Improving the learning process of deep reinforcement learning agents operating in collective heating environments.

Stef Jacobs[a,b,*], Sara Ghane[c], Pieter Jan Houben[b], Zakarya Kabbara[a], Thomas Huybrechts[c], Peter Hellinckx[b], Ivan Verhaert[a]

[a] *EMIB, Faculty of Applied Engineering - Electromechanical Engineering Technology, University of Antwerp, Groenenborgerlaan 171, Antwerp, 2020, Belgium*
[b] *M4S, Faculty of Applied Engineering - Electronics ICT, University of Antwerp, Groenenborgerlaan 171, Antwerp, 2020, Belgium*
[c] *University of Antwerp - imec, IDLab - Faculty of Applied Engineering, Sint-Pietersvliet 7, Antwerp, 2000, Belgium*

Abstract

Deep reinforcement learning (DRL) can be used to optimise the performance of Collective Heating Systems (CHS) by reducing operational costs while ensuring thermal comfort. However, heating systems often exhibit slow responsiveness to control inputs due to thermal inertia, which delays the effects of actions such as adapting temperature set points. This delayed feedback complicates the learning process for DRL agents, as it becomes more difficult to associate specific control actions with their outcomes. To address this challenge, this study evaluates four hyperparameter schemes during training. The focus lies on schemes with varying learning rate (the rate at which weights in neural networks are adapted) and/or discount factor (the importance the DRL agent attaches to future rewards). In this respect, we introduce the GALER approach, which combines the progressive increase of the discount factor with the reduction of the learning rate throughout the training process. The effectiveness of the four learning schemes is evaluated using the actor-critic Proximal Policy Optimization (PPO) algorithm for three types of CHS with a multi-objective reward function balancing thermal comfort and energy use or operational costs. The results demonstrate that energy-based reward functions allow for limited optimisation possibilities, while the GALER scheme yields the highest potential for price-based optimisation across all considered concepts. It achieved a 3%-15% performance improvement over other successful training schemes. DRL agents trained with GALER schemes strategically anticipate on high-price times by lowering the supply temperature and vice versa. This research highlights the advantage of varying both learning rates and discount factors when training DRL agents to operate in complex multi-objective environments with slow responsiveness.

Keywords: Reinforcement Learning, Thermal Inertia, Control Strategy, Discount Factor, Learning Rate Schedule, Collective Heating, PPO

1. Introduction

In the European Union, households account for 26.9% of the final energy use [1] and 28% of total CO_2 emissions [2]. With 80% of energy use in buildings dedicated to Space Heating (SH), Space Cooling (SC) and Domestic Hot Water (DHW), improving energy efficiency of thermal energy supply is critical for meeting the climate goals set for 2050 [3, 4]. Minor improvements in control strategies could reduce global energy use in buildings by up to 30%, highlighting the impact of energy efficiency [5]. Consequently, many countries are adopting energy efficiency standards.

Collective Heating Systems (CHS) are essential in this context, as they can integrate multiple energy vectors to optimise overall energy supply [6]. By connecting multiple thermal end-users with shared generation, CHS benefit from consistent thermal demand and economies of scale, making them ideal for incorporating Heat Pumps (HP) [7]. Electrification with HPs not only improves energy efficiency compared to gas boilers but also supports the transition to renewable-based electricity grids, reducing dependence on fossil fuels.

Despite advances in lowering the distribution temperature in CHS for enhanced energy efficiency, further optimisation of control strategies remains essential [8, 9]. The challenge lies in balancing multiple conflicting objectives, such as minimising energy use, reducing costs, and maintaining thermal comfort. Moreover, the thermal inertia, which delays the effects of control actions, further complicates the optimisation. CHS involve complex, non-linear systems with multiple interconnected components and end-users, each with distinct thermal demands and temperature needs. Therefore, real-time control strategies that anticipate future system requirements and (future) optimisation opportunities, while accounting for time delays and conflicting objectives, are essential.

Traditional rule-based control systems, such as ON/OFF and proportional-integral-derivative (PID), are limited in their ability to optimise multiple objectives and do not employ forward-looking decisions. These limitations have led to a growing

*Corresponding author
Email addresses: Stef.Jacobs@uantwerpen.be (Stef Jacobs), Sara.Ghane@uantwerpen.be (Sara Ghane), PieterJan.Houben@uantwerpen.be (Pieter Jan Houben), Zakarya.Kabbara@uantwerpen.be (Zakarya Kabbara), Thomas.Huybrechts@imec.be (Thomas Huybrechts), Peter.Hellinckx@uantwerpen.be (Peter Hellinckx), Ivan.Verhaert@uantwerpen.be (Ivan Verhaert)

Jacobs Stef, Ghane Sara, Houben Pieter Jan, Kabbara Zakarya, Huybrechts Thomas, Hellinckx Peter, Verhaert Ivan.- Improving the learning process of deep reinforcement learning agents operating in collective heating environments
Applied energy - ISSN 1872-9118 - 384(2025), 125420
Full text (Publisher's DOI): https://doi.org/10.1016/J.APENERGY.2025.125420

research trend towards advanced control methodologies [10]. Model Predictive Control (MPC) and Reinforcement Learning (RL) are emerging as key areas for optimising control strategies in CHS.

1.1. Advanced control: MPC vs. RL

MPC uses a model of the system and disturbance predictions to compute the performance of control actions by solving an optimisation problem over a finite prediction horizon [11]. At each interval, MPC simulates and evaluates the effects of several control actions to minimise a cost function over the defined time horizon. This is particularly beneficial for slow-responsive systems, as it inherently considers the long-term consequences of actions. Key strengths of MPC include its adherence to system constraints and ensuring safety. However, finding the optimal control action requires evaluating all possible scenarios, which can be computationally intensive, especially since MPC does not learn from past situations. Moreover, its performance heavily relies on the accuracy of the simulation model used. These challenges could make the design of an MPC controller more complex for large-scale, stochastic environments, due to the high computational demands and the need for an accurate digital twin of the system [12].

In contrast, model-free RL is a model free machine learning approach where an agent learns sequential decision-making strategies under uncertainty through interaction with its environment, framed as a Markov Decision Process (MDP) [13]. An MDP consists of the state space (S), the set of possible actions (A), the probability distribution of state transitions ($P(s_{t+1}|s_t, a_t)$), and a reward function ($r_t = R(s_t, a_t, s_{t+1})$) [13]. The policy ($\pi$) guides action selection based on the current state s_t and can be either deterministic or stochastic. The objective of the RL agent is to find a policy (π^*) that maximises a discounted sum of rewards: $\sum_{t=0}^{\infty} \gamma^t r_t$, where the discount factor (γ) is a value between 0 and 1 that determines the importance attached to future rewards. In this way, the RL agent learns which actions lead to positive outcomes in which states. By doing so, RL relies on value functions to estimate expected cumulative reward which are related to the current state, action and discount factor. A model-free approach also ensures transferability of learned policies to similar optimisation problems [14]. Deep Reinforcement Learning (DRL) extends RL by using deep neural networks to approximate value functions and policies, enabling it to handle high-dimensional inputs and complex environments by generalising policies across similar states to reduce learning time [15, 16, 17, 18].

DRL also benefits from simulator training before deployment, which reduces online computation time. Although it adapts to complex, non-stationary environments by updating its policy π with new experiences [12, 19], tuning hyperparameters such as the learning rate (α) is crucial for effective learning and to achieve convergence [16, 15]. The learning rate α controls the step sizes taken when updating the weights of neural networks [13]. A high α speeds up learning but may cause instability, while a low α ensures precise updates but slows down the learning process. Moreover, the thermal inertia of CHS poses unique challenges for DRL agents, particularly due to the delayed effects of actions, resulting in complex value functions that need to be learned.

In summary, MPC provides a structured, model-based, and safe approach to control optimisation but lacks adaptability due to the need for an accurate digital twin of the system during deployment. Moreover, it cannot learn from past actions requiring to evaluate all possible scenarios at each interval, making it computational intensive. DRL, while flexible and adaptive to complex environments, faces challenges when learning in slow-responsive systems due to delayed effects of actions [13]. A model-based DRL approach has been proposed [20], but this also increases the computational requirements and relies on a model of the environment during deployment. Therefore, the focus lies on model-free DRL algorithms, which only might use a simulator model of the environment during training, but not during deployment.

Addressing the challenges of model-free DRL requires careful tuning of various hyperparameters, such as discount factor γ and learning rate α, among others. This paper explores the potential of DRL to control the central supply temperature (T_{sup}) of CHS by investigating different tuning schemes for γ and α during training. Hereinafter, the term "hyperparameters" refers specifically to α and γ.

1.2. Related works: DRL in thermal systems

DRL algorithms have been used for load managing, where the effects of thermal inertia on the agent's learning performance are limited. For example, Lissa et al. [21] utilised an RL agent to manage SH and DHW systems, optimising photovoltaic self-consumption and achieving energy savings with a learning rate α of 0.0001 and a discount factor γ fixed at 0.95. Bahrami et al. [22] explored DRL for demand-response in thermal networks, distributing an agent and merging the learned policies from each household. Their approach resulted in similar performance as the centralised actor-critic method with a γ of 0.9. Pinto et al. [23] employed a centralised Soft-Actor Critic (SAC) DRL controller in the CityLearn environment [24] to manage thermal storages connected to four commercial buildings with renewable energy sources. The paper aimed to reduce energy costs and peak demands by training the algorithm over five episodes of 92 days each, with a fixed α of 0.003. The authors highlighted that incorporating future price and weather predictions into the state space allowed the SAC algorithm to quickly converge to a well-performing control policy.

DRL algorithms have also been used to control actuators in thermal systems. For instance, Moriyama et al. [25] demonstrated that model-free RL achieved a 22% reduction in energy costs compared to a built-in model-based control for data centers. The model-based control was a built-in approach available in the Building Controls Virtual Test Bed of EnergyPlus [26]. The used RL algorithm was Trust Region Policy Optimization with a fixed γ of 0.99 and learning rate α of 0.001. Heidari et al. [14] proposed an RL-based control strategy based on double deep Q-learning method for optimal activation of air-source HP connected to DHW storage tank. The RL framework was designed to balance the thermal comfort of end-users, energy ef-

ficiency, and hygiene, while learning the stochastic DHW consumption patterns. The goal was to accelerate training without disturbing the occupants. Hence, it is pre-trained offline, with a stochastic model of occupant behaviour with fixed a learning rate of 0.003 and a 20-week memory buffer to analyse state-action interactions. Finally, transfer learning is used for fine-tuning the policy with real-world data gathered over 27 weeks. This approach realised a 23.8% reduction of energy use, while maintaining thermal comfort and hygiene to the occupants, for a test period of two weeks. Huang et al. [27] adapted the heating curve of a collective SH network in an office building using Q-learning, reducing overheating by including future estimates into the observation space. The optimal learning rate (0.6) and discount factor (0.8) were chosen based on preliminary results. They suggest the use of DRL methodologies to further optimise the performance. Consequently, Chatterjee and Khovalyg [28] provided a comprehensive review of RL applications in dynamically varying indoor temperatures set points of buildings. They suggest that actor-critic methods may be the better choice for controlling temperatures, which was also acknowledged in the technical literature review of Al Sayed et al. [15]. In this regard, Ghane et al. [29] investigated a Proximal Policy optimization (PPO) agent to reduce energy use in a collective space heating network by controlling the central T_{sup}. They explored different weights of the two-objective reward function, with the PPO agent consistently outperforming the traditional heating curve. Their approach was later extended to a cost-efficient PPO agent in [30]. In [31], a PPO algorithm controlled indoor temperature set points, lowering thermal system energy demand while ensuring thermal comfort across multiple zones, with a γ of 0.99 and $\alpha = 0.0005$ for critic network and 0.00015 for the actor network. Although fixed γ and α schedules were used, the PPO agent outperformed value-based methods such as Q-learning and Deep Q-network (DQN).

