Model-Free Deep Reinforcement Learning for Adaptive Supply Temperature Control in Collective Space Heating Systems

SARA GHANE, University of Antwerp - imec, IDLab - Faculty of Applied Engineering, Belgium
STEF JACOBS, EMIB, Faculty of Applied Engineering - Electromechanics, University of Antwerp, Belgium
THOMAS HUYBRECHTS, University of Antwerp - imec, IDLab - Faculty of Applied Engineering, Belgium
PETER HELLINCKX, M4S, Faculty of Applied Engineering - Electronics ICT, University of Antwerp, Belgium
SIEGFRIED MERCELIS, University of Antwerp - imec, IDLab - Faculty of Applied Engineering, Belgium
IVAN VERHAERT, EMIB, Faculty of Applied Engineering - Electromechanics, University of Antwerp, Belgium
ERIK MANNENS, University of Antwerp - imec, IDLab - Department of Computer Science, Belgium

The conventional approach for controlling the supply temperature in collective space heating networks relies on a predefined heating curve determined by outdoor temperature and heat emitter type. This prioritizes thermal comfort but lacks energetic and financial optimization. This research proposes an adaptive supply temperature control in well-insulated dwellings, responsive to diverse environmental parameters. The approach considers variable electricity prices and accommodates different indoor temperature set points in dwellings. The study evaluates the effectiveness of two Deep Reinforcement Learning (DRL) algorithms, i.e. Proximal Policy Optimization (PPO) and Deep Q-Network (DQN), across various scenarios. Results reveal that DQN excels in collective space heating systems with underfloor heating in each dwelling, while PPO proves superior for radiator-based systems. Both outperform the traditional heating curve, achieving up to 13.77% (DQN) and 16.15% (PPO) cost reduction while guaranteeing thermal comfort. Additionally, the research highlights the capability of DRL-based methods to dynamically set the supply temperature based on a cloud of set points, showcasing adaptability to diverse environmental factors and addressing the growing significance of indoor heat gains in well-insulated dwellings. This innovative approach holds promise for more efficient and environmentally conscious heating strategies within collective space heating networks.

CCS Concepts: • **Computing methodologies** → **Control methods**.

Additional Key Words and Phrases: Reinforcement learning, Energy saving, Thermal comfort, Collective space heating system, Variable electricity price, PPO, DQN

1 INTRODUCTION

The utilization of fossil fuels exacerbates Greenhouse Gas (GHG) emissions, air pollution, and resource depletion. In the European Union (EU), the building stock accounts for 40% of final energy use and contributes 36% of the total GHG emissions [45]. This necessitates a concerted effort to adopt energy-efficient and environmental-friendly heating solutions in buildings [5, 9, 10, 45].

Authors' addresses: Sara Ghane, University of Antwerp - imec, IDLab - Faculty of Applied Engineering, Sint-Pietersvliet 7, Antwerp, 2000, Belgium, Sara.Ghane@uantwerpen.be; Stef Jacobs, EMIB, Faculty of Applied Engineering - Electromechanics, University of Antwerp, Groenenborgerlaan 171, Antwerp, 2020, Belgium, Stef.Jacobs@uantwerpen.be; Thomas Huybrechts, University of Antwerp - imec, IDLab - Faculty of Applied Engineering, Sint-Pietersvliet 7, Antwerp, 2000, Belgium, Thomas.Huybrechts@uantwerpen.be; Peter Hellinckx, M4S, Faculty of Applied Engineering - Electronics ICT, University of Antwerp, Groenenborgerlaan 171, Antwerp, 2020, Belgium, Peter.Hellinckx@uantwerpen.be; Siegfried Mercelis, University of Antwerp - imec, IDLab - Faculty of Applied Engineering, Sint-Pietersvliet 7, Antwerp, 2000, Belgium, Siegfried.Mercelis@uantwerpen.be; Ivan Verhaert, EMIB, Faculty of Applied Engineering - Electromechanics, University of Antwerp, Groenenborgerlaan 171, Antwerp, 2020, Belgium, Ivan.Verhaert@uantwerpen.be; Erik Mannens, University of Antwerp - imec, IDLab - Department of Computer Science, Sint-Pietersvliet 7, Antwerp, 2000, Belgium, Erik.Mannens@uantwerpen.be.

In this respect, collective space heating systems offer a sustainable solution by facilitating the integration of Renewable Energy Sources (RES) into thermal energy supply [11, 27, 29, 57]. These systems connect different dwellings through distribution pipelines to shared production units, providing the heat for Space Heating (SH). With respect to heat production, Heat Pumps (HP), as a type of RES, play a vital role in achieving the EU's decarbonization objectives and the aspiration to combat energy poverty by lowering energy bills [23].

However, optimal control in these networks remains challenging and becomes increasingly important and complex as the complexity of the collective space heating systems rises. Traditional supply temperature control methods, such as heating curves or fixed set points ($T_{sup,SP}$), are sub-optimal for two key reasons. Firstly, HP-based heating systems operate more efficiently at low supply temperatures, especially in new buildings with low-temperature emitters and high insulation rates. In such buildings, the traditional heating curve tends to overestimate heat demand at low outdoor temperatures because of factors such as solar gains and internal heat gains that are not considered by the control approach [46]. Consequently, the heating curve should be depicted as a multi-dimensional cloud of $T_{sup,SP}$ points rather than a single trajectory curve. This representation can accommodate variations in multiple parameters such as outdoor temperature and electricity prices, offering a broader range of possibilities compared to a linear representation. This implies that more variables with more complex dynamics should be considered for determining $T_{sup,SP}$ at all times. Secondly, the traditional heating curve overlooks variable electricity prices, which is the trend in countries like Belgium striving for decarbonization and electrification of thermal energy supply. Therefore, control strategies need to adapt to future expected electricity prices to minimise Operational Expenses (OPEX), while ensuring thermal comfort. As can be noted, the flexibility in controlling this $T_{sup,SP}$ depends on intermittent variables, necessitating more complex control strategies to maximize the economic and environmental benefits of collective space heating systems while ensuring indoor thermal comfort.

In this respect, Model Predictive Control (MPC) and Reinforcement Learning (RL) are emerging control methods to solve such control problems. MPC is recognized for its robustness and sample efficiency [3], while RL offers adaptability and the ability to handle uncertainties [3, 37]. Moreover, RL has lower computational complexity at runtime. Recent advancements in Deep Reinforcement Learning (DRL) have made it viable for integration into complex systems such as Heating, Ventilation, and Air Conditioning (HVAC) and collective space heating systems. The DRL methods enable control policies (i.e., strategies) to be learned without requiring detailed knowledge of the system's model. Thus, it does not rely on the complex white-box models anymore once it is trained, hence it is fast in decision-making when it is deployed. Therefore, DRL is used for the control task of this study.

1.1 State of the art

In recent years, several studies have applied (D)RL-based control in thermal networks, specifically in the domain of heating, and have demonstrated promising results in terms of energy savings and maintaining comfort which are reviewed extensively in [2, 17, 30]. These studies have leveraged experiences acquired from the environment to train DRL algorithms. Huang et al. [18] used RL for optimal $T_{sup,SP}$ control in the heating network of an office building. The proposed controller was based on Q-learning method that uses a Q-table of state-action values. The state space contained parameters of both current and future estimates. With a reward function based on indoor comfort scores, they found that when benchmarked against standard Rule-based Controllers (RBCs), RL effectively prevents overheating. In another study by Zhang et al. [56], a framework was developed to train RL agents using the Asynchronous Advantage Actor Critic (A3C) algorithm for controlling an office room's heating system connected to a district heating system. By controlling the mixing

valve's set point, their proposed method reduced the heating demand by 16.7% compared to RBC. Le-Coz et al. [24] proposed a Deep Q-Network (DQN)-based approach for controlling the $T_{sup,SP}$ of a district heating system. Their approach utilized a Recurrent Neural Network (RNN) trained on simulated data to predict indoor temperatures and return temperature. Then, two DQN agents were trained with a single-objective reward function penalizing deviations from a target temperature. One received expert guidance; the other did not. The study demonstrated the potential of RL to achieve its objective, by improving the energy efficiency of district heating systems while providing thermal comfort compared to a baseline strategy. Claessens et al. [4] controlled multiple thermostatically controlled loads in district heating using multi-agent RL, achieving performance improvement.

Furthermore, several other related works have explored the application of RL techniques in various HVAC and energy management scenarios. Ye et al. [55] formulated the energy management problem for a residential Multi-Energy System (MES) as a Markov Decision Process (MDP) and applied a model-free RL approach to provide an optimal energy management strategy, leading to more cost-effective control strategies. Gupta et al. [15] proposed a DQN-based controller that regulated temperature by controlling the on/off status of the thermostat to maintain thermal comfort and optimize energy savings. The controller, trained and verified in simulation, reduced energy consumption and improved thermal comfort over a traditional thermostat. Gao et al. [13] developed a Deep Deterministic Policy Gradient (DDPG)-based framework for controlling the set points for indoor temperature and humidity of the HVAC system in smart buildings, enhancing thermal comfort while reducing HVAC energy use. Du et al. [7] developed a DDPG-based control strategy for the set point control of each room zone of a multi-zone residential HVAC system, saving costs and energy while preserving thermal comfort. For HP operation mode control, Lissa et al. [28] utilized an RL agent which manages the Space Heating (SH) and Domestic Hot Water (DHW) systems, optimizing photovoltaic self-consumption and achieving energy savings. Pinto et al. [36] used a centralized model-free RL controller to manage thermal storages in four commercial buildings, cutting costs and peak demand. Ruelens et al. [38] implemented a model-free RL approach based on fitted Q-iteration for HP set point control in HVAC systems, saving energy over conventional constant set-point strategies. Nagy et al. [34] employed a variant of fitted Q-iteration for SH control by controlling the input power of the HP, outperforming an RBC in energy savings and computation times. Han et al. [16] proposed utilizing Rainbow DQN for operation strategy of multiple HPs, resulting in electricity cost reduction by optimizing demand-charge and energy-charge.

A major limitation in existing research on DRL algorithms for thermal control systems is the dependence on well-known DRL algorithms, such as DDPG and DQN (from the Q-learning family), which have been mainly validated in gaming contexts. Consequently, many studies fail to consider alternative DRL techniques that might be more effective in inherently dynamic environments [1], such as collective heating systems. Exploring and developing these less commonly used approaches, especially policy-based methods such as PPO, and conducting comparative analyses with established methods such as DQN, presents a valuable area for research. This exploration aims to identify and leverage the unique strengths and advantages of each method in the context of complex dynamic environments, such as collective space heating systems.

The key differences between this study and previous ones stand out in several important areas. Previous studies have often overlooked the development of controllers that can effectively manage the dynamic characteristics of collective space heating systems. While some research has focused on individual aspects, such as variable electricity prices or specific heat gains from indoor activities and solar radiation, there is a gap in the literature regarding a comprehensive approach. Specifically, there is a need for a DRL controller that simultaneously addresses these varying factors in well-insulated dwellings to optimize operational cost efficiency and system performance in collective space heating systems. This study addresses these gaps by incorporating these factors into the

real-time decision making to ensure more adaptive and efficient central supply temperature control. Moreover, previous studies have not provided a detailed comparison between off-policy and on-policy DRL methods in the context of supply temperature control in collective heating systems. Specifically, there has been a lack of analysis comparing DQN, a value-based and off-policy method, and PPO, a policy-based and on-policy method. This gap leaves an unclear understanding of the relative effectiveness and applicability of these different DRL approaches for supply temperature control, which our study aims to address.

1.2 Problem statement

As can be noted from subsection 1.1, the main focus of applying (D)RL in collective heating systems has been on providing thermal comfort and energy savings. They often overlook the consideration of changing boundary conditions, such as the necessity of developing a controller capable of adapting to different set points, varying prices and temperature condition, all in conjunction. However, the shift towards HP-based collective space heating systems, whose efficiency depends on supply temperature control, requires a shift in control strategies. Moreover, tariff structures evolving towards variable price settings to encourage demand side management strategies enforce this necessity.

Therefore, this study investigates the applicability and effectiveness of DRL to control the central $T_{sup,SP}$ of a collective space heating system heated by a Geothermal HP (GHP). The focus is on coping with variable boundary conditions, encompassing variable electricity prices and changing indoor temperature set points. The three main contributions of this study can be outlined as follows:

(1) **Adaptive DRL-based controller:** This study presents an adaptive DRL method for optimizing supply temperature in heating systems. Unlike traditional approaches, this method adjusts to various factors, including hourly electricity price fluctuations, outdoor temperature changes, and individual indoor temperature preferences based on occupancy and sleep patterns. The adaptability of this DRL method is enhanced by using a multi-dimensional set of control points instead of a fixed heating curve, improving efficiency and responsiveness. Additionally, the integration of a dynamic tariff structure allows real-time adjustments to operational costs, utilizing forecasted electricity prices for optimal planning. By balancing cost efficiency and thermal comfort, the controller achieves multi-objective optimization, providing significant energy savings and improved occupant comfort. This innovative approach ultimately enhances the performance and sustainability of collective heating systems, offering a more flexible and adaptive solution than conventional methods.

