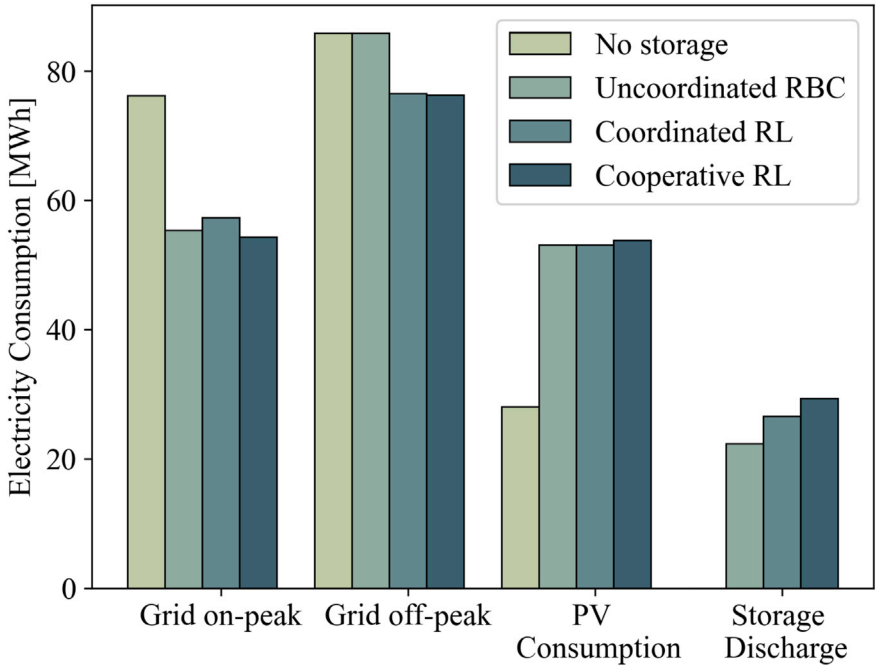

Fig. 10. District energy disaggregation comparison over the entire simulation period (3 months).

Table 8
Results of the MARL controllers deployed on Climate 3A (performance improvement in brackets).

KPI	Climate 3A		
	RBC	Coordinated	Cooperative
Cost [$]	10258	10237 [-0.2%]	9806 [-4.4%]
Peak [kW]	179	174 [-2.4%]	156 [-12.5%]
Daily-Peak [kW]	121	117 [-2.6%]	106 [-11.7%]
Peak-to-average ratio (PAR) [-]	2.61	2.6 [-0.3%]	2.34 [-10.1%]
Daily-PAR [-]	1.77	1.76 [-0.5%]	1.60 [-9.4%]
Self-sufficiency [%]	0.250	0.255 [+2.2%]	0.258 [+3.5%]
Flexibility Factor (FF) [%]	0.65	0.616 [-5.2%]	0.623 [4.1%]

Table 9
Results of the MARL controllers deployed on Climate 5A (performance improvement in brackets).

KPI	Climate 5A		
	RBC	Coordinated	Cooperative
Cost [$]	8946	8856 [-1%]	8874 [-0.8%]
Peak [kW]	150	145 [-3.3%]	145 [-2.7%]
Daily-Peak [kW]	111	98 [-11.7%]	99 [-10.6%]
Peak-to-average ratio (PAR) [-]	2.42	2.4 [-0.6%]	2.42 [+0.2%]
Daily-PAR [-]	1.8	1.63 [-9.3%]	1.65 [-7.9%]
Self-sufficiency [%]	0.270	0.275 [+2.1%]	0.275 [+1.9%]
Flexibility Factor (FF) [%]	0.69	0.645 [-6.4%]	0.649 [-5.9%]

the architectures to different climates. It can be observed that the cooperative approach exhibits better performance in climate zones 2A and 3A, achieving significant advantages (7% and 4%, respectively) in terms of economic costs, while the coordinated architecture performs slightly better in climate zone 5A. This results can be explained noticing that climate 5A is characterised by lower external temperature and solar radiation, that strongly reduce the need of cooling energy, limiting the flexibility provided by the chilled water storage and of the RL control strategy. In general, both architectures achieve better performance with respect to RBC, despite their effectiveness is greatly influenced by climatic conditions. In particular, the main drivers of the problem, district costs and peak, are similar to RBC values shifting the controller from climate 2A to climate 5A, while the controller retains substantial improvement for daily-values, highlighting its stability.