Brandi et al. [32] noted a lack of research on the effects of hyperparameters settings on the performance of control strategies in thermal systems. They conducted a sensitivity analysis on the discount factor (fixed at 0.9, 0.95 and 0.99) and three different weighting factors in the reward function between energy use and indoor thermal comfort, with a fixed learning rate of 0.0001. For each combination of hyperparameters, three DQN agents were trained separately to control the T_{sup} in a space heating network to evaluate the stability of the learning process. Results showed that a higher γ could result in higher rewards but also increased instability.

Despite these advances, training DRL agents for such systems remains complex due to the trade-off between exploration and exploitation [16]. The agent starts without prior knowledge and must explore the state space to identify optimal actions, which is especially challenging in stochastic or delayed-reward systems. To the best of the authors' knowledge, the impact of variable hyperparameter schemes affecting the exploration and exploitation, and thus on the performance of DRL agents in thermal systems, has been poorly investigated. Typically, the learning rate α is fixed at a value between 0.003 and 0.0001, and the discount factor γ is set between 0.8 and 0.99. This underscores the need to investigate methodologies involving adaptive α and γ throughout the training process of DRL agents in slow-responsive systems.

1.3. Scope and Contributions of the Paper

This paper investigates the effectiveness of four training approaches for DRL agents that control T_{sup} in different slow-responsive CHSs. Specifically, we test these training schemes using Proximal Policy Optimization (PPO), a state-of-the-art on-policy algorithm, under different reward functions. The main contributions of this research are threefold.

First, the evaluation includes four distinct learning approaches based on common DRL hyperparameters, namely the discount factor γ and learning rate α. The four approaches include I) a fixed γ and fixed α scheme, II) an increasing γ with a fixed α, III) a fixed γ with a decreasing α, and IV) the new GALER method, which simultaneously increases γ and decreases α during training. Each training scheme consists of five training intervals in which γ and α can be adjusted.

Second, these learning schemes are evaluated across a range of system responsiveness, considering three case studies with varying levels of thermal inertia, which influences how quickly temperature changes occur. We evaluate three case studies: A) a collective space heating network, B) a 2-pipe system with decentralised Booster Heat Pumps (BHP), and C) a 2-pipe system with decentralised Domestic Hot Water (DHW) storages. In these systems, only the central T_{sup} is controlled by the DRL agent, while other actuators are managed by rule-based controllers. The complexity increases for DRL agents operating in systems with high thermal inertia and a large range of potential supply temperatures, as it become less responsive to immediate temperature adjustments.

Finally, this study explores two optimisation objectives that influence performance. First, an energy-based optimisation focusing on minimising energy use while balancing thermal comfort. Second, a price-based optimisation that targets to reduce operational costs under hourly variable electricity tariffs, also while balancing thermal comfort. By adjusting the weights in these weighted-sum reward functions, the priority can be shifted between thermal comfort and either energy efficiency or cost reduction. In these multi-objective optimisations, neither thermal comfort, energy use nor operational costs are considered as strict boundary condition during balancing. Instead of terminating the training when thresholds are violated, extremely low rewards were applied.

This research helps to streamline the hyperparameter tuning process for DRL agent training in slow-responsive networks. We anticipate that the GALER approach (learning scheme IV) will particularly benefit DRL agents with price-based optimisation objectives, as an optimal understanding of time-related effects is more crucial in this context.

2. Material and methods

This Section elaborates on the characteristics of the selected Deep Reinforcement Learning (DRL) algorithm, i.e. Proximal Policy Optimization (PPO), and the impact of learning rate α

and discount factor γ on the learning process. Afterwards, Section 2.3 presents the four investigated learning schemes, followed by Section 2.4 that provides an overview of the experiments to evaluate the efficacy and applicability of the proposed learning schemes. The experiments consist of training a DRL agent for three different collective thermal network concepts with different Markov Decision Process (MDP) formulations and optimisation objectives. Finally, Section 2.5 describes the simulation-based evaluation framework, including the used data sources, simulator models representing the collective heating systems, and the evaluation metrics.

2.1. Deep Reinforcement Learning (DRL)

Deep Reinforcement Learning (DRL) extends traditional Reinforcement Learning (RL) by leveraging neural networks to approximate value functions and policies. To provide a foundational understanding, this Section begins by illustrating the core principles of RL, building further on Section 1.1.

Every policy π that maps states to actions is one of the solutions of an MDP. Through trial and error, the RL agent aims to discover the policy π^* that maximises cumulative discounted rewards, guided by a reward signal $r_t = R(s_t, a_t, s_{t+1})$, which specifies the reward received after transitioning from s_t to state s_{t+1} via action a_t. The current policy from a given state is denoted as $\pi(s)$, and the expected cumulative rewards under $\pi(s)$ are captured by the state-value function ($V^\pi(s)$) and action-value function ($Q(s, \pi(s))$). To balance immediate and long-term rewards, which could encourage exploration of the environment, a discount factor γ (value between 0 and 1) is used in $V^\pi(s)$ and $Q(s, \pi(s))$ [33], as shown in Equation 1.

$$\begin{aligned} V^\pi(s) &= \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1}) | \pi, s_0 = s\Big] \\ Q(s, \pi(s)) &= \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1}) | s_0 = s, a_0 = a\Big] \end{aligned} \tag{1}$$

where t is a given control time step and $\mathbb{E}$ stands for the expected cumulative value, representing the mean value over all possible outcomes, weighted by their probabilities.

The state-value function $V^\pi(s)$ can be expressed in terms of the action-value function $Q(s, \pi(s))$, where the action $a = \pi(s)$ is the best possible action of $\pi(s)$, as shown in Equation2.

$$V^\pi(s) = \max_a Q(s, \pi(s)) \tag{2}$$

The goal of the RL agent is to identify the policy π^* that maximises these value functions:

$$\pi^*(s_0) = \text{argmax}_\pi V^\pi(s_0) = \text{argmax}_a Q(s, \pi(s)) \tag{3}$$

RL can be categorised into three main types based on their approach to identify π^*: value-based, policy-based and actor-critic methods. Value-based methods aim to optimise value functions to find π^*, whereas policy-based methods directly optimise π to maximise expected rewards. Actor-critic methods combine both approaches: the actor updates the policy directly, while the critic estimates the value function to evaluate the action taken by the actor.

In DRL, the state-value function $V^\pi(s)$, action-value function $Q(s, \pi(s))$ and the policy are approximated by neural networks V_ψ, Q_ϕ and π_θ, where ψ, ϕ, and θ are vectors containing the parameters that characterise the neural networks. The policy neural network produces action probabilities (stochastic policy) or a specific action (deterministic policy) based on the state s, while the value neural network approximates cumulative rewards based on state or state-action pairs. These approximations eliminate the need for explicit value computation due to generalising capabilities of learned patterns [16].

The loss function ($\mathcal{L}$) quantifies the difference between the predicted value or action distributions and the target values derived from the environment's feedback. By minimising this loss function, DRL algorithms adjust parameter vectors for more accurate value predictions and better policies over time.

The calculation of $\mathcal{L}$ depends on the DRL algorithm used, which can be either on-policy or off-policy based on the data sampling method [13]. Off-policy methods learn from actions that are outside the current policy, allowing to use the data from older or different policies, while on-policy methods learn from actions taken strictly according to the current policy, followed by iterative evaluation and enhancement of that policy. This study focuses on a state-of-the-art actor-critic algorithm, namely PPO, an on-policy algorithm that balances exploration and exploitation while optimising the policy.

2.2. Proximal Policy Optimization (PPO) Algorithm

PPO is a policy gradient method which aims to achieve a balance between sampling complexity, ease of tuning hyperparameters and implementation. This is done by computing a new policy that maximises the reward while preventing large policy updates and ensure stable policy improvements, using mini-batch stochastic gradient descent (SGD) on the calculated loss function. The gradient of this loss function is typically estimated using the Advantage Function A_t^π, which represents the relative value of taking action a_t over the average action of the current policy in a given state s_t. Typically, $A^\pi(s_t, a_t)$ equals $Q(s_t, \pi(s_t)) - V(s_t)$, but to reduce the variance of policy gradient updates, PPO can leverage Generalized Advantage Estimation (GAE) which combines immediate rewards and estimated future rewards [34].

Two variants are available to calculate the loss function $\mathcal{L}$ of PPO used to train the policy π_θ while ensuring that updates do not deviate excessively from the previous policy $\pi_{\theta_{old}}$, namely Adaptive Kullback-Leibler (KL) Divergence Penalty and Clipped Surrogate Objective, which can also be combined [35].

On the one hand, clipping in PPO restricts how much the new π_θ can deviate from the old one, $\pi_{\theta_{old}}$, by bounding their probability ratio ($pr_t(\theta)$) within a specified range of $[1-\epsilon, 1+\epsilon]$. This is done by modifying the loss function $\mathcal{L}$ through taking the minimum of the unbounded and clipped ratios. The $\mathcal{L}$ is calculated as in Equation 4.

$$\mathcal{L}_t^{\text{CLIP}}(\theta) = \mathbb{E}\left[\min\left(pr_t(\theta)A_t^{\pi}, \text{clip}(pr_t(\theta), 1-\epsilon, 1+\epsilon)A_t^{\pi}\right)\right] \quad (4)$$

where A_t^{π} is the advantage function according to GAE [34], used to estimate the relative value of taking action a_t over the average action of the policy in a given state s_t; the rest is as before.

In practice, PPO can use a combined loss function that incorporates both the clipped surrogate objective and the KL divergence penalty. The KL divergence approach penalises large differences between the old and new policy distributions by adding a penalty term to the loss function [36]. Equation 5 shows the combination of clipping with KL divergence and the loss function for the value neural network:

$$\begin{aligned}\mathcal{L}_t^{\text{PPO}}(\theta) = \mathbb{E}\Big[\ & \mathcal{L}_t^{\text{CLIP}}(\theta) \\ & - c_{KL}\text{KL}\left[\pi_{\theta_{old}}, \pi_{\theta}\right] \\ & + c_{ent}\mathbb{S}[\pi_{\theta}](s_t) \\ & - c_{VF}\mathcal{L}_t^{\text{VF}}(\theta)\ \Big]\end{aligned} \quad (5)$$

where $\mathcal{L}_t^{\text{CLIP}}(\theta)$ the clipping as in Equation 4; $\text{KL}\left[\pi_{\theta_{old}}, \pi_{\theta}\right]$ is the KL divergence between the old and new policy distributions; c_{KL} is a coefficient that scales the KL divergence penalty; $\mathbb{S}[\pi_{\theta}](s_t)$ is the entropy of the policy to encourage exploration; $\mathcal{L}_t^{\text{VF}}(\theta)$ is the value function loss; and c_{VF} and c_{ent} are coefficients for the value function loss and entropy term, respectively.

In summary, the PPO loss function $\mathcal{L}$ aims to optimise the policy by balancing reward maximisation with constraints on the extent of policy changes, using clipping or KL divergence to ensure stability and reliable learning. When these methods are combined, the policy updates are more controlled as clipping directly limit update magnitude and KL divergence ensures updates do not significantly alter the policy distribution.

2.3. Investigated learning schemes

As discussed before, the discount factor γ and learning rate α are two important hyperparameters that influence the convergence and success of the DRL agent to learn the policy π^*. The learning rate α dictates the magnitude of updates to model weights during back-propagation, controlling the specific updates to parameter vectors based on $\mathcal{L}$. This is mathematically represented by Equation 6, where the weights of parameter vector θ are adjusted:

$$\theta = \theta_{old} - \alpha_{\pi} \cdot \nabla_{\theta} \cdot \mathcal{L}_t(\theta) \quad (6)$$

The discount factor γ, on the other hand, determines the agent's priority to long-term rewards versus short-term rewards, as was shown in Equation 1.

This research explores four distinct learning schemes that manipulate these two hyperparameters in different ways to address the challenges posed by thermal inertia in collective heating systems. Each learning scheme is equally computationally expensive to train a DRL agent because the same data sets, training periods and simulator models are used (see Section 2.5).

I) **Fixed hyperparameter scheme:** In this scheme, hyperparameter tuning is employed during preliminary analyses to determine the best values of α (0.00005) and γ (0.99), which are kept constant throughout the five intervals of the training process. The hypertuning was done within the identified range of commonly used values for both the hyperparameters in Section 1.2. This approach serves as a baseline and is used in most of the DRL-related research. It provides a reference point for evaluating the effectiveness of the following adaptive schemes.