(2) **Design of Markov Decision Process (MDP):** The design of an MDP (including state space, action space, and reward function) is a critical aspect of any Artificial Intelligence (AI) application, and directly influences the effectiveness and efficiency of the learning process. Highlighting this design as a novelty emphasizes that our approach addresses specific challenges to improve performance, and provides the foundation for benchmarking and further research. This research introduces innovative elements in the MDP to enhance the optimization of heating systems. The state space incorporates unique variables such as hourly electricity prices for the next day, thermal comfort levels, and immediate rewards, providing comprehensive information for the agent's decision-making process. The action space specifies supply temperature setpoints for the GHP, enabling precise and adaptive control. The reward function is designed to balance two conflicting objectives: minimizing operational costs and reducing thermal discomfort. By integrating fluctuating electricity prices into the reward function, the approach addresses the trade-off between cost efficiency

and occupant comfort, allowing the agent to optimize operations effectively and enhance overall system performance.

(3) **Comparative analysis of DRL techniques:** The study rigorously compares two Deep Reinforcement Learning (DRL) methods: value-based off-policy (DQN) and policy-based on-policy (PPO). This analysis is crucial for understanding which method better reduces operational costs while maintaining thermal comfort. Current RL-based HVAC systems often rely on popular algorithms such as DQN [1]. However, they overlook potentially more effective methods, mostly policy-based, such as PPO for dynamic environments. By comparing DQN and PPO, this research highlights their unique strengths in controlling supply temperature setpoints for GHP-based systems. The findings provide valuable practical insights for researchers and practitioners aiming to apply DRL in similar complex systems.

1.3 Outline of the paper

The rest of the paper is structured as follows. Section 2 begins with an overview of the principles of DRL. Subsequently, the models employed in our simulation environment to represent the collective space heating system are explained. Section 3 outlines our proposed control approach along with details of the training setup. Additionally, the baseline RBC is provided and the Key Performance Indicators (KPIs) used in this research are described. Moving on to Section 4, the conducted experiments are presented and the obtained results are comprehensively analyzed. Finally, Section 5 provides the concluding remarks of this research.

2 BACKGROUND

In this section, an overview of the principles of DRL is provided, serving as a foundational understanding for subsequent methodology section. Following this, the description of the models utilized within our simulation environment are explained, which represent the collective space heating system under study.

2.1 Deep Reinforcement Learning

Reinforcement learning is a paradigm for solving sequential decision making and control problems. Through the repeated interactions of the autonomous RL agent with the surrounding unknown environment, the agent learns the control policy through trial and error [43, 53]. Applying RL to sequential decision making tasks requires the formal description of the environment as Markov Decision Process (MDP). MDP provides a framework for the agent to learn through interaction with the environment and achieve a goal. As a result of this interactive process, the agent learns how to make an optimal decision. This is done by observing the current state of the environment and performing an action (i.e., a control decision) to maximize a reward [25, 35]. This sequential decision making process is presented more clearly in Figure 1. An MDP is defined as a tuple (S, A, P, R, γ), where:

- a set of states $s \in S$ that can be observed in the environment;
- a set of actions $a \in A$;
- the transition probability between states $P : S \times A \times S$;
- the reward function which maps the state and action to the immediate rewards $R : S \times A$;
- the discount factor $\gamma \in [0, 1]$ for determining how much importance we give to the future rewards.

Defining the elements of the MDP is an important step for providing an optimal control. Any change in the definition of MDP elements could lead to a different RL solution, i.e., different control mechanisms.

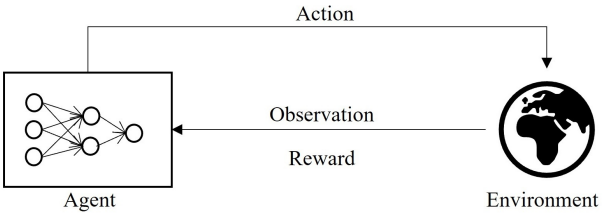


Fig. 1. Sequential decision making with DRL.

The solution of an MDP is a policy π that represents the behavior of an agent and maps the states to actions. Using the policy enables the agent to take an action in each state. The goal of RL is to maximize the agent's reward and reach the optimal policy. However, solely prioritizing immediate rewards leads to a greedy policy, restricting exploration of the environment. The greedy policy constantly chooses the action that is believed to achieve the highest expected reward. Therefore, in an environment with continuing tasks, the agent aims to maximize the expected sum of discounted future rewards. This is included in the state-value function of a state, which represents the performance of a policy in a given state. It is defined as the expected future discounted rewards achieved by the agent [31]:

$$V^{\pi}(s) = \mathbb{E}[\sum_{i=0}^{\infty} \gamma^i r(s_i)|\pi, s_0 = s] \tag{1}$$

Similarly, the action-value function for the policy π is defined as the expected future discounted rewards achieved by the agent that starts from state s, takes action a, and thereafter follows π:

$$Q^{\pi}(s,a) = \mathbb{E}[\sum_{i=0}^{\infty} \gamma^i r(s_i,a_i)|\pi, s_0 = s, a_0 = a] \tag{2}$$

The state-value function can be rewritten in terms of the action-value function:

$$V^{\pi}(s) = max_a Q^{\pi}(s,a) \tag{3}$$

Moreover, an advantage function ($A^{\pi}(s,a)$) is also commonly used to measure the benefit of selecting a particular action a over the average action of the policy at a given state s. $A^{\pi}(s,a)$ is defined in Equation 4.

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s) \tag{4}$$

Additionally, the estimated advantage function, denoted as $\hat{A}$, can be used in the context of Generalized Advantage Estimation (GAE). This function helps in reducing the variance of policy gradient estimates by combining multiple-step returns.

Finding the optimal policy π^* that in each state provides an optimal action, is the goal of an RL agent [19]. The optimal policy can be derived by:

$$\pi^* = argmax_{\pi} V^{\pi}(s) = argmax_{\pi} Q^{\pi}(s,a) \tag{5}$$

To achieve the optimal policy, there are two main RL approaches, namely model-based and model-free. In case of dealing with an environment for which the transition probabilities and the reward function are known, model-based RL is typically used. However, the exact characteristics of the environment in most real-world problems are unknown. Therefore, the agent learns the optimal policy only through the interaction with the unknown environment, without having any prior knowledge of the environment and its dynamics. However, RL encounters challenges with large state and action spaces in real-world applications, making value computation for each state impractical.

Deep Reinforcement Learning (DRL) addresses this by integrating Neural Networks (NNs) to approximate the value function, policy, and/or model. In this study, DQN [32] and PPO [39] are used which are the model-free DRL algorithms.

2.1.1 Deep Q-network (DQN). DQN is an off-policy value-based method that was proposed to overcome the dimension problem of Q-learning [32, 33] that is caused by using a lookup table for large state-action pairs. Being a value-based method, DQN aims to learn the value function and infer the optimal policy by selecting actions that maximize the value function. As an off-policy method, it learns a policy while using a possibly different behavior policy to interact with the environment [22].

DQN uses a NN with parameter θ, i.e., deep Q-network $Q(s, a; \theta)$, to approximate the state-value function in a Q-Learning framework. In DQN, the observed state of the environment (s) is given as an input to the NN, and the Q-values of all actions (a) are produced as outputs which will be used to make the decision. DQN uses the concept of "Experience Replay" to store past experiences in a replay buffer and randomly sample from them during training to avoid the problem of correlated data, which can lead to unstable learning and inefficient convergence.

The goal is to minimize the difference between predicted Q-values and target Q-values ($\widehat{Q}$) with the following loss function at iteration i:

$$L_i(\theta) = \mathbb{E}\left[\left(r_i + \gamma max_{a'} \widehat{Q}(s_{i+1}, a'; \theta^-) - Q(s_i, a_i; \theta)\right)^2\right] \tag{6}$$

where θ and θ^- are the parameters of the Q-network and target $\widehat{Q}$-network, respectively. While powerful, DQN suffers from overestimation bias, instability, and convergence difficulty.

Double DQN [47] addresses overestimation bias by decoupling action selection from Q-value evaluation. The target for Double DQN is calculated as follows:

$$y_i^{DoubleDQN} = r_i + \gamma Q(s_{i+1}, argmax_{a'} Q(s_{i+1}, a'; \theta); \theta^-) \tag{7}$$

To better estimate the value of each state independently of the action taken, thus improving learning efficiency, Dueling DQN is proposed [54] which separately estimates the state-value function $V(s_i; \theta, \nu)$ and the advantage function $\hat{A}(s_i, a_i; \theta, \kappa)$. The Q-value is then calculated as:

$$Q(s_i, a_i; \theta, \kappa, \nu) = V(s_i; \theta, \nu) + \left(\hat{A}(s_i, a_i; \theta, \kappa) - \frac{1}{|\hat{A}|} \sum_{a'} \hat{A}(s_i, a'; \theta, \kappa)\right) \tag{8}$$

where κ and ν are parameters of the advantage and value functions in the Dueling network, respectively.

In this study, an improved version of DQN which is called Dueling Double DQN [54]. This method effectively addresses the aforementioned issues by integrating Double DQN and Dueling DQN, which results in a more accurate and efficient method. The target in Dueling Double DQN is given by:

$$\begin{aligned} y_i^{\text{Dueling Double DQN}} = r_i + \gamma \Bigg[& V(s_{i+1}; \theta^-, \nu^-) + \Big(\hat{A}(s_{i+1}, \arg\max_{a'} Q(s_{i+1}, a'; \theta, \kappa); \theta^-, \kappa^-) \\ & - \frac{1}{|\hat{A}|} \sum_{a'} \hat{A}(s_{i+1}, a'; \theta^-, \kappa^-)\Big)\Bigg] \end{aligned} \tag{9}$$

Then, the loss function is as follows:

$$L_i(\theta) = \mathbb{E}\left[\left(y_i^{DuelingDoubleDQN} - Q(s_i, a_i; \theta, \kappa, \nu)\right)^2\right] \tag{10}$$

We refer to "Dueling Double DQN" as DQN in this paper for the sake of readability and to represent it as part of the broader DQN family as off-policy value-based methods.

2.1.2 Proximal Policy Optimization (PPO). PPO [39] is a well-known model-free DRL algorithm which is from the on-policy policy gradient family. It aims to achieve a balance between sampling complexity, and ease of tuning hyperparameters and implementation. This is done by computing a new policy that minimizes the objective function while having only a slight deviation from the original policy to ensure the stability. Being a policy-based method, it aims to learn the optimal policy directly. As an on-policy method, it learns a policy by using it for selecting actions, followed by iterative evaluation and enhancement of the policy. PPO learns a policy by iteratively performing multiple mini-batch optimizations over a set of trajectories collected from the environment, and stored in a "Trajectory Memory". These trajectories, which include state, action, reward, and next state information, are used to compute the advantage function and optimize the policy.

There are two main variants of PPO: (1) PPO with Kullback–Leibler (KL) Divergence and (2) PPO with Clipping. They both aim to improve the stability and reliability of policy updates.

(1) **PPO with KL Divergence** involves constraining the policy updates by penalizing large deviations from the old policy. Here, the loss function combines the estimated advantage function ($\hat{A}$) with the KL penalty to ensure that updates do not change the policy too drastically. Given the probability ratio of $r_i(\theta) = \frac{\pi_\theta(a_i|s_i)}{\pi_{\theta_{old}}(a_i|s_i)}$, the KL-penalized objective ($L_i^{KL}(\theta)$) is as follows:

$$L_i^{KL}(\theta) = \mathbb{E}\Big[r_i(\theta)\hat{A}_i - c_{KL}KL[\pi_{\theta_{old}}(.|s_i), \pi_\theta(.|s_i)]\Big] \tag{11}$$

where c_{KL} is a coefficient to control the influence of KL divergence penalty.

(2) **PPO with Clipping** modifies the objective function to include a clipping mechanism, which directly limits how much the new policy can deviate from the old one. Given ϵ as a hyperparameter that controls the clipping range, the objective function ($L_i^{CLIP}(\theta)$) is as follows:

$$L_i^{CLIP}(\theta) = \mathbb{E}\Big[min\Big(r_i(\theta)\hat{A}_i, clip(r_i(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i\Big)\Big] \tag{12}$$

When using NNs for approximating value function and policy function, the loss function is redefined by combining the $L_i^{CLIP}(\theta)$ with a value function error term ($L_i^{VF}(\theta) = (V_\theta(s_i) - V_i^{target})^2$). The updated objective is formulated as follows:

$$L_i^{CLIP+VF+\mathbb{S}}(\theta) = \mathbb{E}\Big[L_i^{CLIP}(\theta) - c_{VF}L_i^{VF}(\theta) + c_{ent}\mathbb{S}[\pi_\theta](s_i)\Big] \tag{13}$$

where $\mathbb{S}$ represents the entropy bonus to promote sufficient exploration, and c_{VF} and c_{ent} are coefficients for value function loss and entropy term, respectively.

In this study, to ensure stable and effective learning, we used a combined PPO loss function ($L_i^{PPO}(\theta)$) which is presented in Equation 14.

$$L_i^{PPO}(\theta) = \mathbb{E}\Big[L_i^{CLIP}(\theta) - c_{VF}L_i^{VF}(\theta) + c_{ent}\mathbb{S}[\pi_\theta](s_i) - c_{KL}KL[\pi_{\theta_{old}}(.|s_i), \pi_\theta(.|s_i)]\Big] \tag{14}$$

Here, PPO is employed within an actor-critic framework which integrates a policy network (actor) and a value function network (critic) to optimize the policy. Using actor-critic PPO enables more stable and efficient learning. We refer to "actor-critic PPO" as PPO in this paper for the sake of readability and to represent the on-policy policy-based methods.