6. Discussion

Grid-interactive buildings can exploit energy flexibility to increase the energy efficiency of each individual building and provide advantages to the grid, with a key role in the energy transition. This research aims to exploit different DRL architectures to enhance the energy grid-interaction for a district of buildings. The DRL controllers were designed to act on building active thermal storage systems, with the aim to exploit energy flexibility, minimising the energy cost for both the individual buildings and the entire district. Moreover, the problem involved time-varying electricity tariffs, including a peak-related term, to incentivise a rational use of electricity amongst the different buildings and to favour cooperation and coordination. To assess the performances of the two DRL control architectures, an uncoordinated RBC was introduced as a baseline, due to widespread use of this strategy for thermal storage control and to provide a fair comparison between the different RL architectures. The same information (state-space) was provided to each controller (with the only exception of information specifically related to that architecture). Moreover, the reward function formulation was also conceived with the same objectives, reducing imported electricity and demand peaks and increasing self-consumption.

The control problem was formulated allowing the DRL controllers to exploit forecast information about electricity price and weather for searching the optimal policy. However, despite SAC DRL use of historical data to speed up the training process, the interaction between different buildings, requires a simulation environment for the training of the controllers. Some key observations for the application and scalability of DRL controllers are related to their computational cost and robustness. Considering that the coordinated architecture scales exponentially with the number of buildings, while the cooperative architecture scales linearly, a cooperative architecture may represent the best solution, but as the number of buildings increases, the non-stationarity of the environment can influence the stability of the cooperative control policy. The present work tried to reduce some of the variability associated with DRL controllers performing hyperparameter optimisation, adopting a similar reward function and studying the evolution of the cumulative reward with episodes. However, as highlighted by Fig. A.11, the inherent stochasticity of the coordinated architecture is higher with respect to cooperative architecture. After the training period, the two controllers achieved superior performance compared to the RBC and took advantage of their predictive nature to flatten the load profile, reducing maximum peak and consequently cost. Table 7 demonstrates the advantages of the cooperative controller over the coordinated controller, particularly when considering daily peaks (11% reduction of the cooperative controller compared to a 2% increase of the coordinated controller) and the reduction of energy costs (7% reduction compared to 3%). Moreover, the two RL controllers differ due to PV self-consumption, which is slightly higher for the cooperative controller.

The reward function formulation plays a crucial role to achieve specific objectives, therefore trade-offs between different terms should be carefully considered. In this perspective, the cooperative architecture is more flexible to the formulation of the reward function, designed to represent user needs in particular, while the coordinated architecture should be defined to achieve high-level performance, averaging over single building requirements. For the specific application considered in this paper, cooperative controller proved to perform better since it was able to search a better control policy oriented to the maximisation of self-consumption. On the other hand, in a coordinated architecture, the results obtained for Building 2 and 4 suggest that, despite reducing district costs, some users may experience increased costs, discouraging them from participating in this type of control. Based on this result, we concluded that in heterogeneous context, with different energy systems and users' needs, a cooperative architecture can be more flexible. Furthermore, the work highlighted that despite the relation among the reward function with some of the KPIs, the multi-objective nature of the problem and the different scales analysed makes important to analyse KPIs in addition to the cumulative reward. Indeed, looking only at reward function as performance indicator, the analysis could lack information about the costs faced by individual buildings, as in the case of the coordinated controller. To test the robustness of the learned optimal policy for both architectures, the controllers were deployed in two other climates. Tables 8 and 9 highlight that, despite both controllers performing better than the RBC, the deployment conditions can highly affect maximum peak and PAR, while they do not influence daily controller performances on average (Daily-Peak and Daily-PAR).

6.1. Limitations

A key concern about the comparison between the architectures is whether the conclusions drawn from the current case study can be generalised. For instance, it should be noticed that for Climate 5A, the performance of the coordinated controller is marginally better that of the cooperative controller, not allowing to identify a superior alternative among cooperative and coordinated architectures. Furthermore, the comparison between the two architectures is influenced by the hyperparameter settings, the number of training episodes, the formulation of the reward function, the inherent stochasticity of DRL and the case study itself. As a consequence the findings cannot be considered generalised and thus need further investigations. The study had the aim to produce a fair comparison among the architectures, using the same hyperparameters, except for number of neurons related to the state–action space. Moreover, also the reward function was conceived to have the same structure, despite the different information the controller exploits. Lastly, the aim of the work was to analyse the effect of the two control strategies for the buildings in the district and the district itself. The computational comparison of the two algorithms was beyond the aim of the paper and may represent a limitation that will be addressed by the authors in a future work. However, the influence of the number of buildings on the computational cost and the effectiveness of the control strategies is important to be assessed especially when different architectures are compared.