II) **Learning rate scheme:** This approach is used in some DRL training processes. Here, the γ is fixed at its best value from the hyperparameter tuning (0.99), while the α is reduced at the start of each training interval. The α decreases logarithmically from 0.001 to 0.00001 over the five training intervals: 0.001, 0.0005, 0.0001, 0.00005, 0.00001. The step size can be adapted based on the number of training intervals. This gradual reduction helps the agent to start with more significant updates at the start of learning and to fine-tune the policy as it approaches convergence.

III) **Discount factor scheme:** In this scheme, the α is fixed at its best value from the hyperparameter tuning (0.00005), while γ increases at the start of each training interval. In this research, γ increases from 0.8 to 0.99 over the five training intervals: 0.8, 0.85, 0.9, 0.95, 0.99. This allows the agent to first focus on short-term rewards and gradually shift its focus to long-term strategies as it becomes more familiar with the environment. The fixed learning rate ensures stable updates to the policy function.

IV) **GALER scheme:** This hybrid scheme combines elements of both the learning rate and discount factor schemes. Initially, the agent uses a small γ (0.8) to learn short-term dynamics and a large α (0.001). As training progresses, γ increases to 0.99, while α decreases to 0.00001.

With the GALER scheme, we propose a new learning approach, potentially effective in complex environments with delayed feedback. Early training emphasises short-term rewards by using a larger learning rate and smaller discount factor, which enables efficient exploration and quickly identifies basic patterns, such as the trade-off between temperatures and operational cost. As the training progresses, the learning rate is reduced while the discount factor is increased, slowing down the policy updates to be more attentive to long-term rewards. The agent will, therefore, make finer improvements in strategies and learn advanced system dynamics, such as thermal inertia or optimal recharging of decentralised DHW storage systems. This helps avoid overfitting to immediate rewards in complex environments.

Figure 1 visualises the main principles in PPO (see also Section 2.2) and how the α and γ influence the learning procedure (shown in red in the figure). The α affects the minibatch updates of both π and $V(s)$, while the γ influences values in $V(s)$ and in GAE. Although the learning rate for π and $V(s)$ can be differ-

ent, in this research they are the same in the proposed learning schemes.

2.4. Overview of experiments on RL configurations

In this study, the DRL agent aims to optimise the control of supply temperature (T_{sup}) in a CHS during the heating season. Each of the five training intervals in the training period lasts three months (from November 1 to January 31), which corresponds to the main part of the heating season. This period provides a rich set of data on system behaviour and thermal dynamics, essential for effective training. At the beginning of each training interval, hyperparameters are adjusted according to the specific learning scheme being investigated. After training, the agent's performance is evaluated in February, a month that is not included in the training data.

For all experiments, five rollout workers interact with the environment in parallel, gathering experience during training. After each training episode, which is set at 7 days, the policy (actor network) and value function (critic network) are updated using mini-batch gradient updates from the trajectory data (state-action-reward sequences) collected by all five workers. Once the neural networks are updated, the workers continue interacting with the environment using the updated policy, gathering more data based on the current policy (as PPO is an on-policy method). In this way, the policy is continuously refined, with each worker exploring different actions to collect diverse experiences for future updates.

To evaluate the efficacy of the different learning schemes, a comprehensive set of experiments is designed. These experiments include two optimisation objectives (energy-based and price-based), and three distinct CHSs within newly constructed apartment buildings comprising 24 dwellings ($n_{build} = 24$). The performance differences are expected to be larger for the price-based optimisation, since an additional time-aspect is added due to the variable tariff structure for electricity. Table 1 gives an overview of these variations.

2.4.1. Considered collective heating systems

Each CHS can be represented by the general energy balance equation given in Equation 7:

$$C_{sys}\frac{dT(t)}{dt} = \dot{Q}_{in}(t) - \dot{Q}_{out}(t) \quad (7)$$

where C_{sys} is the system's total thermal inertia [kJ/K], $dT(t)$ is the temperature difference in °C over time interval dt (seconds), $\dot{Q}_{in}(t)$ is the incoming thermal power [kW], and $\dot{Q}_{out}(t)$ the outgoing thermal power [kW] at time t.

As shown in Equation 7, there is a relationship between the system's thermal inertia, C_{sys}, temperature changes, time, and thermal power. For the DRL agent controlling the central T_{sup} over a certain time interval with a given thermal power, the complexity increases when the range of possible T_{sup} set points increases widens and/or as C_{sys} increases. Therefore, the four different learning schemes are evaluated by training DRL agents on three different collective heating systems (concept A, B, and C), each with progressively increasing thermal inertia. This increase is determined by various components within the CHS, such as heat emitters, production units, distribution pipes, and storage tanks. Concept C represents the most complex system, where a wide range of T_{sup} set points are available, and the agent is responsible for managing both indoor thermal comfort and DHW comfort.

In all concepts, newly constructed apartment buildings with 24 dwellings are considered, all equipped with underfloor heating systems with design temperatures of 35°C/30°C. Each underfloor heating system has a passive mixing valve, capping the inlet temperature at 35°C if the central T_{sup} exceeds this value. The primary heat source is a geothermal heat pump (GHP) connected to a thermal energy storage unit with two temperature sensors. These sensors are set to the minimum of the current T_{sup} set point and 45°C. The geothermal source is a borehole thermal energy system. In concept C, a gas boiler is added in series with the GHP to achieve higher temperatures (up to 65°C) required for DHW production. This series configuration is preferred when using a GHP as the primary heat generator [37].

Figure 2 gives a schematically representation of the three considered CHSs, labeled as concept A, B and C. In this figure, the thermal inertia is symbolised by electrical analogy. It also shows where the DRL agent controls the T_{sup} and that this action affects the operation of the central storage tank, central GHP and potential central gas boiler, as they follow the current T_{sup} set point. A detailed description on the three schemes is as follows:

A) **Collective space heating system:** In this concept, all 24 dwellings are connected via a 2-pipe system to a central GHP that serves only the SH demand [29]. This system has the lowest overall thermal inertia because DHW is produced separately within each dwelling. As a result, the DRL agent's control action only affects SH comfort, energy use and operational costs. The thermal inertia to be considered by the DRL agent includes the distribution pipes, the emitters in each dwelling, GHP and central thermal energy storage. Additionally, the range of T_{sup} is limited to 45°C due to the low-temperature underfloor heating systems, further reducing the impact of thermal inertia on the DRL agent's performance compared to systems with larger temperature changes.

B) **2-pipe system with decentralised booster heat pumps (BHP) for DHW:** This system increases the overall thermal inertia because the central GHP now serves as the heat source for both SH and for decentralised BHPs that produce DHW locally in each dwelling [38, 39]. While the DRL agent still fully controls SH, its influence on DHW is indirect, as the BHPs are activated by rule-based controllers. However, the distribution temperature managed by the DRL agent directly affect the Coefficient of Performance (COP) of the BHPs and GHP. A higher T_{sup} improves the COP of BHPs, but reduces the central GHP its COP. The T_{sup} range is constrained by the maximum allowed BHP source temperature, limiting the control possibilities for the DRL agent.

C) **2-pipe system with decentralised DHW storages:** This concept presents the highest thermal inertia due to the con-

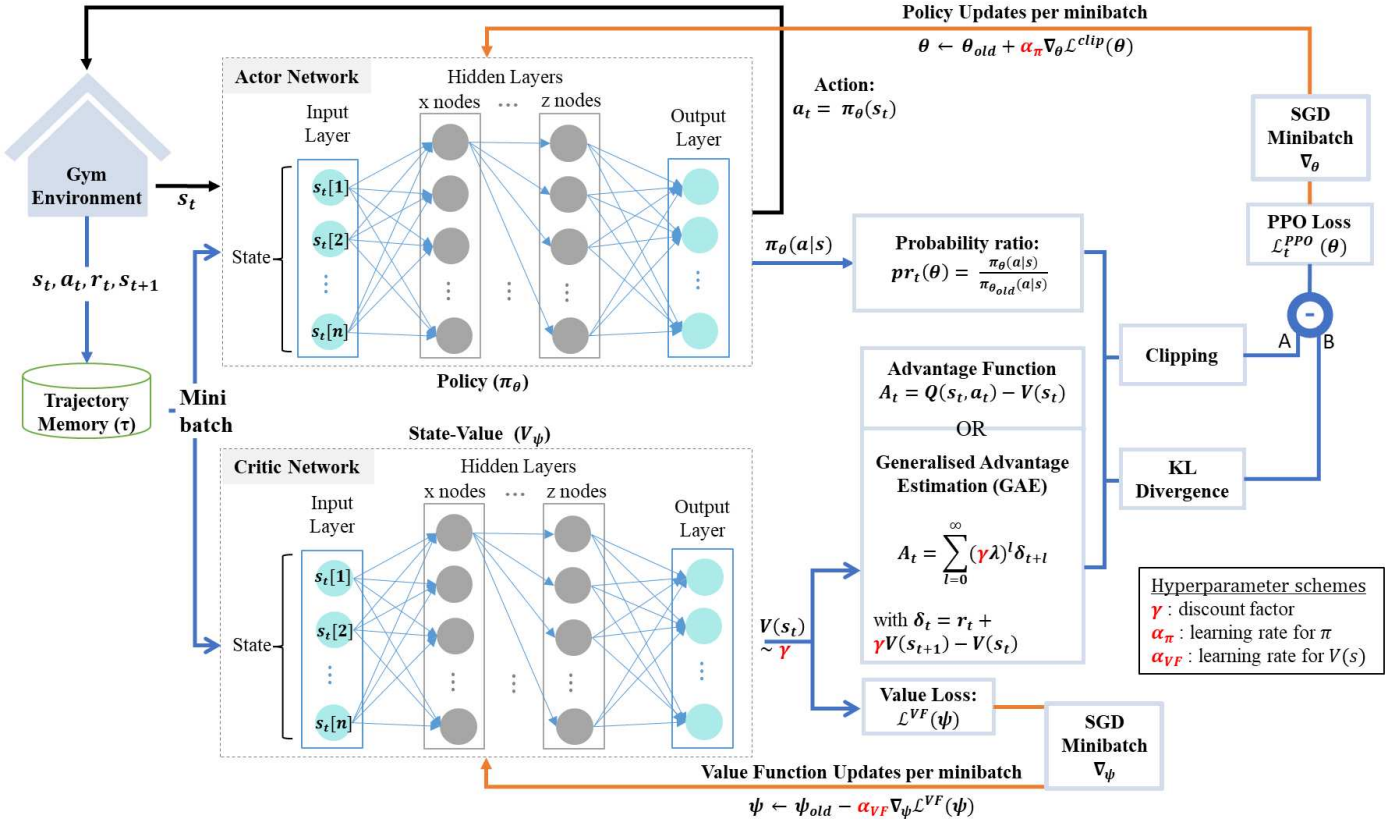

Figure 1: Schematic presentation of the principles of actor-critic PPO and how the α and γ affect the learning of the DRL agent.

Table 1: Overview of experiments. The four learning schemes (I, II, III, and IV) are tested in three environments, where concept A has the lowest thermal inertia impact and concept C the highest. Finally, the objective function is adapted, where the DRL agent should always take account of thermal comfort, but in the first experiments, the energy use should be as low as possible, afterwards, the focus lies on minimising operation costs.