2.2 Simulator of the collective space heating system

In this study, a dynamic simulation environment is employed to emulate the thermal dynamics within a collective space heating system operating in a ten-dwelling apartment building. The thermal models, except for the central storage tank, are based on ordinary linear and non-homogeneous

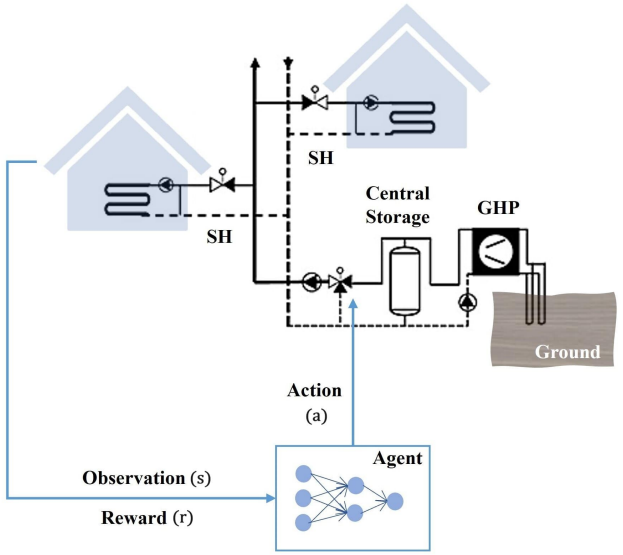


Fig. 2. Overview of the collective space heating system featuring a central GHP connected to a storage tank and ten dwellings where only SH is considered. The SH emitters are either UFH or radiators. In case of radiators, the passive mixing in dwellings is not present. The lower blue section of the image illustrates DRL-based control with either a DQN or a PPO agent.

differential equations of the first order as in [14, 48]. The storage tank is a partial differential equation both in temperature and height. The simulation time step (dt) is 10 seconds. The internal heat gains $\dot{Q}_{internal}$ [W] and occupancy profiles are based on a stochastic profile generator developed in the TETRA-SWW [52] and Install2020 [51] projects with statistical occupancy data of different family types in Belgium.

2.2.1 Overview of considered collective space heating system. The apartment building of this research consists of ten dwellings with identical characteristics in terms of floor area, insulation level and ventilation losses, among others, but with different window orientations and occupancy profiles. These disparities significantly influence their respective heat demands, owing to variations in internal heat gains and expected indoor temperatures. The design heat load, calculated at 21°C indoor and -8°C outdoor, is 2.8 kW for each dwelling, resulting in an overall heat transfer coefficient (UA-value) of 96 W/K. One large GHP, connected to a central storage tank, provides all thermal energy to the collective space heating system. The nominal heating power of the GHP ($\dot{Q}_{(snk,nom)}$) equals the peak heat demand of the apartment building, i.e. 28 kW. The central storage tank is sized to be thermally recharged by the GHP within an hour as was considered optimal according to [48]. The space heating unit inside the dwellings, i.e. the emitter, is either a UFH or a radiator. Figure 2 shows an overview of the case study where all dwellings are equipped with UFH. In this case, the design temperatures are 35°C supply and 30°C return and a passive mixing circuit connects the UFH to the central supply pipe to prevent too high temperatures inside the dwellings. In case of radiators, the design temperatures are 60°C/40°C, for supply and return respectively, and the passive mixing circuits are not present.

2.2.2 Thermal modelling of heat demand. Each dwelling is simulated as three temperature nodes, namely the indoor air temperature (T_{zone} [°C]), with a heat capacity C_{zone} [J/K] related to an indoor

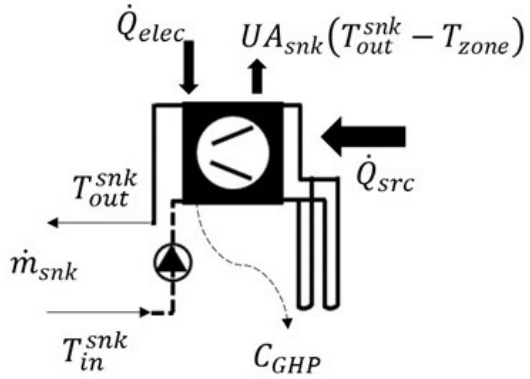


Fig. 3. A schematic overview of the energy balance in the GHP used for modelling its thermal behavior.

air volume of 270 m^3, the temperature of the walls, with a heat capacity of medium-heavy walls [50], and the outdoor temperature (T_{ext} [°C]) based on Belgian weather data [41]. Moreover, two types of heat losses are simulated. First, the transmission losses between both T_{zone} and the wall node, based on an U-value of the inside vertical air layer, and between the wall node and T_{ext}. Second, the ventilation losses of mechanical, balanced ventilation (system D) are considered with a heat recuperation efficiency of 80%. To calculate a ventilation mass flow ($\dot{m}_{vent}$ [kg/s]), a constant ventilation rate of 1 h^{-1} is assumed and an infiltration rate of 4 h^{-1} at 50 Pa according to the *blower door test* [58]. The heat emitter system of each dwelling is simulated as an RC-model with three equal segments, as in [12] and validated in [48]. Each of those segments has a uniform temperature and the outlet of a segment is the inlet temperature of the next segment. The total heat flow from an emitter to the respective thermal zone ($\dot{Q}_{emitter}$) is the sum of the heat flows of the three segments. More information on the used differential equations and more are described in [20].

2.2.3 Model of the central GHP. The central GHP is a grey-box model. First, the dynamic thermal behavior is modelled according to Equation 15, based on the energy balance of Figure 3.

$$\begin{aligned} C_{GHP}\frac{d(T_{out}^{snk})}{dt} &= \dot{Q}_{src} + \dot{Q}_{elec} \\ &\quad - UA_{snk} \cdot (T_{out}^{snk} - T_{zone}) \\ &\quad - c_{p;water} \cdot \dot{m}_{snk} \cdot (T_{out}^{snk} - T_{in}^{snk}) \end{aligned} \tag{15}$$

Where C_{GHP} [J/K] is the GHP's thermal capacity, T_{in}^{snk} and T_{out}^{snk} [°C] are the in- and outgoing temperature, respectively, at the condenser (sink) side of the GHP, T_{zone} is the ambient temperature [°C] of the central production room, UA_{snk} [W/K] is the heat transfer coefficient to surroundings, $c_{p;water}$, the specific heat capacity of water, and $\dot{m}_{snk}$ [kg/s] is the flow rate at condenser side. The energy flows $\dot{Q}_{src}$ and $\dot{Q}_{elec}$ are the heat extracted from the heat source of the GHP and the required electricity, respectively. Second, for a realistic performance map, data of Viessmann reported in [48] is used to define $\dot{Q}_{src}$ and $\dot{Q}_{elec}$ at different source (T_{in}^{src}) and sink temperatures (T_{in}^{snk}). The performance map is scaled to the nominal heating power ($\dot{Q}_{snk,nom}$), which is 28 kW in this research. Equations (16) and (17) represent the fitting of $\dot{Q}_{src}$ and $\dot{Q}_{elec}$ on the performance

map of Viessmann using two two-dimensional second order polynomials.

$$\begin{aligned}\frac{\dot{Q}_{src}}{\dot{Q}_{snk,nom}} &= -58.1 - 313.4 \cdot T_{in}^{src*} + 378.8 \cdot T_{out}^{snk*} \\ &\quad - 31.3 \cdot T_{in}^{src} \cdot T_{out}^{snk*} + 357 \cdot (T_{in}^{src*})^2 \\ &\quad - 297.3 \cdot (T_{out}^{snk*})^2 - 318.6 \cdot T_{out}^{snk*} \cdot (T_{in}^{src*})^2 \\ &\quad + 283.7 \cdot T_{in}^{src*} \cdot (T_{out}^{snk*})^2\end{aligned} \tag{16}$$

$$\begin{aligned}\frac{\dot{Q}_{elec}}{\dot{Q}_{snk,nom}} &= 10.7 - 52.5 \cdot T_{in}^{src*} + 32.2 \cdot T_{out}^{snk*} \\ &\quad + 5 \cdot T_{in}^{src} \cdot T_{out}^{snk*} + 59.3 \cdot (T_{in}^{src*})^2 \\ &\quad - 35.7 \cdot (T_{out}^{snk*})^2 - 51.2 \cdot T_{out}^{snk*} \cdot (T_{in}^{src*})^2 \\ &\quad + 42.4 \cdot T_{in}^{src*} \cdot (T_{out}^{snk*})^2\end{aligned} \tag{17}$$

Where $\dot{Q}_{src}$, $\dot{Q}_{elec}$ and $\dot{Q}_{snk,nom}$ are the same as before, and T_{in}^{src*} and T_{out}^{snk*} are the same temperatures as T_{in}^{snk} and T_{out}^{snk}, but converted to Kelvin and divided by 273 for smaller coefficients.

2.2.4 Other thermal models. The stratified storage tank model consists of a partial differential equation and is based on the type 60 of TRNSYS [40]. The stratified tank is divided in 26 equal water layers, where conduction, advection, heat losses to surroundings and the effects of water density at different temperatures are considered. The control signals of valves and pumps adjust the flow rate directly, without depending on pressure models and hydraulic dependencies. In this respect, the mass flow between the nominal value and 10% of this value is guaranteed to be available. Time delays of control valves are taken into account with a time constant (τ) of 32s as in [48] and the thermal behavior of mixing is according to the mixing rule [6]. The time delay in the pipes is modelled by applying the plug-flow principle [49] and their thermal losses are characterized by an RC-model.

2.2.5 Day-ahead market tariff structure. The day-ahead market price (DAM) for electricity in Belgium is the price established one day before the actual delivery of electricity. It reflects the hourly balance of supply and demand for the next day, determined by the bids submitted by electricity producers and large consumers. The market operator, EPEX SPOT in Belgium, matches these bids to determine the price for each hour [€/MWh] [42].

Small consumers, such as a building management system of an apartment building, do not directly participate in the bidding process. Instead, they follow the dynamic pricing offered by the electricity suppliers, who base their tariffs on the day-ahead market prices. These prices are typically published around noon (12:00 PM CET) each day for the following day, enabling consumers to optimise their energy usage by scheduling intensive energy usages during low prices.

The day-ahead price influences electricity costs for consumers and plays a crucial role in maintaining grid stability. A lower day-ahead price often indicates an abundance of intermittent renewable energy sources, such as wind or solar power. Therefore, using more electricity during these low-price periods can not only reduce operational costs, but also lower CO_2 emissions.

To not rely on contracts of different electricity suppliers, the simulator uses the published day-ahead market price for 2022 in Belgium, which is publicly available [44].

3 METHODOLOGY

3.1 The proposed control approach

This study consists of a collective space heating system at the building level. Two use cases were considered that have different types of emitters, namely UFH and radiator. Radiator is a faster emitter system compared to the UFH due to smaller thermal inertia and potential for higher water temperatures. To enable DRL-based control, for both UFH and radiator use cases, the simulation environment (as described in section 2.2) is integrated into the OpenAI Gym environment. This facilitates the implementation of the proposed RL solutions by using the RLlib toolbox [26], and learning the control policy through agent-environment interactions. An overview of this integrated simulation environment for the UFH is shown in Figure 2. In case of radiator for SH, the configuration is similar, but without passive mixing valves in the dwellings.

In the proposed control approach, two common different indoor temperature set points (i.e., $18^{\circ}C$ and $20^{\circ}C$) are considered that are imposed by the inhabitants. Notably, the indoor temperature set point ($T_{op;SP}$) varies for each dwelling based on occupancy and sleep patterns, with daytime temperatures set at either $18^{\circ}C$ or $20^{\circ}C$, dropping to $T_{op;SP} - 2^{\circ}C$ for the UFH and $16^{\circ}C$ for radiators during night time or when occupants are away. Taking into account these set points during training, which reflect dynamic user behavior and preferences, along with fluctuations in electricity prices sourced from the day-ahead market (DAM), and outdoor temperature, enhances the adaptability of our proposed approach.

To perform the central $T_{sup,SP}$ control, two different type of DRL agents were trained and tested: a DQN (off-policy value-based) and a PPO (on-policy policy gradient). These choices allow us to determine which approach is more effective: optimizing to obtain a good approximation of the expected rewards in different states (DQN) or optimizing the policy directly (PPO). In the following subsection, we propose an MDP formulation to solve this control problem and answer this question.

3.1.1 MDP formulation. The decision-making problem for the DRL agent has to be formalized and designed using various components of MDP in a way that the dynamics of the environment can be captured. The MDP definitions used in the proposed control methods are explained in the following, where UFH and radiator have the same state space and reward function, but differ in the action space.

State space: The state space is designed using different temperature measurements along with other information. It includes the followings:

- s_1: outdoor temperature
- s_2: average indoor set point temperature
- s_3: central T_{sup}
- s_4: time of the day
- s_5: hourly electricity price of the next day
- s_6: thermal comfort
- s_7: immediate reward

Given that by including s_3 the agent knows the current supply temperature, including s_1 and s_2 enables the agent to know which $T_{sup,SP}$ it needs to choose to provide the thermal comfort. That is why it is necessary for the agent to also receive thermal comfort (s_6, which is calculated using Equation 20) as an extra information. Moreover, by providing s_5, the agent will have access to the electricity price of the next 24 hours, which together with time of the day (s_4) helps it in planning. Finally by observing s_7, which is the immediate reward and is calculated using Equation 18 and indicates the quality of the action taken in that state, the agent would know about its current state of fulfilling the objectives. Including information about the objectives of our control task (s_6 and

s_7) in the state space will further help the agent in making decisions [21]. This novel state space, characterized by its compact size, enables the agent to have a good perception of the environment with only a few states, avoiding the complexity in learning, especially in scenarios with many dwellings.