7. Conclusions and future work

The present paper considered the design and application of two different DRL controller architectures, with the aim to compare them with the uncoordinated RBC. The problem was formulated to minimise both energy consumption cost and energy demand peak for a district, while trying to increase renewable electricity self-consumption, using a similar reward functions to benchmark the performances of the two DRL controllers.

Relative to the baseline, the two architectures, coordinated and cooperative, showed a cost reduction of 3% and 7%, respectively, together with a peak reduction of 10% and 14%, respectively. Moreover, both controllers achieved an increase in self-sufficiency and a reduction, over the 3 deployment climates of 10% and 9% for daily-peak and daily-PAR, respectively, demonstrating the robustness of the learnt policies.

In conclusion, both architectures outperformed the uncoordinated RBC, representing a potential alternative approach for grid-interactive district energy management. The research highlighted that, if buildings have heterogeneous objective functions, the multi-agent architecture can capture the preferences of single buildings. This architecture is therefore suitable when dealing with multiple buildings where different energy systems are present or when the operational preferences are specific for each building, such as thermal comfort conditions or renewable electricity over production.

Future work will focus on:

- The application of the presented methodology on a district featuring a larger number of buildings and also considering a model predictive control strategy. In particular, the computational cost at different scale of analysis will be compared to assess the advantages and limitations of each controllers, together with a performance comparison.
- The implementation of a peer-to-peer architecture within a tailored interactive urban environment (e.g., CityLearn). Such an architecture provides additional benefits in the presence of renewable overproduction, enhancing the advantages of cooperation (or coordination) among buildings, thereby allowing a more comprehensive analysis to be carried out.

CRediT authorship contribution statement

Giuseppe Pinto: Conceptualization, Methodology, Software, Investigation, Writing – original draft, Visualization. **Anjukan Kathirgamanathan:** Methodology, Software, Investigation, Formal analysis, Visualization, Writing – original draft. **Eleni Mangina:** Conceptualization, Writing – review & editing, Supervision. **Donal P. Finn:** Conceptualization, Writing – review & editing, Supervision. **Alfonso Capozzoli:** Conceptualization, Methodology, Writing – review & editing, Supervision.

Declaration of competing interest

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

Acknowledgements

Anjukan Kathirgamanathan, Eleni Mangina and Donal Finn gratefully acknowledge that their contribution emanated from research supported by Science Foundation Ireland under the SFI Strategic Partnership Programme Grant Number SFI/15/SPP/E3125.