	#	Analysed variants			
Learning scheme	(4):	I: fixed	II: adaptive α	III: adaptive γ	IV: GALER
CHS	(3):	A: Coll. SH	B: 2-pipe BHP		C: 2-pipe storage
Objective	(2):	Energy-based		Price-based	

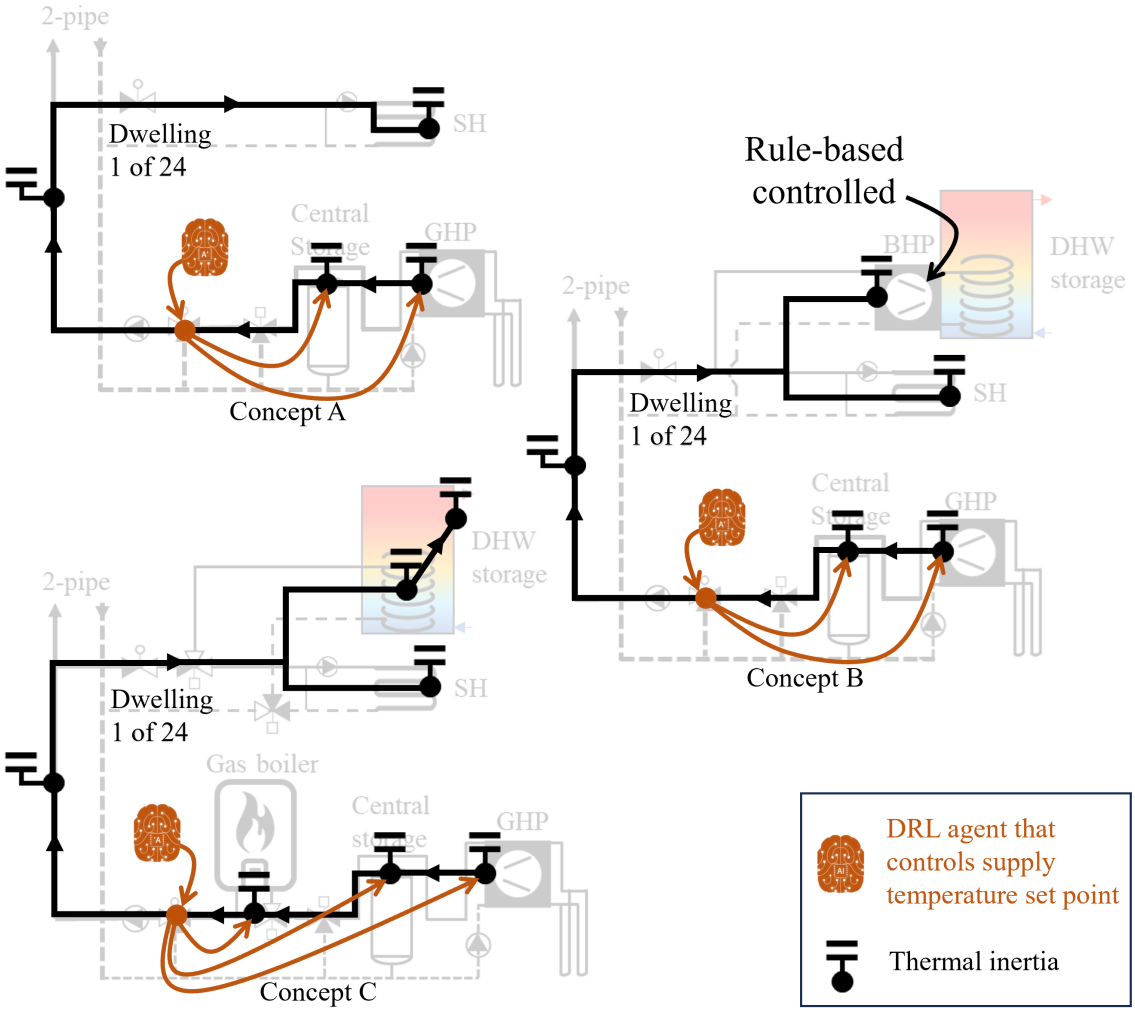

Figure 2: Schematic overview of the three CHSs showing the thermal inertia affecting the DRL agent (represented as capacitance's by electrical analogy). **(A)** collective space heating system, lowest thermal inertia with the lowest thermal inertia; **(B)** 2-pipe system with decentralised booster heat pumps; and **(C)** 2-pipe system with decentralised DHW storage tanks, with the highest thermal inertia. The DRL agent controls the T_{sup} in all concepts. The central GHP, storage and gas boiler (in concept C) follow the DRL agent's imposed set point.

nection of decentralised DHW storage tank and an underfloor heating system for SH in each dwelling to the 2-pipe distribution system [40]. The central DRL agent controls both SH and DHW comfort, as well as energy use and costs. Managing DHW introduces additional challenge because the agent must decide when to increase T_{sup} above 55 °C to charge the DHW storage tanks. This is particularly challenging given the infrequent nature of DHW consumption (only 3% of the day [41, 42]) and the extended charging times of the storage tanks. The system also incorporates a gas boiler in series with the GHP to provide the higher temperatures required for DHW. The range of T_{sup} is extended to 65 °C, increasing the impact of thermal inertia on the DRL agent's decision-making process.

2.4.2. MDP formulations

To formalise the DRL agent's decision-making problem, the problem is framed as an MDP, capturing the environment's dynamics. This Section is an outline of how the MDP components are structured for energy-based and price-based optimisation scenarios.

Parameters in state space: The state space design provides the DRL agent with essential information for effective decision-making. The core states $s_t[1-6]$ are included in all experiments, while additional states ($s_t[B1]$, $s_t[C1]$, and $s_t[C2]$) are specific to certain system concepts. The state space is further tailored based on the optimisation objective: energy-based ($s_t[E1]$) or price-based ($s_t[P1]$, $s_t[P2]$).

1. General observations for all experiments:
   - $s_t[1]$: Outdoor temperature [°C]
   - $s_t[2]$: Average indoor operative temperature error [°C]
   - $s_t[3]$: T_{sup} at central production [°C]
   - $s_t[4]$: T_{sup} at top of distribution pipe [°C]
   - $s_t[5]$: Time of the day [h]
   - $s_t[6]$: Central mass flow rate [kg/s]
2. Concept-specific observations:
   - Concept B: $s_t[B1]$, number of working BHPs
   - Concept C: $s_t[C1]$, number of cold DHW storages according to upper sensor; $s_t[C2]$, number of cold DHW storages according to bottom sensor
3. Optimisation-specific observations:
   - Energy-based: $s_t[E1]$, total primary energy use [kWh]
   - Price-based: $s_t[P1]$, hourly electricity price of the next 12 hours [€/MWh]; $s_t[P2]$, hourly gas price of the next 12 hours [€/MWh], only for concept C.

Including $s_t[3]$ and $s_t[4]$ helps the agent understand distribution delays and system response time, while $s_t[1]$ and $s_t[2]$ provide insight necessary to maintain indoor thermal comfort. The inclusion of states $s_t[B1]$, $s_t[C1]$, and $s_t[C2]$ allows the agent to account for DHW comfort and the additional system thermal inertia in concepts B and C. Concept A requires no additional state, since the states for SH are general.

State $s_t[5]$, representing the time of day, enhances the agent's ability to anticipate typical demand patterns and plan its actions accordingly. State $s_t[6]$ indicates periods of high or low demand, helping to manage system load efficiently. In energy-based optimisation, $s_t[E1]$ is essential for minimising primary energy use. Similarly, for price-based optimisation, including electricity price data for the next 12 hours ($s_t[P1]$) and, if applicable, the gas price data ($s_t[P2]$) enables the agent to reduce operational costs.

Action spaces: The discrete action space of the DRL agent is defined by the set points for T_{sup}, while other actuators are managed by rule-based controllers. The action time interval of the DRL agent (Δt_{action}) to select a specific temperature value is set at 10 minutes. The actions are constrained to ensure T_{sup} remains within safe limits, preventing any disruption to the functioning of the system.

In concept A, the action space includes 20°C, 25°C, 30°C, 35°C, 40°C, and 45°C, a typical range of T_{sup} for underfloor heating systems. For concept B, the 45°C option is excluded from this set of temperatures to avoid exceeding the maximum inlet temperature of the decentralised BHPs (42°C). In concept C, a 65°C option is added to concept A's action space to facilitate recharging the decentralised DHW storage tanks.

Reward function: The reward function (Equation 8) is a weighted sum that balances four key objectives, each normalised between 0 and 1: minimising the OPEX, reducing Primary Energy (PE) use, lowering indoor thermal discomfort, and ensuring DHW comfort. These metrics are evaluated over the last 10 minutes, corresponding to the agent's action interval, Δt_{action}.

$$
\begin{aligned}
r_t \;= f_1 \cdot & \frac{OPEX_t - OPEX_{t,max}}{OPEX_{t,min} - OPEX_{t,max}} \\
+ f_2 \cdot & \frac{PE_{t,max} - PE_t}{PE_{t,max}} \\
+ f_3 \cdot & \frac{1}{1 + e^{c_1(\widehat{SH}_t - c_2)}} \\
+ f_4 \cdot & \frac{1}{1 + e^{c_3(\widehat{DHW}_t - c_4)}}
\end{aligned}
\tag{8}
$$

The OPEX term, $OPEX_t$, reflects the energy costs incurred from electricity and gas usage accounting the variable tariff structures. The values $OPEX_{t,max}$ and $OPEX_{t,min}$ represent the theoretical maximum and minimum costs over the same period, based on the highest or lowest possible energy use of the system's production units. Note that these minimum and maximum values can be negative in case of negative tariffs. Similarly, PE_t and $PE_{t,max}$ are computed using primary energy conversion factors: factor 1 for gas and 2.5 for electricity, according to the Belgian energy market. For energy-based optimisation, $f_1 = 0$, and in case of price-based optimisation, $f_2 = 0$.

For comfort-related terms, a Sigmoid function is employed. The indoor thermal comfort, $\widehat{SH}_t$, is expressed through the average Room Temperature Lack (RTL) across all dwellings, see Equation 12 for more details. This RTL is here normalised to

a scale where the minimum and maximum RTL values correspond to 0 and $0.5 \cdot n_{build} \cdot \frac{\Delta t_{action}}{3600 s/h}$, respectively. The parameters c_1 and c_2 are 0.1 and 50, respectively.

The DHW comfort term, $\widehat{DHW}_t$, varies depending on system concept:

- Concept A: This term is omitted ($f_4 = 0$) because DHW is not supplied by the CHS, but is produced separately.
- Concept B: $\widehat{DHW}_t$ is the average time DHW temperature falls below 40°C during consumption (as defined by the $t_{DHW;\,dc}$ KPI of Section 2.5.3). Parameters c_3 and c_4 are set to 1.5 and 5, respectively. If no DHW is consumed across all dwellings in the past 10 minutes, this term defaults to 1; otherwise, its value is adjusted based on the DHW use simultaneity factor S (Equation 9)[43].
- Concept C: $\widehat{DHW}_t$ correlates with the state $s_t[C2]$, representing the percentage of DHW storage tanks requiring recharging. If no recharging is necessary and the agent sets T_{sup} to 65°C, this term is 0. If T_{sup} is lower, then $C_3 = -0.1$ and $C_4 = 30\%$. Conversely, if recharging is necessary and the agent sets T_{sup} to 65°C, $C_3 = 0.1$ and $C_4 = 40\%$. If T_{sup} is lower, this term is 0.

$$S = \frac{1}{\sqrt{n_{build} - 1}} + 0.17 \tag{9}$$

Since each term is normalised between 0 and 1, the sum of the weights f_1, f_2, f_3, f_4 should equal 1 to maintain a balanced trade-off between the different objectives in the reward function.

2.5. Simulation-based evaluation framework

The evaluation framework employed in this study is adapted from previous research [39] to systematically compare the performance of different control strategies. The framework is comprehensive, encompassing boundary conditions, simulation models, and key performance indicators (KPIs) that collectively provide a robust basis for assessing the efficacy of the DRL agents in high-thermal inertia environments.

2.5.1. Boundary conditions of case study

Weather profile: The weather files contain outdoor temperature measurements [°C], and solar radiation measurements [W/m²] for each cardinal direction (North, East, South and West) to account for the spatial variability of solar gains on the building envelope. To ensure temporal consistency with the simulation time step, linear interpolation is applied when the resolution of weather data does not align with the simulation's 10-second time step. For training, an average weather profile representing Belgian climate conditions from 2001 to 2020 was used [44]. For testing was conducted using specific weather data from the year 2009 in Uccle, Belgium [45]. In these weather files, November until January were used for training, while February is used for testing. This approach ensures that the DRL agents are evaluated under different circumstances than their training data.