Action space: The discrete action space represents the GHP supply temperature set points. Each action corresponds to setting the GHP supply temperature to a specific value. Selecting an action means adjusting the GHP supply temperature to one of these values, which allows precise and adaptive control over the collective heating system. The actions are set in a way that the supply temperature is always within a certain limit to assure it does not disturb the functioning of the collective space heating system.

Given the fact that UFH and radiator have different requirements, their action spaces are defined separately. Each type of heating system is experimented with two action spaces:

- Radiator
  - Action_1: This vector represents a series of possible supply temperatures for the GHP that is connected to the dwellings equipped with radiator, ranging from 20°C to 65°C, in increments of 5°C. The specific set points included are [20, 25, 30, 35, 40, 45, 50, 55, 60, 65].
  - Action_2: This vector is similar to Action_1 but excludes the highest set point (65°C). The set points included are [20, 25, 30, 35, 40, 45, 50, 55, 60]. This variation allows us to study the impact of excluding the highest temperature on cost savings and comfort.
- UFH
  - Action_1: This vector represents a series of possible supply temperatures for the GHP that is connected to the dwellings equipped with the UFH, ranging from 20°C to 45°C, in increments of 5°C. The specific set points included are [20, 25, 30, 35, 40, 45].
  - Action_2: This vector is similar to Action_1 but excludes the highest set point (45°C). The set points included are [20, 25, 30, 35, 40]. This variation allows us to examine the impact of excluding the highest temperature on cost savings and comfort.

Besides, it is also investigated how the agent's control interval (10 minutes vs. 15 minutes) impacts its effectiveness.

Reward function: The reward function (Equation 18), which is a critical component of our model, is designed to balance two main conflicting objectives: minimizing operational costs, i.e. OPEX, related to the GHP, and minimizing indoor temperature discomfort for the end users. This dual-objective reward function is novel in its approach, as it directly addresses the trade-offs between cost efficiency and occupant comfort, which are often competing priorities in heating systems.

Another novel aspect of our reward function design is its incorporation of fluctuating electricity prices rather than fixed prices, which adds to the complexity of the learning problem. Using fluctuating electricity prices allows the agent to capitalize on lower prices, further enhancing cost efficiency. The reward function is formulated as a linear combination of these two objectives, where the weight of the reward function is denoted as β which strikes a balance between the two objectives. For every time step i, the OPEX (Equation 19) is defined as the current electricity consumption of the GHP's compressor ($\dot{Q}_{elec,i}$), multiplied by the current day-ahead market price ($DAM_{elec,i}$).

The Belgian day-ahead market (DAM) price of 2022, available from [8], is considered as this year contains lots of variations. The maximum average power consumption of the GHP ($\dot{Q}_{elec,max}$) during the whole considered time frame for reward calculation is 5.6 kW. Similarly, the maximum and minimum cost for electricity in €/kWh during the considered time frame are denoted as

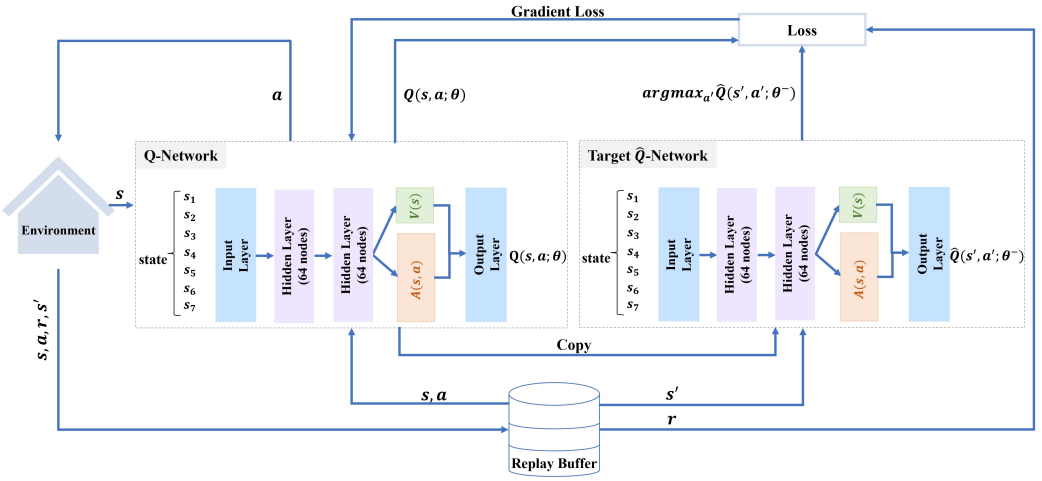


Fig. 4. The supply temperature control method for a collective space heating system using DQN. The *Replay Buffer* is a storage mechanism that collects past experiences, from which samples are drawn to update the networks.

$DAM_{elec,max}$ and $DAM_{elec,min}$, respectively. The indoor temperature comfort is based on the Room Temperature Lack (RTL) of the considered time frame and calculated as in [48]. The RTL^R_{max} equals a temperature deviation of 1°C throughout the whole considered time frame for reward calculation.

$$R_i = \beta \cdot \frac{\dot{Q}_{elec,max} \cdot DAM_{elec,max} - OPEX}{\dot{Q}_{elec,max} \cdot (DAM_{elec,max} - DAM_{elec,min})} + (1-\beta) \cdot \sqrt{1 - \left(\frac{RTL_i}{RTL^R_{max}}\right)^2} \tag{18}$$

$$OPEX = \dot{Q}_{elec,i} \cdot DAM_{elec,i} \tag{19}$$

To identify the optimal value for β, preliminary experiments were conducted using grid search. We tested a range of values for $\beta = (0.1, 0.3, ..., 0.9)$ to find the best balance. Our preliminary results indicated that $\beta = 0.3$ provided the best trade-off.

3.2 Training setup in the collective space heating system

In this paper, two DRL algorithms are employed: PPO (Algorithm 1) and DQN (Algorithm 2). Both algorithms are used to train the supply temperature control strategy for a collective space heating system, which are shown in Figure 4 and Figure 5. These figures illustrate the structure of the proposed training process for both methods, including the inputs to the NNs and the corresponding loss calculation procedures for them.

Each method undergoes training for a total of 100 episodes using the hyperparameters that are listed in Table 1. Grid search using Ray Tune was employed to identify hyperparameter settings. This approach allows for systematic exploration of the hyperparameter space to find the best-performing configuration for our DRL models. The hyperparameters were chosen based on best practices and are as follows: the discount factor γ was either set to 0.95 or 0.99, the learning rate was either 0.001 or 0.00001, and the number of neurons in each layer was 32 or 64. β was 0.3. The choice of β will be elaborated in-depth in section 4.1.

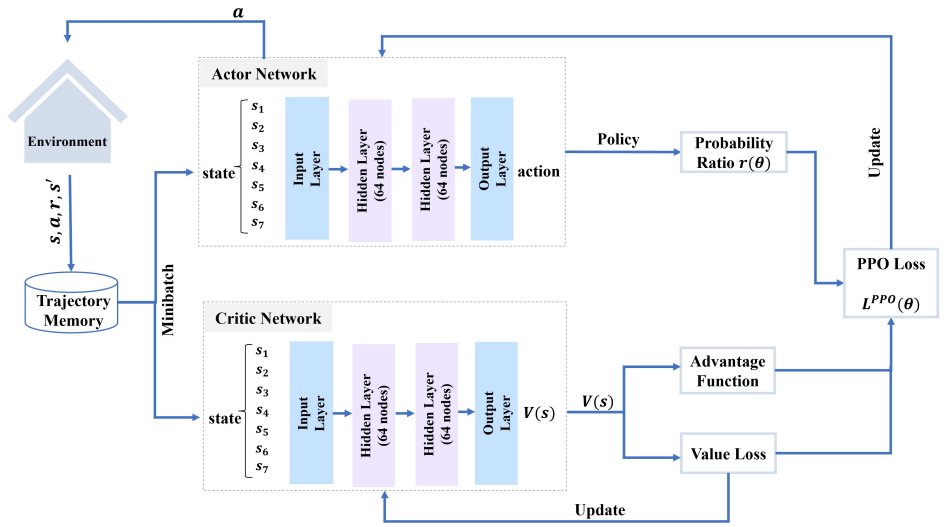


Fig. 5. The supply temperature control method for a collective space heating system using PPO. Here, *Trajectory Memory* is a storage space that stores tuples of (s, a, r, s'), and is used to sample minibatches of data for training.

Table 1. Hyperparameters used for training of the DRL agents.

Hyperparameter	Value
Discount factor γ	0.99
Learning rate	0.00001
Number of hidden layers	2
Number of neurons in each layer	64
Reward function's weight β	0.3

Each episode of training is based on 4 months of data with high heating demand during fall-winter period, starting from November 1 to February 28. The testing is done in the same period, using a different weather profile.

During training, the collective space heating system is affected by various conditions, such as fluctuating electricity prices, weather conditions, and user behavior, which includes different indoor temperature set points ($18°C$ and $20°C$). Exposing the agent to these conditions enhances its adaptability. The training and testing weather profiles belong to the same region in Belgium. Specifically, the training weather profile is an average of weather data over 20 years, while the testing weather profile is based on data from a single year. This distinction allows us to assess the agent's adaptability to varying weather conditions. Figure 6 illustrates the temperature differences between these two weather profiles over the same period. The temperature differences between two weather profiles is up to $15°C$. The variable electricity price is the Belgian DAM price of year 2022. The minimum, maximum and standard deviation of the tariff profile are €-30/MWh, €665.01/MWh, and €112.52/MWh, respectively.

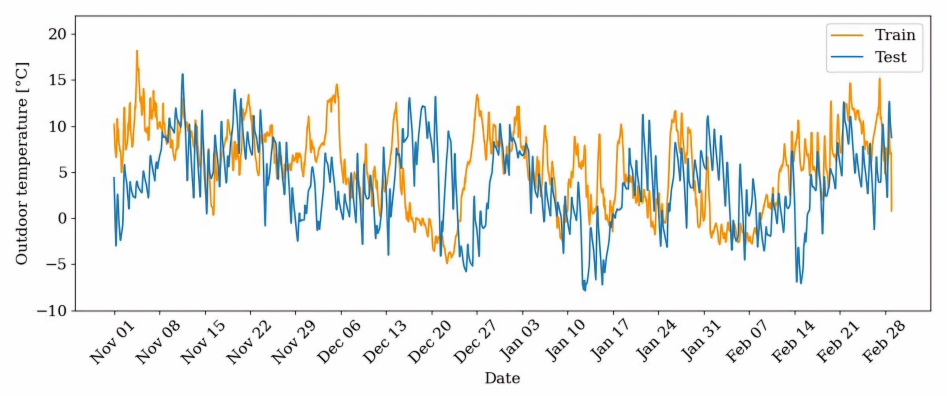


Fig. 6. Outdoor temperature trends during training and testing, from November 1st to the end of February.

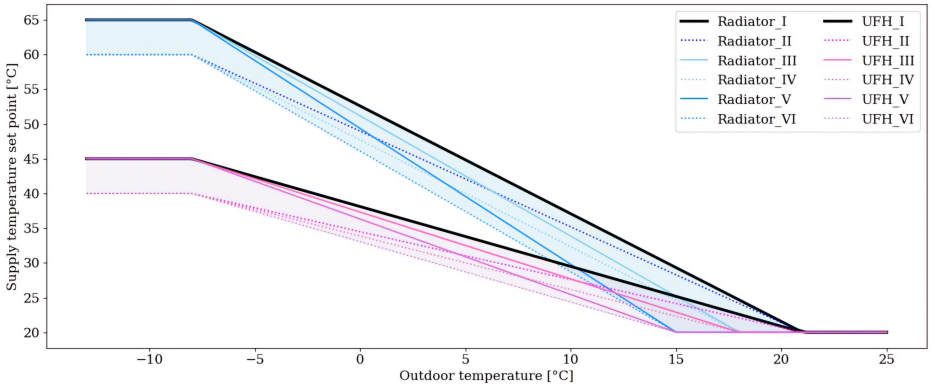


Fig. 7. Visualisation of heating curve principle for both a system with radiators (in shades of blue) and with UFH (in shades of pink). Based on outdoor temperature (T_{ext}), a supply temperature set point ($T_{sup,SP}$) is determined for the collective space heating system. The baseline heating curve in our experiments for both the UFH and radiator is shown in black.

To determine the best performing DRL algorithm for controlling a collective space heating system, i.e., optimizing for a close approximation of expected rewards in different states (DQN) or to optimize the policy directly (PPO), a series of experiments were conducted:

- First, the impact of the agent's control interval (i.e., action taking time interval) was examined by looking into a time interval of 10 minutes and 15 minutes. This is important due to the thermal inertia in the environment controlled by the DRL agent.
- Next, the influence of indoor set point temperature on the thermal comfort, cost- and energy-savings was explored. This provides an insight into the adaptability of the agents when faced with some unpredictable changes.
- Finally, the potential to further optimize the agents was studied by removing the highest temperature from Action_1, i.e, action space Action_2.