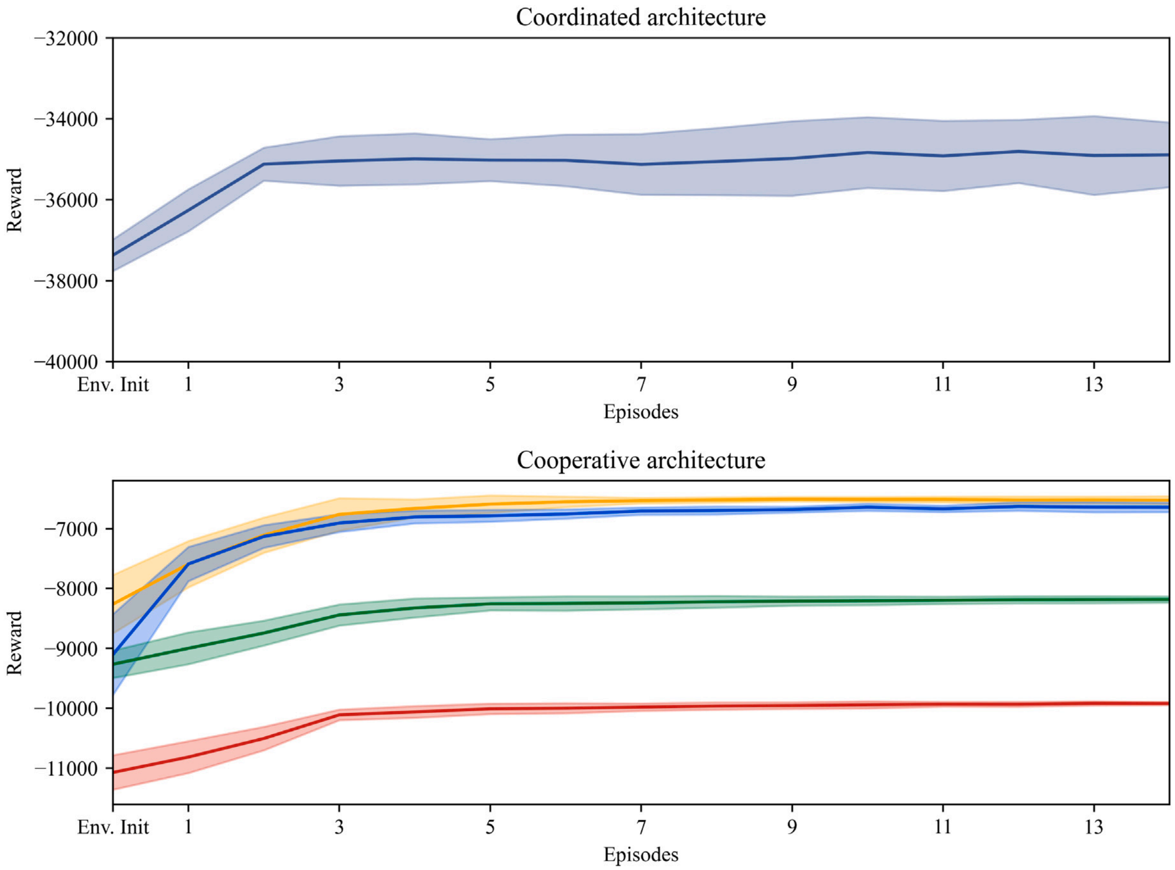

Fig. A.11. Evolution of the reward function with episodes.

Table A.10
Settings of the DRL hyperparameters for coordinated and cooperative architectures.

Hyperparameter	Coordinated controller	Cooperative controller	Search Space
DNN architecture	4 Layers (2 hidden)	4 Layers (2 hidden)	–
Neurons per hidden layer	256	64	[64,128,256]
DNN Optimiser	Adam	Adam	–
Batch size	512	512	–
Learning rate (λ)	0.001	0.001	[0.001,0.005,0.01]
Discount rate (γ)	0.99	0.99	[0.9,0.95,0.99]
Decay rate (τ)	0.005	0.005	[0.001,0.005,0.01]
Temperature coefficient (α)	0.05	0.05	[0.01,0.05,0.1]

Table A.11
Settings of the control problem hyperparameters for coordinated and cooperative architectures.

Hyperparameter	Coordinated controller	Cooperative controller
Learning starts	2208	2208
Target model update	1	1
Episode Length	2208 Control steps	2208 Control steps
Training Episodes	5	5

Appendix A. Deep reinforcement learning hyperparameters

Table A.10 list the SAC hyperparameters for the two architectures, along with the optimisation space analysed. Table A.11 shows the hyperparameter of the control problems, along with the final configuration selected to perform the analysis, while Fig. A.11 displays the evolution of the reward function with the number of episodes.

References

[1] European Commission. The European green deal. In: Communication from the commission to the european parliament, the european council, the council, the european economic and social committee and the committee of the regions. 2019, p. 24.

[2] Victoria M, Zhu K, Brown T, Andresen GB, Greiner M. Early decarbonisation of the European energy system pays off. Nature Commun 2020;11(1):1–9.

[3] Lund PD, Lindgren J, Mikkola J, Salpakari J. Review of energy system flexibility measures to enable high levels of variable renewable electricity. Renew Sustain Energy Rev 2015;45:785–807.

[4] Villar J, Bessa R, Matos M. Flexibility products and markets: Literature review. Electr Power Syst Res 2018;154:329–40.

[5] Cochran J, Miller M, Zinaman O, Milligan M, Arent D, Palmintier B, et al. Flexibility in 21st century power systems. Technical report, 21st Century Power Partnership; 2014, p. 14.

[6] Warren P. A review of demand-side management policy in the UK. Renew Sustain Energy Rev 2014;29:941–51.

[7] Klein K, Herkel S, Henning H-M, Felsmann C. Load shifting using the heating and cooling system of an office building: Quantitative potential evaluation for different flexibility and storage options. Appl Energy 2017;203:917–37.

[8] Economidou M, Laustsen J, Ruyssevelt P, Staniaszek D. Europe's buildings under the microscope. Technical report, Buildings Performance Institute Europe (BPIE); 2011, p. 130.