Energy tariff structures: The energy tariffs used in the simulations are essential for evaluating the cost-effectiveness of the control strategies. For electricity, the day-ahead market prices from 2022 in Belgium were utilised, reflecting real-time hourly variability in energy costs. Since these prices are published the day before, these can be used by DRL agents to optimise their planning. Gas prices were modeled using the TTF103 price signal of 2022, ensuring that the economic performance of the DRL agents is assessed against current market conditions. These tariffs are expressed in €/kWh and are applied uniformly across all simulations.

Building characteristics: The simulated apartment building is a modular assembly comprising 24 dwellings with 12 distinct configurations. Each dwelling is characterised by its thermal properties, including transmission losses, ventilation rates, and solar gains. The building's envelope is designed with a U-value of 0.24 W/m²K for walls and HR++ windows with a U-value of 1.1 W/m²K and a g-factor of 0.6 [46]. These parameters are aligned with current Flemish building regulations for new constructions. The floor areas of the dwellings range from 88 to 104 m², with window areas constituting 21% of the floor area, oriented towards different cardinal directions to account for varying solar exposure. Ventilation is provided by a type D mechanical balanced system with 80% heat recuperation efficiency, bypassed when the outdoor temperature exceeds 16°C. These specifications result in design heat loads ranging from 1.5 to 2.4 kW for an indoor temperature set point of 21°C and outdoor temperature of -8°C.

Occupant profiles: Occupant behaviour is modeled using the stochastic profile generator developed in the TETRA-SWW and Instal2020 projects [47, 48]. These profiles include internal heat gains [W], occupancy presence, DHW consumption [kg/s] at 60°C, and indoor temperature set point schedules. The profiles in a specific building are linked to each other, ensuring realistic and consistent simulation of occupant-driven energy use.

This generator allows to choose for each dwelling between nine different family types, ranging from 1 to 4 inhabitants, and the number and type of tapping points [49]. These nine family types result from a survey held on 700 dwelling in Belgium during the Instal2020 project. The 24 families living in the apartment building of this research are a mix of these nine family types.

2.5.2. Simulator models

The dynamic simulation environment, implemented in Python, is designed to replicate the thermal dynamics of CHSs. The models in the simulator are grounded in previous work Van Riet [37], Jacobs et al. [40, 39] and capture the transient thermal behaviour of system components using first-order, linear, ordinary, and non-homogeneous differential equations, as was also done in [50, 51]. The general form of these equations is expressed as:

$$\frac{dy(t)}{dt} = -a(t)y(t) + b(t) \tag{10}$$

where $y(t)$ represents the integrand, t is time, and $a(t)$ and $b(t)$ are constants within each time step (Δt). The explicit solution

of these equations, based on a zero-order hold, ensures accurate effects of time delays on domestic hot water (DHW) and indoor thermal comfort.

Each dwelling is modeled as a thermal zone with three temperature nodes: emitter surface temperature, indoor air temperature and indoor wall surface temperature. These nodes are interconnected through thermal capacities and resistances, accounting for heat exchanges via transmission losses, ventilation, internal heat gains and solar gains. The model's granularity is enhanced by dividing emitters into nine segments, following the methodology outlined in [40, 37].

The stratified storage tank model, adapted from TRNSYS Type 60 [52] and described in [37, 40], simulates heat transport and stratification using partial differential equation in temperature and along the height [53]. It assumes a number of homogeneous volume layers with a uniform temperature. The captured thermodynamics include conduction, advection, heat losses, and heat gains from a potential internal coil heat exchanger. A temperature-inverse algorithm was added to account for temperature-dependent water density.

In case of a thermal energy storage containing technical water, water flows through the top or bottom of the storage tank to represent charging and discharging, respectively. In case of DHW storages, cold domestic water of 10°C enters the tank at the bottom during a DHW demand and the DHW extraction at the top. The sizing of the internal coil heat exchanger in DHW storage tanks was fitted to laboratory measurements and manufacturer specifications [43].

The heat production dynamics are governed by the following equation (11):

$$C_{prd}\frac{dT_{out}}{dt} = \dot{Q}_{prd} - UA_{prd}\cdot(T_{out}-T_{amb}) - \dot{m}_{prd}\cdot c_p\cdot(T_{out}-T_{in}) \quad (11)$$

where C^{prd} is the thermal capacity [J/K] at temperature T_{out} [°C], $\dot{Q}_{prd}$ is the source heat [W], UA_{prd} is the heat transfer coefficient [W/K] to the surroundings (T_{amb} set at 20 °C), c_p is the distribution medium's specific heat capacity equal to 4187 J/kgK for water-based systems, and $\dot{m}_{prd}$ is the mass flow rate through the production unit [kg/s], entering at temperature T_{in} [°C].

The $\dot{Q}_{prd}$ of different production units are fetched using non-linear regression fits based on manufacturer data for various temperatures. The GHP and gas boiler are as in [37], and the data for the BHP models is based on *Alpha Innotec WWB21 of Nathan Systems*, with separate lumped capacities for the condenser and evaporator [39].

The pump operations are simplified by assuming mass flow availability within the nominal value and 10% of this value. Time delays of control valves are modeled with a time constant of 32 seconds [37] and the mixing rule is applied for three-way valves and mixing points. The time delay in the pipes and distribution losses are captured using the plug-flow model and RC-model, respectively [54].

2.5.3. Key Performance Indicators

The evaluation framework employs four primary Key Performance Indicators (KPIs) to assess the performance of DRL agents trained with varying hyperparameter schemes. These KPIs are chosen to balance between occupant comfort, energy efficiency, and economic cost.

Thermal comfort-related KPIs:

1. The average Room Temperature Lack (RTL) [Kh/day]: This metric, defined in [37], quantifies the indoor thermal discomfort in dwellings by measuring the deviation of the operative indoor temperature (T_{op}) from the set point temperature ($T_{op;SP}$). The operative temperature represents the perceived temperature by inhabitants due to convection and radiation. The KPI calculation is as follows:

$$RTL = \frac{1}{n_{build}}\sum_{n=1}^{n_{build}}\left[\int_{t_1}^{t_2}(T_{op;SP}(n) - (T_{op}(n)+e_{tol}))_+ dt\right] \quad (12)$$

   where e_{tol} is the comfort tolerance, n_{build} is the number of dwellings, i.e. 24.

2. The relative duration of lacking DHW temperature ($t_{DHW;dc}$) [%]: This KPI measures the percentage of DHW consumption time during which the temperature is below 40°C, indicating user discomfort. The lower the percentage, the better the performance of the system in maintaining user comfort. In case of concept A, this KPI is not calculated.

Energy-related KPI:

3. Primary Energy use (PE) [kWh]: This KPI calculates the total primary energy consumed by the system, including electricity and fossil fuels. For electricity, a conversion factor of 2.5 is applied to account for the Belgian grid. This KPI is directly targeted by DRL agents in the energy-based optimisation.

Economic KPI:

4. Operational Expenses ($OPEX$) [€]: OPEX represents the total cost associated with energy consumption, calculated based on the tariff structures for electricity and gas defined in Section 2.5.1. This KPI is directly targeted by DRL agents in the price-based optimisation.

To determine the most effective DRL agent across multiple system KPIs, a holistic scoring method, adapted from [39], is used. This score integrates the KPIs, allowing for a balanced comparison of thermal comfort, energy use, and cost-effectiveness, by dividing each KPI of a specific DRL agent by its respective median value (denoted as μ) for normalised and dimensionless KPIs. The advantage of using the median in the normalisation method is that it preserves the relative differences in performance between distinct concepts.

Based on this normalisation methodology, the weighted sum of the KPIs equals the KPI^*, as shown in Equation 13 for DRL agent Z.

$$KPI^*(Z) = -\Big(f_1 \cdot \frac{OPEX(Z)}{\mu_{OPEX}} + f_2 \cdot \frac{PE(Z)}{\mu_{PE}} + f_3 \cdot \frac{RTL(Z)}{\mu_{RTL}} + f_4 \cdot \frac{t_{DHW;\,dc}(Z)}{\mu_{t_{DHW;\,dc}}}\Big) \quad (13)$$

where the weights f_1, f_2, f_3, and f_4 for DRL agent Z are the same as in the reward function of experiment Z (Equation 8).

Finally, the Volume Weighted Average Supply Temperature (VWAST) [°C] is used to give insight on effects of actions of DRL agents. This KPI is calculated according to Equation 14 showing the general tendency of the control actions.

$$VWAST = \frac{\sum_{i=1}^{n_{sim}} T_{sup;i} \cdot \dot{m}_i}{\sum_{i=1}^{n_{sim}} \dot{m}_i} \quad (14)$$

where $T_{sup;i}$ and $\dot{m}_i$ are the central supply temperature set point and flow rate through the 2-pipe distribution system at time step i, respectively, and n_{sim} the total number of time steps in the testing period (= 259200).

3. Results and discussion

3.1. Energy-based optimisation

The energy-based optimisation results, as presented in Table 2, offer an initial assessment of how the DRL agents, trained with different learning schemes, perform when controlling T_{sup} without considering OPEX (f_1 = 0). High discomfort values (≥ 6 Kh/day for RTL and ≥ 10% for $t_{DHW;\,dc}$) are given in red, while the overall performance (KPI^*) is colour-coded within each concept, with green the best performance and red the worst. For each concept, the weighting factors of the reward function (see Equation 8), which are also used for the KPI^* calculations, are provided in the respective grey rows. DRL agents trained for Concept A employ weights derived from previous research [29], while the weights for agents in concept B and C were optimised during preliminary tests to achieve the best performance of the DRL agent.

Since the DRL agents do not learn about variable prices, the optimisation possibilities are limited to adjusting temperature set points to improve thermal comfort at the lowest possible primary energy use. The results reveal that learning schemes II and III occasionally struggle to achieve this balance, as evidenced by higher thermal discomfort and low VWAST values. In contrast, scheme I and scheme IV do not always lead to the best performance, but are more consistent across different concepts.

Focusing on concept A, learning schemes II and IV exhibit the best KPI^* values of -0.72 and -0.75, respectively. Scheme III, however, fails to optimise properly, with an excessively high RTL of 19.25 Kh/day and a poor KPI^* of -6.74. There is a clear relation between higher VWAST and better-balanced objectives, though this comes at the cost of increased PE use.

Concept B requires the DRL agent to balance PE, RTL, and $t_{DHW;\,dc}$, with most focus on PE (f_2 = 0.6). Here, weighting factor f_4 is 0.1, since DHW comfort is mostly affected by the decentralised rule-based activation strategy of the BHP in each dwelling rather than the central DRL agent controlling the T_{sup}. Except for scheme II, all schemes perform similarly, with a narrow range of KPI^* values from -0.96 to -1.04, indicating the agent's consistent ability to balance the three objectives. However, the overall potential for optimisation is rather limited when not controlling the decentralised BHPs. Although, the GALER training scheme led to worse overall performance than schemes I and III, it still provided adequate RTL values at lower PE use. This indicates it was able to understand the environment dynamics.

In concept C, DHW comfort is emphasised by f_4 = 0.4, since this involves long-term thinking for the agent to foresee sufficient heat in the decentralised DHW storages. On the one hand, schemes I and IV perform well, achieving a KPI^* of -0.88 and -0.91, respectively. Scheme I, in particular, maintains a high T_{sup} as indicated by a VWAST of 60.7°C, with T_{sup} above 50°C for 70.8% of time. This reflects the priority given to DHW comfort, as the decentralised DHW storages can only be recharged when T_{sup} is above 50°C. On the other hand, scheme II exhibits poor performance, with a high $t_{DHW;\,dc}$ of 77.21% and a lower KPI^* of -17.79, indicating a failure to ensure DHW comfort. This suggests that the DRL agent trained with scheme II was not able to learn the complex dynamics of concept C.

In summary, the energy-based optimisation offers useful insights into the behaviour of DRL agents trained with the different schemes. While scheme I generally shows reliable performance across all concepts, schemes II and III sometimes fail to achieve acceptable results. The inconsistent performance of these schemes can be attributed to struggling in balancing long-term and short-term rewards. For instance, an increase in γ without decreasing α (scheme III) might result in too drastic updates for systems with smaller thermal inertia (concept A), while decreasing the learning rate α in scheme II without adapting the γ could cause a lack of understanding the long-term system dynamics in concept B and C that include DHW production. However, when the DRL agents are successful, the performance differences tend to be small, highlighting the limited optimisation potential in these scenarios without anticipating on the dynamic prices. These findings lay the foundation for the more complex price-based optimisation, where including the dynamic pricing into the state space and optimisation objectives offer more opportunities and a greater challenge for the DRL agent.