In this regard, a total of 12 simulations were performed with DRL: 6 simulations with DQN and 6 simulations with PPO for UFH and radiator. Then the results are compared to the RBC as the baseline approach (Subsection 3.3).

3.3 Baseline control approach

The baseline for testing the DRL agents is the heating curve, which is a widely used RBC for SH systems. In most buildings, the heating curve is only a preset diagram, where a certain supply temperature set point ($T_{sup,SP}$) is selected, based on T_{ext}. In Figure 7, the heating curve that serves as a baseline in our experiments for both the UFH system and radiator system is shown in black. This control approach is denoted as RBC in our experiments. The rationale for choosing this heating curve is rooted in standard practices for maintaining indoor comfort. Traditionally, a heating curve is defined based on the type of emitter and is calibrated for an indoor temperature set point of 21°C. When the outdoor temperature matches this set point (21°C), it indicates that there are no heat losses to the outside environment, making it unnecessary for the heating system to provide additional heat. Consequently, the supply temperature is set to 21°C, effectively comparable to shutting down the system as no additional heating is required.

The upper point on the heating curve is determined based on the design supply temperature for the UFH system, typically set at 40°C. An additional 5°C is added to this value to account for distribution losses and to provide a safety margin, which is a common practice in real heating systems. This brings the supply temperature to 45°C at the design condition of -8°C, which is the standard outdoor temperature used for heat loss calculations in Belgium. Thus, the curve intersects at -8°C and 45°C which is the maximum allowable temperature for UFH systems and ensures that indoor comfort is maintained even in extreme conditions.

Similarly, for radiator systems, the design supply temperature is typically set at 60°C. Adding a 5°C margin for distribution losses and safety, the supply temperature is set to 65°C at the design condition of -8°C. This ensures that the system can handle the highest heat demand scenarios while maintaining efficiency and comfort. Thus, the curve for radiators intersects at -8°C and 65°C, ensuring that the system provides adequate heating even during the coldest conditions.

However, to strengthen our evaluations, we will also compare the results of some other heating curves (as shown in Figure 7) against the best DRL agents for both UFH and radiators in section 4.

3.4 Key Performance Indicators (KPIs)

In this study, seven KPIs are used to compare the performance of the DRL agent to the RBC, but can also be used to compare the different DRL agents. A total of two energy-related KPIs, one cost-related KPI, and two comfort-related KPIs are used for general evaluation of the DRL agents. For the comparative analysis between DQN and PPO, two new KPIs are introduced, namely the Objective Score Ratio (OSR) and $OS^{P95}_{A<B}$.

The two energy-related KPIs are the following. First, the total consumed Primary Energy of the system (PE_{use} [kWh]), being the electricity use of the central GHP converted to primary energy with a conversion factor of 2.5 valid for the Belgian grid. The objective is to save energy compared to the RBC, which is denoted as *energy saving* [%] in the results. Second, the Primary Energy Ratio (PER) is the total efficiency in terms of PE, i.e. the ratio of PE_{useful} to PE_{use}. PE_{useful} [kWh] is the useful energy for SH. A higher PER indicates a higher Coefficient of Performance (COP) of the GHP and lower heat losses. To measure the operational costs related to the GHP, we used OPEX (Equation 19) expressed in €, which is calculated according to the Belgian DAM price. The objective is to save costs compared to the RBC, which is denoted as *cost saving* [%] in the results.

The two comfort-related KPIs are the following. First, the average Room Temperature Lack (RTL_{avg}) [Kh/day], as in [48], is used as a metric to quantify the average indoor thermal comfort

of the dwelling. The comfort calculations in each dwelling are based on the operative indoor temperature, which represents the perceived temperature by inhabitants due to convection and radiation. For each dwelling n the average number of Kh per day that the indoor air temperature is below the indoor temperature set point is calculated ($RTL(n)$). The RTL_{avg} is the average value of all $RTL(n)$. Second, in addition to the RTL_{avg}, RTL_{max} is also considered which equals the maximum $RTL(n)$ of all dwellings and gives an indication of the highest average of discomfort among the dwellings. $RTL(n)$ is calculated as in Equation 20.

$$RTL(n) = \int_{t_1}^{t_2} (T_{op;SP}(n) - (T_{op}(n) + e_{tol}))_+ dt \tag{20}$$

Where t_1 and t_2 are the start time and end time of the testing phase, respectively, $T_{op;SP}(n)$ the indoor temperature set point at time i for dwelling n, $T_{op}(n)$ the actual indoor operative temperature at time i for dwelling n, and e_{tol} the tolerance for indoor temperature comfort calculation set at 0.5°C.

Finally, we defined two new KPIs, the Objective Score Ratio (OSR) and $OS^{P95}_{A<B}$, to assess our methods' performance in achieving two objectives: minimizing GHP-related OPEX and maximizing indoor temperature comfort for end-users, compared to the RBC. Given that the Objective Score (OS) is equal to the reward that is achieved after taking every action in the test phase, the OSR is computed as the percentage of times a method (A) has a higher OS than the other method (B) during testing. The OSR can be calculated as follows:

$$OSR = \frac{OS_{A>=B}}{OS_{total}} \times 100\% \tag{21}$$

Where $OS_{A>=B}$ is the number of times A performs better than B in terms of OS during testing, and OS_{total} is total number of testing data points. If a method has an OSR greater than 50%, it is seen as more successful in achieving the objectives. Besides, to assess the performance during the times where A has lower OS than B, $OS^{P95}_{A<B}$ is introduced, to represent the average OS during 95% of time that $OS_{A<B}$. Since the reward, and consequently OS, falls within the range of [0, 1], a higher $OS^{P95}_{A<B}$ is an indication of a better performance during 95% of time when $OS_{A<B}$. Hence, if this KPI is higher for a method, it suggests that the method does not experience underperformance compared to the other during such periods. Using the 95th percentile ensures that the KPI is based on the vast majority of data (95%), making it representative of typical performance. At the same time, excluding the extreme 5% of cases makes this KPI robust to outliers and anomalies, ensuring that the assessment is not influenced by rare values. Thus, using the 95th percentile strikes a balance between excluding extreme cases and retaining enough data to provide a representative KPI.

4 RESULTS AND DISCUSSION

This section presents and discusses the results throughout the fall-winter period, characterized by high demand for space heating. The performance of DRL-based proposed methods are compared to the RBC, i.e., heating curve.

Table 2 provides a comprehensive overview of experiment outcomes of both DQN and PPO, all conducted under the same testing conditions. The table is divided into two sections, one for collective space heating systems with UFH in the dwellings (in blue) and another for radiators in the dwellings (in green). Experiments marked with *, use the smaller Action_2 as the action space, while the rest use Action_1. Each experiment corresponds to a specific DRL agent, trained with control (i.e., decision-making) intervals of either 10 minutes or 15 minutes. The table displays four performance metrics, including *energy saving* (%) and *cost saving* (%), both reported relative to RBC, along with the RTL_{avg} and RTL_{max} (Kh/day). The *set point* (°C) column is the indoor

Table 2. Experiment outcomes for DQN and PPO agents under similar testing condition for UFH (blue columns) and radiator (green columns). Four performance metrics are shown, including energy saving (%) and cost saving (%), both relative to RBC, along with the RTL_{avg} and RTL_{max} (Kh/day). The control interval's unit is in minutes and the set point (°C) column represents the indoor temperature set point imposed by occupants during day time or presence at home. Bold texts highlight the best performance in savings, while red signifies unfavorable outcomes. Experiments with Action_2 are marked with *.

			UFH				Radiator			
Name	Control interval (minute)	Set point (°C)	Energy saving (%)	Cost saving (%)	RTL_{avg} (Kh/day)	RTL_{max} (Kh/day)	Energy saving (%)	Cost saving (%)	RTL_{avg} (Kh/day)	RTL_{max} (Kh/day)
DQN_1	10	18	21.14	14.81	4.08	8.81	-23.41	-23.5	0.96	1.31
		20	22.38	24.21	9.97	22.18	-21.04	-20.84	2.5	3.3
DQN_1*	10	18	10.26	12.96	0.93	1.68	44.75	45.11	34	44.73
		20	8.9	6.75	1.22	2.96	50.53	51.83	55.44	66.63
DQN_2	15	18	1.68	6.5	0.49	0.82	-23.42	-23.44	0.96	1.27
		20	2.73	0.36	0.94	2.47	-21.19	-20.82	2.52	3.3
DQN_2*	15	18	**12.47**	**13.77**	1.57	2.94	51.93	52.78	44.67	56.78
		20	**10.82**	**9.59**	2.13	4.55	56.88	58.69	66.79	79.33
PPO_1	10	18	4.89	3.87	1.03	1.85	7.75	7.11	1.7	2.2
		20	7.11	11.22	2.61	5.77	9.12	8.61	5.3	6.78
PPO_1*	10	18	19.72	14.97	5.57	11.12	8.18	7.53	1.64	2.11
		20	22.84	28.89	14.13	28.28	9.2	8.36	5.02	6.76
PPO_2	15	18	1.15	-0.19	0.61	1.03	**13.85**	**13.79**	2.23	3.26
		20	3.39	5.26	1.56	3.36	**15.4**	**16.15**	7.6	9.43
PPO_2*	15	18	7.45	6.57	1.77	3.44	11.98	11.24	1.84	2.49
		20	9.46	11.06	3.89	9.27	12.37	12.12	6.01	7.92

temperature set point imposed by occupants. The RTL up to 12 Kh/day is considered as comfortable, which is equivalent to a temperature variation of only 0.5°C from the set point throughout the day. The average RTL of RBC for UFH is 0.29 and 0.51 Kh/day, while for radiators, it is 1.12 and 3.13 Kh/day for indoor set points of 18°C and 20°C, respectively. The cost- and energy-savings of the top-performing agents are highlighted in bold for both UFH and radiators, while the red color signifies an unfavorable outcome. If an agent fails in any of these metrics, it is considered as a failed experiment in our analysis. For instance, although DQN_1 and PPO_1* for UFH, achieve more than 20% energy savings and more than 14% cost savings, they are not bolded as the best ones because of the following reason: at 20°C, while the RTL_avg is below 12 for DQN_1 and slightly more than 12 for PPO_1*, the RTL_max is above 22 for both, which is high and undesirable. This high maximum RTL indicates high thermal discomfort in one of the dwellings, which is a critical factor to consider in evaluating overall performance.

The best performing agent is PPO_2 for the collective space heating system with radiators, and DQN_2* with UFH. Both have a 15-minute control interval, indicating that this is the best time interval to set. To exemplify the practical significance of cost savings, consider a scenario where the GHP operates with an OPEX of 1000 euros with RBC control strategy. In this context, our RL agents, DQN_2* and PPO_2, demonstrate noteworthy financial saving potential. The control strategy of DQN_2* can save up to 137.7 euros when deployed with UFH, while with radiators PPO_2 can save up to 161.5 euros. On the contrary, the worst performing agents are all DQN experiments in collective heating system with radiators. It should be noted that even though DQN_1 achieved the highest energy savings for UFH with an interval of 10 minutes, it has a high RTL_{max}, which means a high thermal discomfort. Therefore, it is not considered as an overall good performing agent.

It is clear that in the dynamic environment of this study, on-policy PPO shows more consistency in performance compared to off-policy DQN. The PPO consistently outperformes DQN in case of radiators, while DQN failed each experiment with radiators. This indicates the poor adaptability of DQN to a faster response systems (i.e., radiators). In case of UFH, PPO generally achieves better results, with only one exception that DQN_2* outperforms PPO_2*.

The reasons behind the performance gap between DQN as an off-policy and PPO as an on-policy method in collective space heating system control can be attributed to several factors. On-policy methods, such as PPO, generally demonstrate better performance in environments demanding extensive exploration, e.g., collective space heating systems. PPO adopts an on-policy approach to train a stochastic policy, meaning it explores by sampling actions based on the most recent version of its stochastic policy. As training progresses, the policy tends to become less random, driven by the update rule's encouragement to exploit more. However, this can lead to a local optima rather than the global optimum. In contrast, off-policy methods (e.g., DQN) prioritize exploiting known optimal actions, potentially overlooking alternative strategies.

Therefore, DQN favors the high temperature action more often. This resulted in DQN_2 having slight amount of energy savings for UFH and failing to provide savings for the radiator use case. It is observed that by removing the high temperature action (i.e., using Action_2), DQN is able to save cost and energy for both UFH and radiator, neglecting the thermal comfort in case of radiator. It it is apparent that UFH, with its slower response time and smaller action space, was better suited for DQN_2*, while the rapidly changing dynamics of the radiator cannot be captured by DQN and posed a challenge for it, which ultimately resulted in poor adaptability of DQN. Additionally, PPO directly learns the policy which allows it to tolerate some degree of inaccuracy in value function estimation. This contrasts with off-policy DQN, which heavily depends on precise value function approximations for action selection. In situations where accurately learning the value function is challenging, which often occurs in a complex and dynamic system like collective space heating systems, DQN may struggle, thus favoring on-policy methods such as PPO.

Henceforth, the discussions will be narrowed to DQN_2* (UFH) and PPO_2 (radiator), as they stand out as the most energy- and cost-effective options among the experiments, while having an acceptable RTL.