[9] 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(3).

[10] Sun Y, Wang S, Xiao F, Gao D. Peak load shifting control using different cold thermal energy storage facilities in commercial buildings: A review. Energy Convers Manag 2013;71:101–14.

[11] Hao H, Middelkoop T, Barooah P, Meyn S. How demand response from commercial buildings will provide the regulation needs of the grid. In: 2012 50th annual allerton conference on communication, control, and computing, allerton 2012. 2012, p. 1908–13.

[12] Kathirgamanathan A, De Rosa M, Mangina E, Finn DP. Data-driven predictive control for unlocking building energy flexibility: A review. Renew Sustain Energy Rev 2020;135(January 2021):110120.

[13] 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(2020):110248.

[14] Vázquez-Canteli JR, Nagy Z. Reinforcement learning for demand response: A review of algorithms and modeling techniques. Appl Energy 2019;235(2018):1072–89.

[15] Foo. Eddy YS, Gooi HB, Chen SX. Multi-agent system for distributed management of microgrids. IEEE Trans Power Syst 2015;30(1):24–34.

[16] Santos G, Pinto T, Morais H, Sousa TM, Pereira IF, Fernandes R, et al. Multi-agent simulation of competitive electricity markets: Autonomous systems cooperation for European market modeling. Energy Convers Manag 2015;99:387–99.

[17] Khan MW, Wang J, Ma M, Xiong L, Li P, Wu F. Optimal energy management and control aspects of distributed microgrid using multi-agent systems. Sustain Cities Soc 2019;44:855–70.

[18] Karavas C-S, Kyriakarakos G, Arvanitis KG, Papadakis G. A multi-agent decentralized energy management system based on distributed intelligence for the design and control of autonomous polygeneration microgrids. Energy Convers Manag 2015;103:166–79.

[19] Mohamed MA, Jin T, Su W. Multi-agent energy management of smart islands using primal-dual method of multipliers. Energy 2020;208:118306.

[20] Li C, Jia X, Zhou Y, Li X. A microgrids energy management model based on multi-agent system using adaptive weight and chaotic search particle swarm optimization considering demand response. J Clean Prod 2020;262:121247.

[21] Jin S, Wang S, Fang F. Game theoretical analysis on capacity configuration for microgrid based on multi-agent system. Int J Electr Power Energy Syst 2021;125:106485.

[22] Qiu D, Ye Y, Papadaskalopoulos D, Strbac G. Scalable coordinated management of peer-to-peer energy trading: A multi-cluster deep reinforcement learning approach. Appl Energy 2021;292:116940.

[23] Samadi E, Badri A, Ebrahimpour R. Decentralized multi-agent based energy management of microgrid using reinforcement learning. Int J Electr Power Energy Syst 2020;122:106211.

[24] 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.

[25] Pinto G, Deltetto D, Capozzoli A. Data-driven district energy management with surrogate models and deep reinforcement learning. Appl Energy 2021;304:117642.

[26] Xiong L, Li P, Wang Z, Wang J. Multi-agent based multi objective renewable energy management for diversified community power consumers. Appl Energy 2020;259:114140, URL https://www.sciencedirect.com/science/article/pii/S0306261919318276.

[27] Labeodan T, Aduda K, Boxem G, Zeiler W. On the application of multi-agent systems in buildings for improved building operations, performance and smart grid interaction – a survey. Renew Sustain Energy Rev 2015;50:1405–14, URL https://www.sciencedirect.com/science/article/pii/S1364032115005638.

[28] Etedadi Aliabadi F, Agbossou K, Kelouwani S, Henao N, Hosseini SS. Coordination of smart home energy management systems in neighborhood areas: A systematic review. IEEE Access 2021;9:36417–43.

[29] Sutton AG, Richard S. Barton, reinforcement learning: an introduction. 2nd ed.. Cambridge, Massachusetts: MIT Press; 2014.

[30] Henze GP, Schoenmann J. Evaluation of reinforcement learning control for thermal energy storage systems. HVAC R Res 2003;9(3):259–75.

[31] Liu S, Henze GP. Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory: Part 1. Theoretical foundation. Energy Build 2006;38(2):142–7.

[32] Liu S, Henze GP. Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory: Part 2: Results and analysis. Energy Build 2006;38(2):148–61.

[33] Dalamagkidis K, Kolokotsa D, Kalaitzakis K, Stavrakakis G. Reinforcement learning for energy conservation and comfort in buildings. Build Environ 2007;42(7):2686–98.