3.2. Price-based optimisation

Table 3 presents the KPIs for DRL agents trained to minimise OPEX while maintaining thermal comfort across three different concepts. The same lay-out is used as in Table 2. Given the electricity tariff is based on hourly day-ahead market prices, effective optimisation requires strategic planning of T_{sup} adjustments, using the system's thermal inertia to benefit from lower price periods. The same weighting factors for reward and KPI^* calculations were used as in Section 3.1.

Table 2: Energy-based optimisation: overview of KPIs for testing the trained central PPO agent controlling the T_{sup}. Each row represents a different learning scheme for a specific concept (A, B, C), and the first row of each concept indicates the used weights in the reward function (Equation 8) and the KPI^* calculation (Equation 13). The higher the KPI^*, the better the DRL agent learned to balance the different objectives. The VWAST represents the weighted average T_{sup} set by the agent.

		KPIs in KPI^*					
Con-cept	Learning scheme	OPEX [€]	PE [kWh]	RTL [Kh/day]	$t_{DHW;\,dc}$ [%]	KPI^* [-]	VWAST [°C]
A		$f_1 = 0$	$f_2 = 0.3$	$f_3 = 0.7$	$f_4 = 0$		
	I	336.07	5109.78	2.85	-	-1.26	28.8
	II	393.54	5882.68	1.16	-	**-0.72**	33.8
	III	238.2	3590.32	19.25	-	-6.74	24.8
	IV	376.28	5605.2	1.27	-	-0.75	33
B		$f_1 = 0$	$f_2 = 0.6$	$f_3 = 0.3$	$f_4 = 0.1$		
	I	636.75	9337.92	1	2.12	-0.963	35.7
	II	473.28	7208.85	15.6	2.27	-4.4	26.3
	III	626.28	9047.6	1.03	2.32	**-0.96**	33.5
	IV	608.87	8882.4	1.42	2.28	-1.04	32.2
C		$f_1 = 0$	$f_2 = 0.4$	$f_3 = 0.2$	$f_4 = 0.4$		
	I	1806.05	17097.22	1	1.22	**-0.88**	60.7
	II	988.37	10168.25	4.64	77.21	-17.79	31.7
	III	1577.75	15461	1.06	2.34	-1.09	53.4
	IV	1757.44	16795.14	1.02	1.37	-0.91	57.1

Table 3: Price-based optimisation: overview of KPIs for testing the trained central PPO agent controlling T_{sup}. Each row represents a different learning scheme for a specific concept (A, B, C), and the first row of each concept indicates the used weights in the reward function (Equation 8) and the KPI^* calculation (Equation 13). The higher the KPI^*, the better the DRL agent learned to balance the different objectives. The VWAST represents the weighted average T_{sup} set by the agent.

		KPIs in KPI^*					
Con-cept	Learning scheme	OPEX [€]	PE [kWh]	RTL [Kh/day]	$t_{DHW;\,dc}$ [%]	KPI^*	VWAST [°C]
A		$f_1 = 0.3$	$f_2 = 0$	$f_3 = 0.7$	$f_4 = 0$		
	I	345.51	4932.33	2.71	-	-1.07	28.3
	II	346.68	5128.24	2.29	-	-0.95	29.8
	III	180.41	2926.11	38.6	-	-10.97	21.4
	IV	331.05	5278.2	1.97	-	**-0.85**	30.3
B		$f_1 = 0.6$	$f_2 = 0$	$f_3 = 0.3$	$f_4 = 0.1$		
	I	640.87	8894.95	1.1	2.31	-1.02	32.4
	II	579.95	8669.59	1.8	2.28	-1.15	30.4
	III	623.85	9134.82	1.13	2.16	-1.01	34.1
	IV	611.53	9077.44	1.07	2.09	**-0.98**	33.1
C		$f_1 = 0.4$	$f_2 = 0$	$f_3 = 0.2$	$f_4 = 0.4$		
	I	1361.35	13600.3	1	12.44	-1.21	30.1
	II	889.91	9366.36	3.68	91.04	-5.73	33.3
	III	1571.11	15388.42	1.15	2.8	-0.79	48
	IV	1764.82	16795.21	1	1.41	**-0.74**	57.8

The most consistent observation across all concepts is the superior performance of the GALER training method (learning scheme IV) in balancing the objectives. This is indicated by the lowest KPI^* values of -0.85, -0.98, and -0.74 for concepts A, B and C, respectively. This highlight the ability of DRL agent trained with the GALER approach to prioritise long-term objectives more effectively than the other learning schemes. Specifically, in concept A, the GALER approach resulted in lower OPEX and larger PE use. This indicates the GALER learning scheme resulted in a DRL agent that effectively raised T_{sup} during low-price periods and lowered it during high-price periods. Despite the better balancing, it generally results in higher VWAST, indicating that the T_{sup} is generally higher during high demand times than for the DRL agents trained with other learning schemes, which is also reflected in the thermal comfort.

Focusing on concept A, where the focus lies on indoor thermal comfort ($f_3 = 0.7$), learning scheme IV outperformed schemes I and II, achieving a lower OPEX of €331.05 compared to €345.51 and €346.68, respectively. Additionally, the average RTL was also improved, with scheme IV achieving 1.97 Kh/day compared to 2.71 and 2.29 for schemes I and II, respectively. This means the GALER scheme saved 4% of costs in combination with an 27% increase in thermal comfort, compared to the fixed learning scheme, while it also used the most PE. The increased energy use results from the DRL agent's strategic actions: adjusting the T_{sup} in anticipation of future price trends by setting it in the opposite direction of expected price changes. Figure 3 illustrates the decision-making of this DRL agent, where the black arrows highlight this observed trend.

In contrast, scheme III, which used an increasing γ and a fixed α, failed to deliver adequate thermal comfort, with a high RTL of 38.6 Kh/day/dwelling, due to maintaining a consistently low T_{sup} (VWAST = 21.4°C). This highlights a key drawback of neglecting thermal inertia in complex environments: optimising only for immediate cost savings can lead to undesirable outcomes, such as uncomfortably low indoor temperatures.

In concept B, the DRL agent trained according to schemes I, II, and IV has very similar performance with KPI^* values ranging from -1.02 to -0.98. This reflects a similar balance of the different objectives in the reward function. Indeed, the differences in OPEX, RTL and $t_{DHW;\,dc}$ are less pronounced in concept B due to the presence of rule-based controllers that activate the decentralised BHPs. This constrained the DRL agent's optimisation capabilities to only a small improvement in BHP's efficiency and a minor influence on $t_{DHW;\,dc}$. Nonetheless, the performance of the GALER approach again suggests that the combined adjustment of learning rate and discount factor allowed for more nuanced and effective learning to target this small optimisation potential, as it resulted in the lowest OPEX and highest thermal comfort values.

In concept C, the reliance on a gas boiler, which operates at a fixed tariff of €114.04/MWh and has lower efficiency than BHPs, results in significantly higher OPEX and PE use. Despite these higher costs, the decentralised DHW storages and the wider range of possible supply temperatures lead to more pronounced performance differences between the learning schemes. In particular, scheme II exhibits a $t_{DHW;\,dc}$ of 91.04% of tap time, which is comparable to the DHW discomfort of the DRL agent trained with scheme II for the energy-based optimisation in concept C. This indicates that a decreasing learning rate α with fixed γ confuses the DRL agent and limits its ability to capture the slow responsiveness of this complex environment. In contrast, the GALER scheme achieves the highest KPI^* value of -0.74. Adjusting the γ and α allows the DRL agent to efficiently manage the T_{sup} to minimise the OPEX while providing acceptable levels of DHW comfort.

These results underscore the importance of using adaptive learning rate and discount factor schemes for DRL agents to manage the supply temperature of complex thermal systems. Balancing operational cost, thermal comfort, and energy efficiency requires dynamic approaches, and the GALER scheme demonstrates the best performance, improving KPI^* by 3% to 15% compared to the other schemes, excluding the outliers where DRL agents violated thermal comfort constrains.

3.3. Analysis of the learning process

Figure 4 provides further insights into the training process of the DRL agent trained with learning schemes IV and II in concept C and the price-based optimisation objective. The x-axis shows the training time steps, which are divided into five training intervals, each representing 91 days (from November 1 to January 31), marked in different colours. The y-axis represents in **a)** the total loss ($\mathcal{L}_t^{\text{PPO}}(\theta)$), in **b)** the maximum cumulative reward achieved during a 7-day episode by one of the five parallel rollout workers in each training interval, and in **c)** the mean cumulative reward across all five workers during the 7-day episodes in each training interval.

To assess the success and convergence of the trained agents, the $\mathcal{L}_t^{\text{PPO}}(\theta)$ should stabilise around zero, indicating minimal deviation from the old policy. Meanwhile, both the maximum and mean cumulative rewards should increase across the training intervals, reflecting the agent's ability to discover a better policy.

In graph **a)**, the $\mathcal{L}_t^{\text{PPO}}(\theta)$ for the GALER scheme (learning scheme IV) converges quickly towards zero, indicating small policy updates and effective convergence of the policy. In contrast, for scheme II, the loss remains high in each episode, indicating instability which was also reflected in poor overall performance ($KPI^* = -5.73$). The fixed γ for a DRL agent trained for concept C likely contributes to this instability, as it is crucial for the agent to learn the system dynamics in order to effectively manage T_{sup} for recharging decentralised DHW storages. This becomes even more complex with a variable price structure. This indicates that the γ should not be fixed for cases involving wide temperature ranges and large responsibility for the central DRL agent.

Additionally, the mean (graph **b)**) and maximum (graph **c)**) cumulative reward during the 7-day episodes support this conclusion. Both rewards increase throughout the training process, with the GALER scheme (learning scheme IV) consistently achieving higher cumulative rewards than scheme II. This indicates, the DRL agents find better policies. However, a decline in the mean reward is observed within each training in-

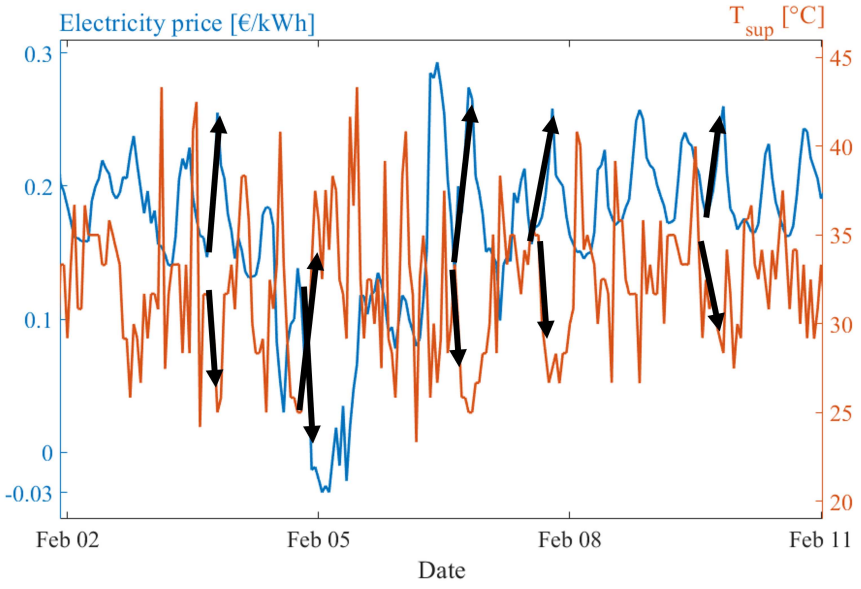

Figure 3: Price-based optimisation: the hourly electricity price (in blue on left axis) and T_{sup} (in orange on right axis) are visualised for DRL agent trained with the GALER scheme in concept A. It clearly shows that the DRL agent consistently sets lower supply temperatures during high-price times and vice versa.

terval (starting in November and ending in January) for both learning schemes. This can be attributed to the increasing heat demand over this period: in November, the total heat demand is 8.9 MWh for SH and 2.9 MWh for DHW, while in December and January, it rises to 13.2 MWh and 13.3 MWh for SH and 3.0 MWh and 3.2 MWh for DHW, respectively. As the heat demand increases, it becomes more challenging for the DRL agent to maintain high rewards, particularly as higher T_{sup} are required to comply with thermal comfort requirements. Consequently, the reward component related to primary energy use or OPEX often drops to zero.