It was expected that the smaller Action_2 action space (experiments with *) would improve the cost and energy savings, but might lead to higher RTL (i.e., thermal discomfort). This assumption was true for all experiments, except PPO_2* for radiator use case. Even though both PPO_2 and PPO_2* outperformed the RBC, PPO_2* did not have higher savings and resulted in a lower RTL compared to the PPO_2. To make it more clear, the results of PPO_2 and PPO_2* for radiator are compared by dividing the OPEX of the GHP into two groups: low price (below €4) and high price (between €4 to €8). Then, the frequency of providing heat in each of these periods was counted. It turned out that PPO_2* provides heat during low price times 97.9% of the time, while this frequency for PPO_2 was 98.84% of the time. This along with providing a lower average T_{sup} by PPO_2 (1.71°C

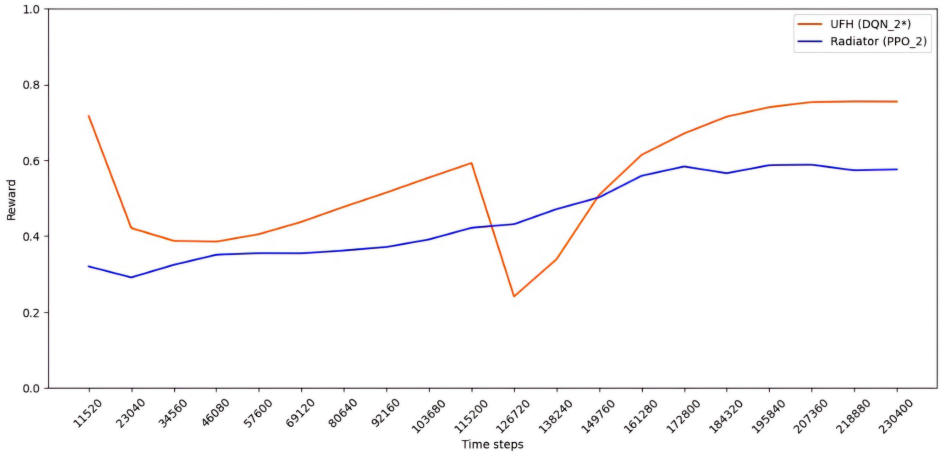


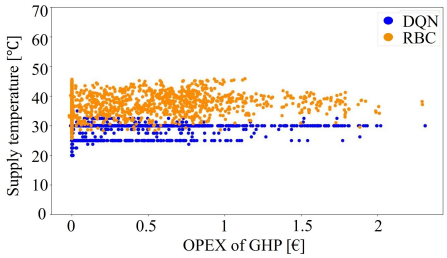
Fig. 8. Reward plot of training with training weather profile (from November 1 to February 28) for DQN_2* and PPO_2 for UFH and radiators, respectively. The reward axis shows the average of rewards across 5 runs, normalized to a scale from 0 to 1.

lower than PPO_2*) justifies the overall higher cost- and energy-savings by PPO_2. It can be that PPO_2 learned to provide high temperature at right times. Therefore, having a lower maximum $T_{sup,SP}$ is not always equal to more savings.

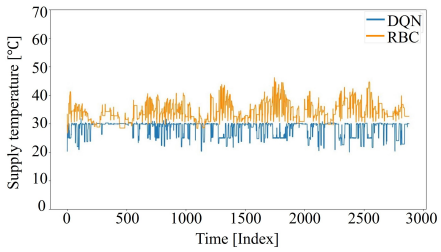
To measure the effectiveness of our proposed control strategies in achieving the specified reward function objectives, the OSR is calculated [%]. Remarkably, across both 18°C and 20°C, DQN_2* achieved an average OSR of 69.7%, while PPO_2 obtained 65%. Considering the diverse weather profiles and fluctuating electricity prices experienced by the DRL agents during testing, these high OSRs confirm the competent adaptability of our DRL agents in successfully achieving the reward function objectives. To evaluate DQN_2* (UFH) and PPO_2 (radiator) during the periods they (DQN or PPO as RL methods) had lower OS than RBC, $OS^{P95}_{RL<RBC}$ is calculated. It turned out that DQN_2* had a $OS^{P95}_{DQN<RBC}$ of 0.31 which is higher than $OS^{P95}_{RBC<DQN}$ of 0.27 for the RBC. The same holds true for the PPO_2, which has a $OS^{P95}_{PPO<RBC}$ of 0.36, while $OS^{P95}_{RBC<PPO}$ is 0.28 for the RBC. This indicates that the DRL methods still fulfill the objectives of the control better than the RBC during the time they have a lower OS.

Figure 8 shows the learning curves of DQN_2* for UFH (red) and PPO_2 for radiators (blue), by showing their respective rewards during training. The rewards are averaged over 5 runs and are normalized to a range of 0 and 1. A key observation from the figure is the dip in the reward curve for DQN_2* around the middle of the training process. This minimum point occurred precisely when the set point was changed from 18°C to 20°C. This suggests that DQN_2* initially struggled to adapt to the new set point. This reaction could be attributed to the method's difficulty in recalibrating its strategy to the updated conditions, which led to a temporary degradation in performance. Despite this initial setback, DQN_2* demonstrates a gradual recovery in performance, but it shows a potential weakness in handling sudden changes in task parameters. In contrast, PPO_2 displayed a steady increase in rewards throughout the training process, unaffected by the setpoint change. This consistent improvement suggests that PPO_2 better handles dynamic environments, and maintains stable performance without the fluctuations observed in DQN_2*.

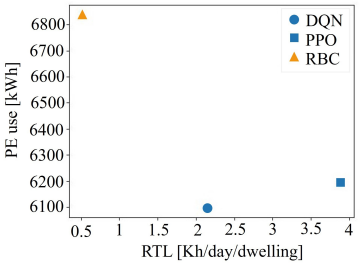
Figure 9 and Figure 10 display additional information about the performance of the DQN_2* (UFH) and PPO_2 (radiator), respectively. In both figures, subfigure (a) shows the hourly average of



(a) Hourly average supply temperature set point vs. hourly OPEX sum.

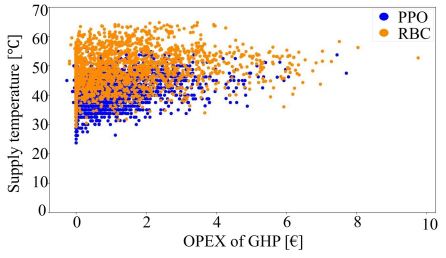


(b) Hourly average of supply temperature (T_{sup}) vs. time.

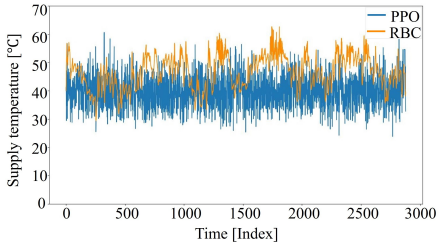


(c) Primary energy consumption of the system (PE_{use}) vs. RTL. PPO denotes PPO_2* and DQN denotes DQN_2*.

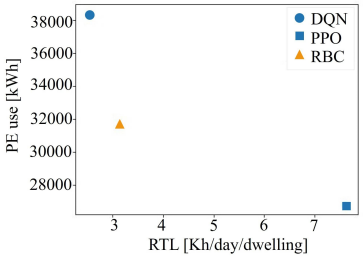
Fig. 9. DQN_2* agent controlling the $T_{sup,SP}$ of the collective space heating system with UFH during testing. The interval between taking actions is 15 minutes and the indoor temperature set point is 20°C.



(a) Hourly average supply temperature set point vs. hourly OPEX sum.

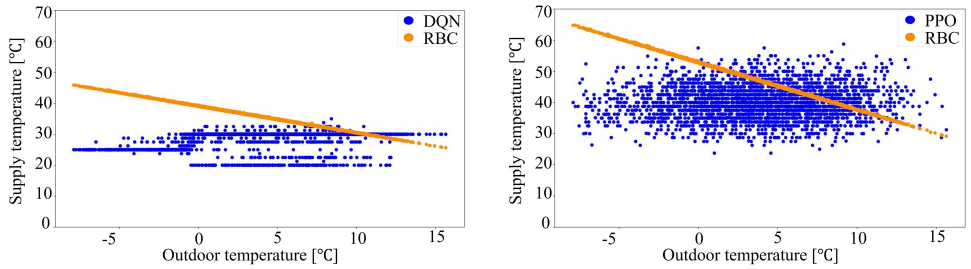


(b) Hourly average of supply temperature (T_{sup}) vs. time.



(c) Primary energy consumption of the system (PE_{use}) vs. RTL. PPO denotes PPO_2 and DQN denotes DQN_2.

Fig. 10. PPO_2 agent controlling the $T_{sup,SP}$ of the collective space heating system with radiator during testing. The interval between taking actions is 15 minutes and the indoor temperature set point is 20°C.



(a) UFH (b) Radiator

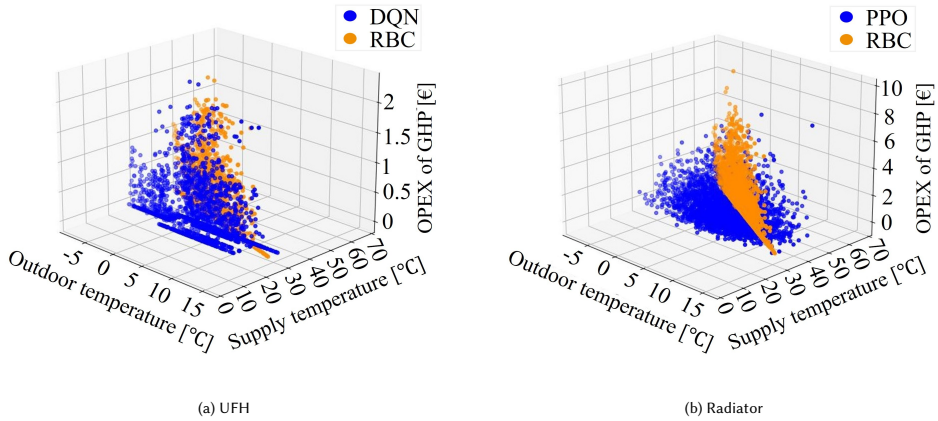
Fig. 11. Outdoor temperature vs. $T_{sup,SP}$. Comparison of the control strategy between RBC and best performing DRL agents. Subfigures (a) and (b) feature DQN_2* (UFH) and PPO_2 (radiator), respectively. The labeled supply temperature in these figures represents $T_{sup,SP}$.

supply temperature set point($T_{sup,SP}$) vs. hourly sum of OPEX. Subfigure (b) plots the hourly average of supply temperature (T_{sup}) vs. time, where maintaining a lower T_{sup} generally corresponds to less PE_{use}. Finally, subfigure (c) illustrates the primary energy consumption of the system (PE_{use}) vs. RTL, where both PE_{use} and RTL ideally should be minimized.

To assess whether the DRL agents have effectively learned to optimize the T_{sup} control while minimizing operational costs of GHP, Figure 9(a) and Figure 10(a) are presented. To secure savings, the control mechanism must meet two key criteria: firstly, it must maintain a lower average T_{sup}, and secondly, it should deliver this supply during periods of lower pricing. Referring to subfigure (a) in both Figure 9 and 10, it is evident that both DQN_2* and PPO_2 outperformed RBC by meeting both of these criteria more effectively. Furthermore, Figure 9 (b) and (c) clearly demonstrate that DQN_2* opts for a lower T_{sup} for the UFH, prioritizing energy savings and reduced OPEX, with a minor trade-off in indoor thermal comfort, where the RTL remains comfortably below 12 Kh/day (Table 2). The same holds true for the PPO_2 with radiator, where T_{sup} of the RBC is generally higher than the agent. As observed in Figure 10 (c), DQN for radiator (DQN_2 experiment) deliberately maintains a lower RTL by consuming an additional average of 11075.23 kWh of primary energy compared to the PPO_2 agent, i.e., 43.26% more energy. This behavior resulted in the failure of DQN_2. Additionally, more fluctuations are observed in T_{sup} of the agents, primarily attributed to their awareness of dynamic environmental conditions, particularly observations from the state space, such as s_1, s_2, and s_5.

Besides PE use, PER is calculated to quantify energy efficiency. PPO_2 has a higher efficiency with a PER of 1.37, compared to RBC's PER of 1.2 for radiators. Similarly, for UFH, DQN_2* achieves a PER of 2.48, surpassing RBC's 2.18. These results underscore the better efficiency of DRL over RBC. This enhanced efficiency is mainly due to the lower $T_{sup,SP}$ for most of the time, facilitated by the design of state space. By integrating information on indoor temperature set points (s2) and thermal comfort (s6), the supply temperature can reduce more frequently. This reduces heat losses during distribution, but it also increases the Coefficient of Performance (COP) of the central GHP.