[34] Mason K, Grijalva S. A review of reinforcement learning for autonomous building energy management. Comput Electr Eng 2019;78:300–12.

[35] Azuatalam D, Lee W-L, de Nijs F, Liebman A. Reinforcement learning for whole-building HVAC control and demand response. Energy AI 2020;2:100020.

[36] Wang Z, Hong T. Reinforcement learning for building controls: The problem, opportunities and challenges. Appl Energy 2020;269(1):300.

[37] Kazmi H, Suykens J, Balint A, Driesen J. Multi-agent reinforcement learning for modeling and control of thermostatically controlled loads. Appl Energy 2019;238:1022–35.

[38] Nagarathinam S, Menon V, Vasan A, Sivasubramaniam A. MARCO - multi-agent reinforcement learning based control of building HVAC systems. In: Proceedings of the eleventh ACM international conference on future energy systems. e-Energy ’20, New York, NY, USA: Association for Computing Machinery; 2020, p. 57–67.

[39] Kofinas P, Dounis A, Vouros G. Fuzzy Q-learning for multi-agent decentralized energy management in microgrids. Appl Energy 2018;219:53–67.

[40] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning. In: 4th international conference on learning representations. 2016.

[41] Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: 35th international conference on machine learning, Vol. 5. 2018, p. 2976–89.

[42] Zhang B, Hu W, Cao D, Li T, Zhang Z, Chen Z, et al. Soft actor-critic –based multi-objective optimized energy conversion and management strategy for integrated energy systems with renewable energy. Energy Convers Manag 2021;243:114381.

[43] Coraci D, Brandi S, Piscitelli MS, Capozzoli A. Online implementation of a soft actor-critic agent to enhance indoor temperature control and energy efficiency in buildings. Energies 2021;14(4).

[44] Biemann M, Scheller F, Liu X, Huang L. Experimental evaluation of model-free reinforcement learning algorithms for continuous HVAC control. Appl Energy 2021;298:117164.

[45] Pinto G, Piscitelli MS, Vázquez-Canteli JR, Nagy Z, Capozzoli A. Coordinated energy management for a cluster of buildings through deep reinforcement learning. Energy 2021;229:120725.

[46] Vázquez-Canteli JR, Kämpf J, Henze G, Nagy Z. CityLearn V1.0: An openAI gym environment for demand response with deep reinforcement learning. In: BuildSys 2019 - proceedings of the 6 ACM international conference on systems for energy-efficient buildings, cities, and transportation. 2019, p. 356–7.

[47] Deltetto D, Coraci D, Pinto G, Piscitelli MS, Capozzoli A. Exploring the potentialities of deep reinforcement learning for incentive-based demand response in a cluster of small commercial buildings. Energies 2021;14(10).

[48] Dhamankar G, Vazquez-Canteli JR, Nagy Z. Benchmarking multi-agent deep reinforcement learning algorithms on a building energy demand coordination task. In: RLEM 2020 - Proceedings of the 1st international workshop on reinforcement learning for energy management in buildings and cities. 2020, p. 15–9.

[49] Tuyls K, Weiss G. Multiagent learning: Basics, challenges, and prospects. AI Mag 2012;33(3):41–52.

[50] Wong A, Bäck T, Kononova AV, Plaat A. Multiagent deep reinforcement learning: Challenges and directions towards human-like approaches. 2021, ArXiv, arXiv:2106.15691.

[51] Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, et al. Automatic differentiation in PyTorch. 2017.

[52] Kathirgamanathan A, Mangina E, Finn DP. Development of a soft actor critic deep reinforcement learning approach to a virtual large office building for Harnessing energy flexibility. Energy AI (Under Rev.) 2021;1–32.

[53] 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.

[54] Vazquez-Canteli JR, Dey S, Henze G, Nagy Z. CityLearn: Standardizing research in multi-agent reinforcement learning for demand response and urban energy management. 2020.

[55] Entergy. Entergy electricity tariff. 2020, Available at https://cdn.entergy-texas.com/userfiles/content/price/tariffs/eti_gs-tod.pdf.

[56] Clauß J, Finck C, Vogler-finck P, Beagon P. Control strategies for building energy systems to unlock demand side flexibility – a review. In: Proc. Of BS2017: 15th conference of international building performance simulation association, San Fransisco, USA, Aug 7-9. San Fransisco; 2017.