3.4. Influence of reward weights in price-based optimisation

In the previous analyses, the reward weights were selected from preliminary tests to find the overall best performing DRL agent. However, such weights are user-dependent. In order to examine the effects of this variability, the following discussion establishes a wider range of reward weights. Figure 5 illustrates the performance of the four learning schemes (I, II, III, IV) applied to controlling T_{sup} in concept C under a price-based optimisation objective ($f_2 = 0$). A total of 19 variations in reward weights assigned to different terms in the reward function are analysed to evaluate their impact on learning performance. Rows represent the weight assigned to OPEX reduction (f_1), while columns correspond to f_4, the weight assigned to DHW comfort. The weight for SH comfort (f_3) is computed as $f_3 = 1 - f_1 - f_4$. Since indoor thermal comfort is generally easier to maintain than cost reductions or DHW requirements, f_3 remains between 0 and 0.4, except in one extreme case where $f_3 = 1$. Light blue-coloured boxes indicate a $t_{DHW;dc} \geq 7.5\%$.

In Figure 5**a)**, which presents results for smaller f_4 values ($f_4 \in [0, 0.4]$), the predominance of blue boxes indicates that assigning a low weight to f_4 often leads to DHW discomfort. Specifically, for $f_4 \leq 0.2$, 22 out of 28 trained DRL agents exhibit DHW discomfort for more than 7.5% of the tap time. This demonstrates that the reward function effectively reflects the desired objectives but underscores the necessity of properly balancing the reward weights to satisfy all comfort boundaries.

In Figure 5**b)**, showing higher f_4 values ($f_4 \in [0.5, 1]$), learning schemes III and IV generally outperform schemes I and II. For $f_4 \geq 0.3$, the GALER scheme (scheme IV) performs best in 5 out of 12 cases, while scheme III outperforms in the remaining 7 cases. However, learning scheme III performs worst when $f_1 = 0.2$ and $f_4 = 0.4$, where the high RTL of 9.37 Kh/day and second-highest $t_{DHW;dc}$ (only scheme II performs worse) lead to suboptimal results. In scenarios where scheme IV does not perform best, its performance remains close to the performance of the best-performing scheme, except in only two cases:

1. When $f_1 = 0.5$ and $f_4 = 0.3$, where the DRL agent fails to maintain $t_{DHW;dc}$ below 7.5%.
2. When $f_1 = 0.2$ and $f_4 = 0.6$, where the high RTL of 3.3 Kh/day reduces the overall performance, even though $t_{DHW;dc}$ (3.5%) is the best among all schemes for these weights.

While varying hyperparameter schemes are uncommon, studies that adopt this approach typically focus on adjusting the learning rate. However, learning scheme II violates DHW comfort requirements in 14 out of the 19 tested reward weight configurations. This indicates that varying the learning rate during training without varying the discount factor may not be suitable for optimisation problems involving CHS.

In summary, learning schemes with a varying γ (schemes III and IV) consistently outperform fixed-γ schemes (I and II) for concept C with a price-based optimisation objective. This confirms previous findings that a non-fixed γ is advantageous for scenarios with wide temperature ranges, large thermal inertia, and huge control responsibility assigned to the central DRL

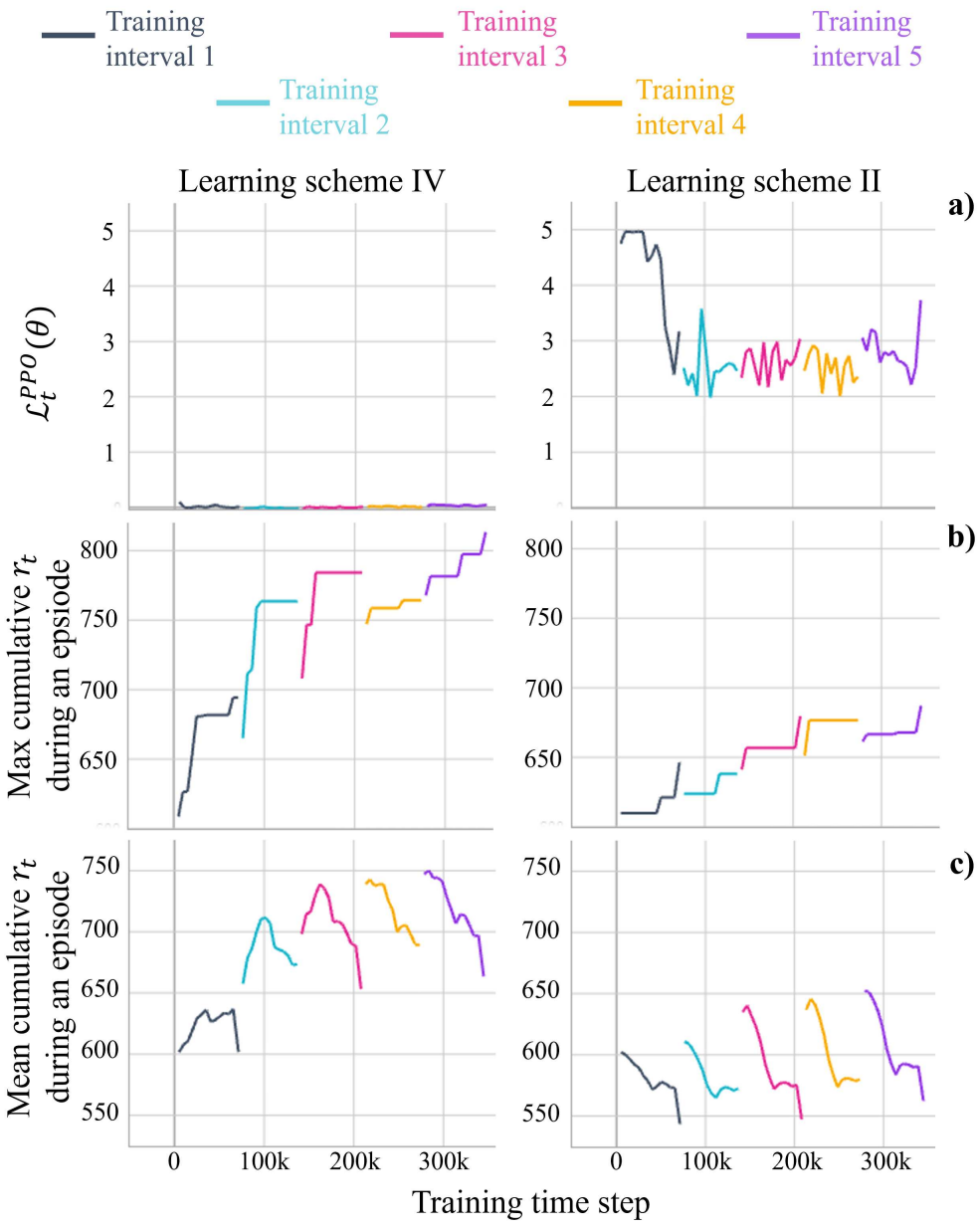

Figure 4: Insights on the training process are visualised for DRL agent trained with learning schemes IV (best performance, left graphs) and II (worst performance, right graphs) in concept C with an price-based optimisation objective. In **a)** the total loss ($\mathcal{L}_t^{\text{PPO}}(\theta)$) during each episode for the five training intervals is given. The five training intervals are presented in different colours. Figure **b)** shows the maximum cumulative reward that one of the five rollout workers obtained during the episodes of the respective training interval, while graph **c)** shows the mean reward during each episode of all five rollout workers.

agent.

4. Conclusion and future research

4.1. Conclusions

This paper investigated the impact of different learning schemes for adjusting the learning rate and discount factor on the performance of a PPO agent controlling the supply temperature in various collective heating systems. The systems considered were selected based on increasing impact of thermal inertia, which reduces the responsiveness of the system to control actions.

The energy-based optimisation provided valuable insights into the behaviour of DRL agents trained under different schemes. While scheme I consistently performed well across all system concepts, schemes II and III occasionally struggled to balance exploration vs. exploitation and long-term vs. short-term rewards. In contrast, for price-based optimisation, the GALER scheme consistently demonstrates the best performance, improving KPI^* by 3% to 15% compared to the other schemes, excluding the outliers where DRL agents violated thermal comfort constrains. In concept A, the DRL agent's strategic actions were illustrated by adjusting the T_{sup} in anticipation of future price trends, moving in the opposite direction of expected price changes. These results highlight the importance of varying both the learning rate and discount factor when training DRL agents in complex, multi-objective environments, particularly when the system dynamics lead to slow responsiveness.

In a broader context, the DRL agents and training schemes developed in this study have practical implications for managing collective heating systems in residential buildings. Many district heating networks currently operate under similar optimisation frameworks, aiming to minimise energy use or costs while ensuring user comfort. Smart thermostats and building energy management systems allow to employ DRL techniques to balance multi-objectives. However, challenges remain, particularly in accounting for external factors such as weather and user behaviour, which can complicate the realisation of optimisation potential.

4.2. Future Research

An adaptive GALER scheme, where the α and γ are varied based on system characteristics and reward function should be investigated to further optimise the performance of DRL agents. Now the values of α and γ were predefined during preprocessing (based on preliminary training), which already improved the learning performance in most cases. However, an adaptive discount factor scheme as proposed by Kim et al. [55] for a TETRIS game could be used, where the advantage function is used as an indicator for setting the value of the current discount factor during training. When the advantage function is low, the discount factor could be decreased due to overestimation, otherwise it could be increased. A similar approach for the learning rate could be developed.

To further improve the understanding of delayed-rewards and thermal inertia in the system, the DRL agent could be enhanced by implementing a long short-term memory (LSTM) network in front of the policy neural network. Zou et al. [56] combined an LSTM with Deep Deterministic Policy Gradient to improve the simulation of actual operation in multiple air handling units. However, this requires a myriad of data to train, so it would be interesting to investigate the potential of a GALER scheme with an LSTM. With respect to the BHPs environment, a multi-agent approach as in [57] should be investigated to replace the rule-based activation controllers to enable better demand response strategies.

Funding

This research was supported by a PhD fellowship of the Research Foundation Flanders (FWO) [1S08624N].

CRediT Authorship Contribution Statement

Stef Jacobs: Writing - original draft, Conceptualization, Validation, Data Curation, Investigation, Methodology, Visualization, Formal Analysis, Funding Acquisition, Project administration. **Sara Ghane:** Writing - review & editing, Conceptualization, Validation, Investigation, Methodology, Formal Analysis. **Pieter Jan Houben:** Writing - review & editing, Methodology. **Zakarya Kabbara:** Writing - review & editing. **Thomas Huybrechts:** Writing - review & editing, Supervision, Resources. **Peter Hellinckx:** Writing - review & editing, Supervision, Funding Acquisition, Conceptualization. **Ivan Verhaert:** Writing - review & editing, Supervision, Resources, Funding Acquisition, Project administration, Conceptualization.

Declaration of Competing Interest

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

Data Availability

The data presented in this study will be made available on request from the corresponding author.

References

[1] European Commission - Eurostat, energy balances, https://ec.europa.eu/eurostat/cache/infographs/energy_balances/enbal.html, 2022. Datasheet - updated 29 January 2024.

[2] M. González-Torres, L. Pérez-Lombard, J. F. Coronel, I. R. Maestre, D. Yan, A review on buildings energy information: Trends, end-uses, fuels and drivers, Energy Reports 8 (2022) 626–637. doi:10.1016/j.egyr.2021.11.280.

[3] European Commission and Directorate-General for Climate Action, Going climate-neutral by 2050 : a strategic long-term vision for a prosperous, modern, competitive and climate-neutral EU economy, Publications Office, 2019. doi:doi/10.2834/02074.