To compare the $T_{sup,SP}$ control of heating curve and the top-performing DRL agents, the control strategies from DQN_2* (UFH) and PPO_2 (radiators) are plotted in Figure 11 and Figure 12. In both figures, subfigure (a) and (b) are related to the UFH and radiator use cases, respectively. Figure 11 presents T_{ext} vs. $T_{sup,SP}$. It can be seen that the $T_{sup,SP}$ control with both DQN_2* and PPO_2 resembles a cloud of set points rather than a line in RBC. This cloud-like pattern in the set points



(a) UFH (b) Radiator

Fig. 12. Comparison of $T_{sup,SP}$ control strategy between RBC and best performing DRL agents. Subfigures (a) and (b) feature DQN_2* (UFH) and PPO_2 (radiator), respectively. The labeled supply temperature in these figures represents $T_{sup,SP}$.

is attributed to the DRL agent's consideration of factors beyond just T_{ext}.This varied approach is essential for cost and energy saving while ensuring comfort levels.

To extend the comparison, the correlation between $T_{sup,SP}$, OPEX of GHP and outdoor temperature is shown as a 3D plot in Figure 12. Considering the experimental results detailed in Table 2, Figure 12 clearly indicates that traditional $T_{sup,SP}$ control, solely based on T_{ext}, is no longer suitable. Instead, the optimized control strategies proposed by both agents suggest a dynamic range of $T_{sup,SP}$, predominantly located closer to the lower end of the *OPEX of GHP* axis in the figure, representing reduced OPEX. Note that the OPEX is higher for radiators as they require higher temperatures which results in a lower COP and higher electricity consumption compared to the UFH. Given the agents' exposure to a different weather profile for testing and fluctuating electricity prices during testing, it is evident that our DRL-based control strategies adapted well to these dynamic conditions, all without explicit knowledge of the system model.

4.1 Reward function weight (β) sensitivity analysis

In this subsection, the sensitivity of the reward function weight, β, which is crucial in balancing the objectives of cost minimization and thermal comfort within our system, is examined. Extensive experimentation was conducted using grid search across potential β values of 0.1, 0.3, 0.5, 0.7, and 0.9. This analysis is aimed at determining how different weights affect the system's ability to manage the volatile nature of costs (i.e., Day-Ahead Market prices) while maintaining satisfactory thermal comfort.

Through our comprehensive sensitivity analysis, we identified that a β value of 0.3 provided the optimal balance between cost minimization and thermal comfort. This optimal setting is highlighted in Table 3, where the comparative performance of various β values is presented. The table is organized into two sections: one highlights collective space heating systems with UFH in the dwellings, colored in blue, and the other focuses on systems with radiators in the dwellings, shown in green. The cost and energy savings achieved by the top-performing β value of the reward function for both UFH and radiators are shown in bold. Additionally, β values that led to unfavorable outcomes are marked in red. Among all experiments with various β values, only $\beta = 0.1$ for Radiator (PPO_2) performed slightly worse than our baseline RBC, while the rest succeeded. Particularly, it

Table 3. Experimental results for β sensitivity analysis for UFH (DQN_2*) and Radiator (PPO_2). Four performance metrics are shown, including energy saving (%) and cost saving (%), both relative to RBC, along with the RTL_{avg} and RTL_{max} (Kh/day). The set point (°C) column represents the indoor temperature set point imposed by occupants during day time or presence at home. Bold texts highlight the best performance in savings, while red signifies unfavorable outcomes.

		UFH (DQN_2*)				Radiator (PPO_2)			
β	Set point (°C)	Energy saving (%)	Cost saving (%)	RTL_{avg} (Kh/day)	RTL_{max} (Kh/day)	Energy saving (%)	Cost saving (%)	RTL_{avg} (Kh/day)	RTL_{max} (Kh/day)
0.1	18	5.27	5.93	0.61	1.11	0.3	-0.07	1.4	1.87
	20	5.58	2.28	1.07	2.93	1.06	0.85	3.96	5.56
0.3	18	**12.47**	**13.77**	1.57	2.94	**13.85**	**13.79**	2.33	3.26
	20	**10.82**	**9.59**	2.13	4.55	**15.4**	**16.15**	7.6	9.43
0.5	18	5.3	6.77	0.61	1.17	11.4	11.61	1.9	2.61
	20	4.31	1.63	0.91	2.13	12.54	13.15	6.21	8.26
0.7	18	12	13.01	1.3	2.67	8.32	8.28	1.59	2.09
	20	10.49	8.84	1.97	4.24	8.45	8.29	4.97	6.83
0.9	18	9.93	13.3	0.81	1.45	11.66	11.64	2	2.73
	20	8.58	6.3	0.91	2.08	12.79	13.4	6.51	8.44

is observed that variations in β can significantly influence system performance; however, $\beta = 0.3$ consistently emerged as the optimal value in our sensitivity analysis. This ensures that the desired trade-off between minimizing costs and maintaining thermal comfort is achieved by the agents, as demonstrated by the results presented in Table 3.

A linear relationship between changes in β and the results is not observed; for instance, increasing β does not necessarily lead to increased cost savings. This can be attributed to two main factors:

(1) Given the volatile nature of DAM prices that are used in cost calculations, the interaction between cost and RTL is complex and non-linear. So, a higher β does not necessarily guarantee a lower cost and vice versa. According to our experiments, weight of $\beta = 0.3$ ensures that both objectives are reasonably satisfied without overly prioritizing one over the other, which could lead to suboptimal performance.

(2) The volatility of the DAM price was best managed with $\beta = 0.3$. This value allowed the agent to adapt to price fluctuations while maintaining thermal comfort, resulting in overall cost-effective performance.

4.2 Comparison of heating curves against DRL agents

In addition to the primary evaluation metrics, a comparative analysis of alternative heating curves is conducted to further validate the effectiveness of our top performing DRL agents: DQN_2* (UFH) and PPO_2 (Radiator). This comparison is done based on the heating curves that are shown in Figure 7.

Figure 13 consists of two plots comparing the performance of the best DRL agents against different heating curves in terms of OPEX versus average RTL for UFH (a) and radiators (b). In the

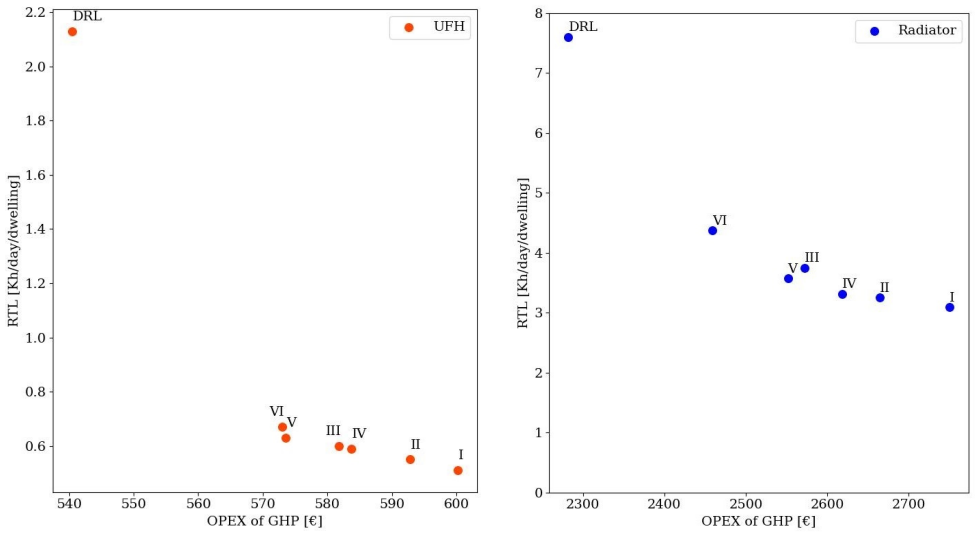


Fig. 13. Comparison of DRL agents (DQN_2* for UFH, PPO_2 for radiators) with various heating curves (Figure 7). DRL agents achieve lower OPEX with acceptable average RTL, while heating curves offer superior indoor thermal comfort (lower RTL) at higher costs.

UFH plot (a), the DRL agent (DQN_2*) achieves the lowest OPEX (€540.47) with an RTL_{avg} of 2.13 kWh/day/dwelling, which although higher than the heating curves, remains within an acceptable range for thermal comfort (RTL below 12 is desirable). Heating curve I, with the lowest RTL_{avg} (0.51 kWh/day/dwelling) and higher OPEX (€600.22), offers the best comfort. This heating curve is designed to maintain indoor comfort across a range of outdoor temperatures. In the radiator plot (b), the DRL agent (PPO_2) also achieves the lowest OPEX (€2281.99) with an RTL_{avg} of 7.6 kWh/day/dwelling, which is well within the desirable range. Again, the heating curve I provides the most stable temperature control at the highest OPEX (€2750.41), reflecting its design for maximum comfort across a range of outdoor temperatures.

These figures demonstrate that while DRL agents, such as DQN_2* and PPO_2 are optimized for cost efficiency, they manage to maintain a reasonable level of thermal comfort. The DRL agents' ability to adjust dynamically to fluctuating conditions, including varying electricity prices, allows them to operate at lower costs without severely compromising comfort. However, this approach contrasts with traditional heating curves which are inherently more rigid, focusing on maintaining indoor comfort under all conditions. This rigidity, while ensuring indoor thermal comfort, does not account for variations in electricity prices, leading to higher OPEX. The consequence of this traditional approach is evident in the higher OPEX, as these heating curves do not optimize for variable electricity prices. The traditional curves operate under the assumption of constant energy costs, leading to an over-provision of heat during periods when it might not be necessary or when cheaper heating strategies could be employed. The DRL agents, on the other hand, are designed to adapt in real-time, reducing operational costs by adjusting supply temperature based on price fluctuations and real-time demand.

Thus, as it is observed from Figure 13, this comparison highlights the trade-offs between traditional heating curves and modern DRL-based control strategies. While traditional curves offer

robust comfort, they do so at a higher operational cost. In contrast, DRL agents offer a more cost-effective solution, adapting to external variables, e.g., electricity prices, yet still providing adequate comfort levels. This difference emphasizes the potential of DRL approaches in optimizing heating control strategies for both cost and thermal comfort, particularly in dynamic environments.

5 CONCLUSION

This paper presents two adaptive DRL-based approaches, off-policy value-based (DQN) and on-policy policy-based (PPO), for controlling the supply temperature of a GHP connected to a collective space heating system. The innovative feature resides in their ability to maintain efficiency amidst dynamic environmental changes and balance competing objectives of thermal comfort, cost, and energy savings. The proposed approach integrates indoor temperature set points reflecting user preferences, alongside electricity price fluctuations and outdoor temperature in the design of the MDP to enhance the adaptability.

Based on 4 months of fall-winter data, 12 experiments with UFH and radiators were conducted using DQN and PPO which varied in control intervals. The DRL agents were tested on a different weather profile than training, and were compared to an RBC using different KPIs. It was expected that the smaller action space (Action_2) would save more energy and costs since it does not contain the highest set point of Action_1. However, this assumption was challenged by the PPO_2* of radiator, indicating that lower maximum set points do not always translate to more cost and energy savings. Successful savings also depend on the agent's ability to supply high temperatures at the right times, i.e. at times of low electricity prices with a heat demand expected in the near future. The results also indicate that DQN excels in UFH whereas PPO performs best for radiator while keeping the thermal comfort. Both DQN and PPO achieve significant cost savings compared to the RBC, with DQN reaching up to 13.77% and PPO up to 16.15%, alongside notable energy savings of 12.47% and 15.4%, respectively. Moreover, it is observed that on-policy learning with PPO generally achieves better performance consistency than off-policy learning with DQN. Additionally, unlike the RBC's rigid outdoor temperature-dependent approach, DRL-based methods dynamically adjust $T_{sup,SP}$ based on a cloud of set points, which is essential for achieving higher savings.

Future works should encompass the control of thermostats in dwellings to enhance comfort and achieve additional cost and energy savings. This is an interesting way to increase the flexibility by being fully in control of the thermal demands where the boundaries for indoor thermal comfort should be respected (e.g., maximum $1°C$ deviation from the set point). Moreover, strategies for adapting to significant environmental changes, such as accommodating occupants with different behavioral patterns or adjusting to alterations in dwelling insulation, should also be explored. Additionally, investigating the feasibility of transferring policies to similar environments is worthwhile. Finally, efforts should focus more on addressing privacy risks linked to data-driven solutions like DRL to facilitate their real-world adoption.

ACKNOWLEDGMENTS

This research is partially funded by a PhD fellowship of the Research Foundation Flanders (FWO) [1S08622N].

REFERENCES

[1] Khalil Al Sayed, Abhinandana Boodi, Roozbeh Sadeghian Broujeny, and Karim Beddiar. 2024. Reinforcement learning for HVAC control in intelligent buildings: A technical and conceptual review. *Journal of Building Engineering* 95 (2024), 110085. https://doi.org/10.1016/j.jobe.2024.110085

[2] Kari Alanne and Seppo Sierla. 2022. An overview of machine learning applications for smart buildings. *Sustainable Cities and Society* 76 (2022), 103445.

[3] Javier Arroyo, Carlo Manna, Fred Spiessens, and Lieve Helsen. 2022. Reinforced model predictive control (RL-MPC) for building energy management. *Applied Energy* 309 (2022), 118346.

[4] Bert J Claessens, Dirk Vanhoudt, Johan Desmedt, and Frederik Ruelens. 2018. Model-free control of thermostatically controlled loads connected to a district heating network. *Energy and Buildings* 159 (2018), 1–10.

[5] European Commission. 2020. Renovation wave. https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/renovation-wave_en

[6] Margot De Pauw, Freek Van Riet, Jef De Schutter, Simon Binnemans, Jeroen Van der Veken, and Ivan Verhaert. 2018. A methodology to compare collective heating systems with individual heating systems in buildings. In *The REHVA Annual Meeting Conference: Low Carbon Technologies in HVAC*. REHVA, Brussels, Belgium.

[7] Yan Du, Helia Zandi, Olivera Kotevska, Kuldeep Kurte, Jeffery Munk, Kadir Amasyali, Evan Mckee, and Fangxing Li. 2021. Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning. *Applied Energy* 281 (2021), 116117.

[8] Elexys. 2022. *Spot BelpexDAM prices for electricity in Belgium.* https://my.elexys.be/MarketInformation/SpotBelpex.aspx Accessed: 17-01-2023..

[9] European Commission. 2020. A Renovation Wave for Europe - greening our buildings, creating jobs, improving lives. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1603122220757&uri=CELEX:52020DC0662

[10] European Commission. 2021. A European Green Deal: Striving to be the first climate-neutral continent. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en

[11] European Environment Agency. 2023. Greenhouse Gas Emissions from Energy Use in Buildings in Europe. https://www.eea.europa.eu/ims/greenhouse-gas-emissions-from-energy

[12] Joshua Fong, Jerry Edge, Chris Underwood, Andy Tindale, and Steve Potter. 2015. Performance of a dynamic distributed element heat emitter model embedded into a third order lumped parameter building model. *Applied Thermal Engineering* 80 (Apr 2015), 279–287. https://doi.org/10.1016/j.applthermaleng.2015.01.067

[13] Guanyu Gao, Jie Li, and Yonggang Wen. 2019. Energy-efficient thermal comfort control in smart buildings via deep reinforcement learning. *arXiv preprint arXiv:1901.04693* (2019).

[14] Sara Ghane, Stef Jacobs, Wim Casteels, Christian Brembilla, Siegfried Mercelis, Steven Latré, Ivan Verhaert, and Peter Hellinckx. 2021. Supply temperature control of a heating network with reinforcement learning. In *2021 IEEE International Smart Cities Conference (ISC2)*. IEEE, 1–7.

[15] Anchal Gupta, Youakim Badr, Ashkan Negahban, and Robin G Qiu. 2021. Energy-efficient heating control for smart buildings with deep reinforcement learning. *Journal of Building Engineering* 34 (2021), 101739.

[16] Gwangwoo Han, Hong-Jin Joo, Hee-Won Lim, Young-Sub An, Wang-Je Lee, and Kyoung-Ho Lee. 2023. Data-driven heat pump operation strategy using rainbow deep reinforcement learning for significant reduction of electricity cost. *Energy* 270 (2023), 126913.

[17] Mengjie Han, Ross May, Xingxing Zhang, Xinru Wang, Song Pan, Da Yan, Yuan Jin, and Liguo Xu. 2019. A review of reinforcement learning methodologies for controlling occupant comfort in buildings. *Sustainable Cities and Society* 51 (2019), 101748.

[18] Chenzi Huang, Stephan Seidel, Fabian Paschke, and Jan Bräunig. 2022. A reinforcement learning approach for optimal heating curve adaption. In *2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, 1–4.

[19] Yongming Huang, Chunmei Xu, Cheng Zhang, Meng Hua, and Zhengming Zhang. 2019. An overview of intelligent wireless communications using deep reinforcement learning. *Journal of Communications and Information Networks* 4, 2 (2019), 15–29.

[20] Stef Jacobs, Margot De Pauw, Senne Van Minnebruggen, Sara Ghane, Thomas Huybrechts, Peter Hellinckx, and Ivan Verhaert. 2023. Grouped Charging of Decentralised Storage to Efficiently Control Collective Heating Systems: Limitations and Opportunities. *Energies* 16, 8 (2023). https://doi.org/10.3390/en16083435

[21] Stef Jacobs, Sara Ghane, Ali Anwar, Siegfried Mercelis, Peter Hellinckx, and Ivan Verhaert. 2022. Reinforcement learning based mass flow and supply temperature control for combined heat distribution. In *IECON 2022 – 48th Annual Conference of the IEEE Industrial Electronics Society*. 1–6. https://doi.org/10.1109/IECON49645.2022.9968547

[22] Abhyuday Jagannatha, Philip Thomas, and Hong Yu. 2018. Towards high confidence off-policy reinforcement learning for clinical applications. In *CausalML Workshop, ICML*.

[23] Joint Research Centre. 2023. Residential heating: heat pumps would knock down energy consumption and emissions. https://joint-research-centre.ec.europa.eu/jrc-news-and-updates/residential-heating-heat-pumps-would-knock-down-energy-consumption-and-emissions-2023-06-21_en.

[24] Adrien Le-Coz, Tahar Nabil, and Francois Courtot. 2020. Towards optimal district heating temperature control in china with deep reinforcement learning. *arXiv preprint arXiv:2012.09508* (2020).

[25] Lei Lei, Yue Tan, Kan Zheng, Shiwen Liu, Kuan Zhang, and Xuemin Shen. 2020. Deep reinforcement learning for autonomous internet of things: Model, applications and challenges. *IEEE Communications Surveys & Tutorials* 22, 3

(2020), 1722–1760.
[26] Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Ken Goldberg, Joseph Gonzalez, Michael Jordan, and Ion Stoica. 2018. RLlib: Abstractions for distributed reinforcement learning. In *International Conference on Machine Learning*. PMLR, 3053–3062.
[27] Peter Lichtenwoehrer, Susanna Erker, Franz Zach, and Gernot Stoeglehner. 2019. Future compatibility of district heating in urban areas—a case study analysis in the context of integrated spatial and energy planning. *Energy, sustainability and society* 9, 1 (2019), 1–12.
[28] Paulo Lissa, Conor Deane, Michael Schukat, Federico Seri, Marcus Keane, and Enda Barrett. 2021. Deep reinforcement learning for home energy management system control. *Energy and AI* 3 (2021), 100043.
[29] Henrik Lund, Sven Werner, Robin Wiltshire, Svend Svendsen, Jan Eric Thorsen, Frede Hvelplund, and Brian Vad Mathiesen. 2014. 4th Generation District Heating (4GDH): Integrating smart thermal grids into future sustainable energy systems. *Energy* 68 (2014), 1–11.
[30] Karl Mason and Santiago Grijalva. 2019. A review of reinforcement learning for autonomous building energy management. *Computers & Electrical Engineering* 78 (2019), 300–312.
[31] Nina Mazyavkina, Sergey Sviridov, Sergei Ivanov, and Evgeny Burnaev. 2021. Reinforcement learning for combinatorial optimization: A survey. *Computers & Operations Research* 134 (2021), 105400.
[32] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).
[33] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.
[34] Adam Nagy, Hussain Kazmi, Farah Cheaib, and Johan Driesen. 2018. Deep reinforcement learning for optimal control of space heating. *arXiv preprint arXiv:1805.03777* (2018).
[35] Sindhu Padakandla. 2021. A survey of reinforcement learning algorithms for dynamically varying environments. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–25.
[36] Giuseppe Pinto, Marco Savino Piscitelli, José Ramón Vázquez-Canteli, Zoltán Nagy, and Alfonso Capozzoli. 2021. Coordinated energy management for a cluster of buildings through deep reinforcement learning. *Energy* 229 (2021), 120725.
[37] Naren Srivaths Raman, Adithya M Devraj, Prabir Barooah, and Sean P Meyn. 2020. Reinforcement learning for control of building HVAC systems. In *2020 American Control Conference (ACC)*. IEEE, 2326–2332.
[38] Frederik Ruelens, Sandro Iacovella, Bert J Claessens, and Ronnie Belmans. 2015. Learning agent for a heat-pump thermostat with a set-back strategy using model-free reinforcement learning. *Energies* 8, 8 (2015), 8300–8318.
[39] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
[40] Solar Energy Laboratory Univ. of Wisconsin-Madison. 2009. TRNSYS 17 Volume 4 Mathematical Reference: Type 60 (Stratified fluid storage tank with internal heat exchangers). , 390–396 pages.
[41] Solar Energy Laboratory Univ. of Wisconsin-Madison. 2014. TRNSYS 17 Volume 8 Weather Data.
[42] EPEX SPOT. 2021. Basics of the Power Market. https://www.epexspot.com/en/basicspowermarket
[43] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement Learning: An Introduction*. The MIT Press.
[44] Unicorn systems. 2022. Entsoe Transparency platform. https://transparency.entsoe.eu/transmission-domain/r2/dayAheadPrices/
[45] Agne Toleikyte, REINA Juan Carlos Roca, Jonathan Volt, Johan Carlsson, Lorcan Lyons, Andrea Gasparella, Derck Koolen, Matteo De Felice, Dalius Tarvydas, Veronika Czako, Georgios Koukoufikis, Anna Kuokkanen, and Simon Letout. 2023. *The Heat Pump Wave: Opportunities and Challenges*. Scientific analysis or review. Luxembourg (Luxembourg). https://doi.org/10.2760/27877
[46] Paul Van den Bossche, Jeroen Van der Veken, Sébastien Pecceu, , Sara Verheyleweghen, and Stijn Verbeke. 2022. 'Power gap' in Heat Load Calculations - EN12831-1 versus monitoring and simulation results. In *Proceedings of CLIMA 2022 conference: 14th HVAC World Congress*. TUDelft, Rotterdam, The Netherlands, 1–8. https://doi.org/10.34641/clima.2022.134
[47] Hado Van Hasselt, Arthur Guez, and David Silver. 2016. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.
[48] Freek Van Riet. 2019. *Hydronic design of hybrid thermal production systems in buildings*. University of Antwerp.
[49] Freek Van Riet, Gunther Steenacker, and Ivan Verhaert. 2018. A new approach to model transport delay in branched pipes. In *10th International Conference on System Simulation in Buildings*. 1–11.
[50] Vlaams Energie – en Klimaatagentschap. 2022. Energiebesluit, bijlage V – Bepalingsmethode EPW 2022. https://www.vlaanderen.be/epb-pedia/epb-regelgeving/energiebesluit-en-bijlagen/energiebesluit-bijlage-v

[51] VLAIO. 2021. Instal 2020 project: Integraal ontwerp van installaties voor sanitair en verwarming (Dutch). *VIS 135098 (2014-2018)* (2021). https://www.instal2020.be
[52] VLAIO. 2021. Productie en distributie van Sanitair warm water: selectie en dimensionering (Dutch). *TETRA 120145 (2012-2014)* (2021). https://www.tetra-sww.be/
[53] Zhe Wang and Tianzhen Hong. 2020. Reinforcement learning for building controls: The opportunities and challenges. *Applied Energy* 269 (2020), 115036.
[54] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. 2016. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*. PMLR, 1995–2003.
[55] Yujian Ye, Dawei Qiu, Xiaodong Wu, Goran Strbac, and Jonathan Ward. 2020. Model-free real-time autonomous control for a residential multi-energy system using deep reinforcement learning. *IEEE Transactions on Smart Grid* 11, 4 (2020), 3068–3082.
[56] Zhiang Zhang, Adrian Chong, Yuqi Pan, Chenlu Zhang, and Khee Poh Lam. 2019. Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning. *Energy and Buildings* 199 (2019), 472–490.
[57] Bingwen Zhao, Yu Jin, Wan Li, and Hanyu Zheng. 2022. Analysis on the Technical Situation and Applied Difficulties of District Heating Load Forecasting. *Thermal Engineering* 69, 6 (2022), 464–472.
[58] Xiaofeng Zheng, Edward W. Cooper, Joe Mazzon, Ian Wallis, and Christopher J. Wood. 2019. Experimental insights into the airtightness measurement of a house-sized chamber in a sheltered environment using blower door and pulse methods. *Building and Environment* 162 (2019), 106269. https://doi.org/10.1016/j.buildenv.2019.106269

A ALGORITHMS

Algorithm 1 Actor-Critic Proximal Policy Optimization Algorithm

Initialize trajectory memory TM
Initialize the policy network (actor) and the value network (critic)
Initialize N with the total number of (parallel) actors
Set hyperparameters (refer to Table 1)
for episode = 1 to MaxEpisodes **do**
    for actor = 1 to N **do**
        Run policy $\pi_{\theta_{\text{old}}}$ in environment for T time steps
        Store collected trajectories in trajectory memory TM
        Compute advantage estimates $\hat{A}_1$ to $\hat{A}_T$
    end for
    Compute loss using Equation 14
    Update $\theta_{\text{old}} \leftarrow \theta$
end for

Algorithm 2 Dueling Double DQN Algorithm

Initialize replay buffer RB
Initialize primary Q-network (Q) with random weights θ
Initialize target Q-network (Q') with weights $\theta' = \theta$
Set hyperparameters (refer to Table 1)
for episode = 1 to MaxEpisodes **do**
    Reset the environment and observe initial state s_0
    for time step = i to T **do**
        With probability ϵ, select a random action a_i
        Otherwise, select $a_i = \arg\max_a Q(s_i, a; \theta)$
        Execute action a_i, observe reward r_i and next state s_{i+1}
        Store transition (s_i, a_i, r_i, s_{i+1}) in replay buffer RB
        Sample a random minibatch of transitions from RB
        Compute $y_i^{\text{Dueling Double DQN}}$ according to Equation 9
        Compute loss using Equation 10
        if time step mod C == 0 **then**
            Update the target network: $\theta' \leftarrow \theta$
        end if
    end for
end for