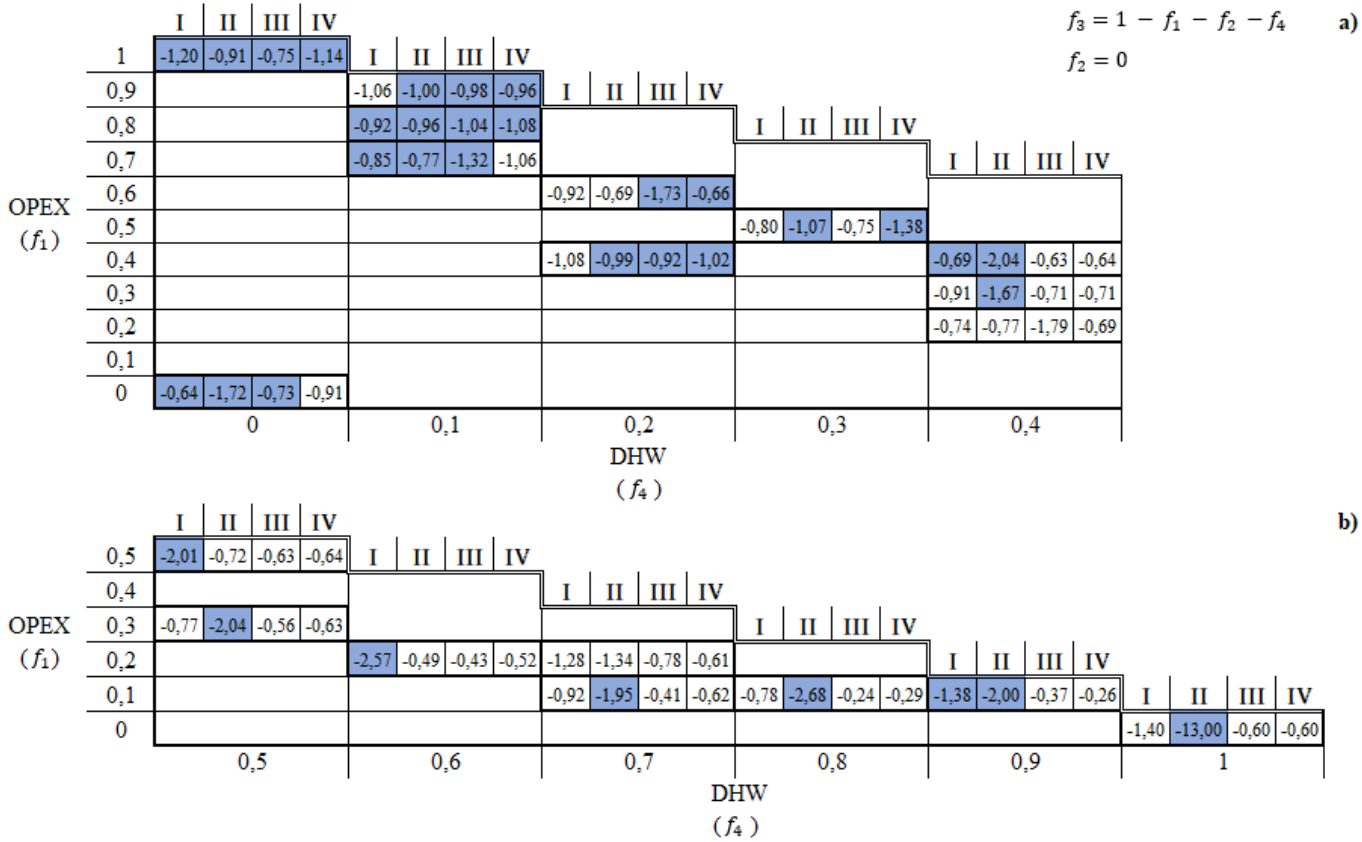

Figure 5: KPI^* values of DRL agents trained under four learning schemes (I, II, III, IV) for concept C with a price-based optimisation objective ($f_2 = 0$) with different reward weights. Rows represent variations in f_1, and columns show KPI^* values for each learning scheme at different f_4. The weighting factor f_3 equals $1 - f_1 - f_4$. In **a)**, KPI^* values for $f_4 \in [0, 0.4]$ are shown, while **b)** presents results for $f_4 \in [0.5, 1]$. Light blue boxes denote cases where $t_{DHW;\,dc} \geq 7.5\%$.

[4] B. Mathiesen, N. Bertelsen, N. Schneider, L. García, S. Paardekooper, J. Thellufsen, S. Djørup, Towards a decarbonised heating and cooling sector in Europe: Unlocking the potential of energy efficiency and district energy, Aalborg Universitet, 2019. Report.

[5] P. H. Shaikh, N. B. M. Nor, P. Nallagownden, I. Elamvazuthi, T. Ibrahim, A review on optimized control systems for building energy and comfort management of smart sustainable buildings, Renewable and Sustainable Energy Reviews 34 (2014) 409–429. doi:10.1016/j.rser.2014.03.027.

[6] W. Zhong, J. Chen, Y. Zhou, Z. Li, X. Lin, Network flexibility study of urban centralized heating system: Concept, modeling and evaluation, Energy 177 (2019) 334–346. doi:10.1016/j.energy.2019.04.081.

[7] H. Lund, B. Möller, B. V. Mathiesen, A. Dyrelund, The role of district heating in future renewable energy systems, Journal of Energy 35 (2010) 1381–1390. doi:10.1016/j.energy.2009.11.023.

[8] M. Sayegh, P. Jadwiszczak, B. Axcell, E. Niemierka, K. Bryś, H. Jouhara, Heat pump placement, connection and operational modes in european district heating, Energy and Buildings 166 (2018) 122–144. doi:10.1016/j.enbuild.2018.02.006.

[9] H. Lund, P. A. Østergaard, M. Chang, S. Werner, S. Svendsen, P. Sorknæs, J. E. Thorsen, F. Hvelplund, B. O. G. Mortensen, B. V. Mathiesen, C. Bojesen, N. Duic, X. Zhang, B. Möller, The status of 4th generation district heating: Research and results, Energy 164 (2018) 147–159. doi:10.1016/j.energy.2018.08.206.

[10] United Nations Environment Programme, 2022 global status report for buildings and construction: Towards a zero-emission, efficient and resilient buildings and construction sector, 2022-11.

[11] J. Jansen, F. Jorissen, L. Helsen, Mixed-integer non-linear model predictive control of district heating networks, Applied Energy 361 (2024) 122874. doi:https://doi.org/10.1016/j.apenergy.2024.122874.

[12] J. Arroyo, C. Manna, F. Spiessens, L. Helsen, Reinforced model predictive control (rl-mpc) for building energy management, Applied Energy 309 (2022) 118346.

[13] S. Russel, P. Norvig, Artificial Intelligence: A Modern Approach, Pearson, 2019.

[14] A. Heidari, F. Maréchal, D. Khovalyg, An occupant-centric control framework for balancing comfort, energy use and hygiene in hot water systems: A model-free reinforcement learning approach, Applied Energy 312 (2022) 118833. doi:https://doi.org/10.1016/j.apenergy.2022.118833.

[15] K. Al Sayed, A. Boodi, R. Sadeghian Broujeny, K. Beddiar, Reinforcement learning for hvac control in intelligent buildings: A technical and conceptual review, Journal of Building Engineering 95 (2024) 110085. doi:10.1016/j.jobe.2024.110085.

[16] R. Nian, J. Liu, B. Huang, A review on reinforcement learning: Introduction and applications in industrial process control, Computers & Chemical Engineering 139 (2020) 106886. doi:10.1016/j.compchemeng.2020.106886.

[17] Y. Bao, Y. Zhu, F. Qian, A deep reinforcement learning approach to improve the learning performance in process control, Industrial & Engineering Chemistry Research 60 (2021) 5504–5515. doi:10.1021/acs.iecr.0c05678.

[18] F. Elmaz, U. Di Caprio, M. Wu, Y. Wouters, G. Van Der Vorst, N. Vandervoort, A. Anwar, M. E. Leblebici, P. Hellinckx, S. Mercelis, Reinforcement learning-based approach for optimizing solvent-switch processes, Computers & Chemical Engineering 176 (2023) 108310. doi:10.1016/j.compchemeng.2023.108310.

[19] N. S. Raman, A. M. Devraj, P. Barooah, S. P. Meyn, Reinforcement learning for control of building hvac systems, in: 2020 American Control Conference (ACC), IEEE, 2020, pp. 2326–2332.

[20] R. S. Sutton, A. G. Barto, Reinforcement learning: an introduction, The MIT Press, 2018. URL: http://incompleteideas.net/book/the-book-2nd.html.

[21] P. Lissa, C. Deane, M. Schukat, F. Seri, M. Keane, E. Barrett, Deep reinforcement learning for home energy management system control, Energy and AI 3 (2021) 100043. doi:10.1016/j.egyai.2020.100043.

[22] S. Bahrami, Y. C. Chen, V. W. S. Wong, Deep reinforcement learning for demand response in distribution networks, IEEE Transactions on Smart Grid 12 (2021) 1496–1506. doi:10.1109/TSG.2020.3037066.

[23] G. Pinto, M. S. Piscitelli, J. R. Vázquez-Canteli, Z. Nagy, A. Capozzoli, Coordinated energy management for a cluster of buildings through deep reinforcement learning, Energy 229 (2021) 120725. doi:10.1016/j.energy.2021.120725.

[24] J. R. Vázquez-Canteli, J. Kämpf, G. Henze, Z. Nagy, Citylearn v1.0: An openai gym environment for demand response with deep reinforcement learning, in: Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, BuildSys '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 356–357. doi:10.1145/3360322.3360998.

[25] T. Moriyama, G. De Magistris, M. Tatsubori, T.-H. Pham, A. Munawar, R. Tachibana, Reinforcement learning testbed for power-consumption optimization, in: Methods and Applications for Modeling and Simulation of Complex Systems, Springer Singapore, Singapore, 2018, pp. 45–59.

[26] P. Haves, P. Xu, The building controls virtual test bed - a simulation environment for developing and testing control algorithms, strategies and systems, in: Building Simulation 2007: 10th Conference of IBPSA, volume 10, IBPSA, Beijing, China, 2007, pp. 1440–1446.

[27] C. Huang, S. Seidel, F. Paschke, J. Bräunig, A reinforcement learning approach for optimal heating curve adaption, in: 2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA), IEEE, 2022, pp. 1–4.

[28] A. Chatterjee, D. Khovalyg, Dynamic indoor thermal environment using reinforcement learning-based controls: Opportunities and challenges, Building and Environment 244 (2023) 110766. doi:https://doi.org/10.1016/j.buildenv.2023.110766.

[29] S. Ghane, S. Jacobs, W. Casteels, C. Brembilla, S. Mercelis, S. Latré, I. Verhaert, P. Hellinckx, Supply temperature control of a heating network with reinforcement learning, in: 2021 IEEE International Smart Cities Conference (ISC2), 2021, pp. 1–7. doi:10.1109/ISC253183.2021.9562966.

[30] S. Ghane, S. Jacobs, T. Huybrechts, P. Hellinckx, S. Mercelis, I. Verhaert, E. Mannens, Model-free deep reinforcement learning for adaptive supply temperature control in collective space heating systems, ACM Trans. Intell. Syst. Technol. (2024). URL: https://doi.org/10.1145/3709010. doi:10.1145/3709010, just Accepted.

[31] Z. Li, Z. Sun, Q. Meng, Y. Wang, Y. Li, Reinforcement learning of room temperature set-point of thermal storage air-conditioning system with demand response, Energy and Buildings 259 (2022) 111903. doi:10.1016/j.enbuild.2022.111903.

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

[33] N. Mazyavkina, S. Sviridov, S. Ivanov, E. Burnaev, Reinforcement learning for combinatorial optimization: A survey, Computers & Operations Research 134 (2021) 105400.

[34] J. Schulman, P. Moritz, S. Levine, M. Jordan, P. Abbeel, High-dimensional continuous control using generalized advantage estimation, arXiv preprint arXiv:1506.02438v6 (2018).

[35] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347 (2017).

[36] J. Shlens, Notes on kullback-leibler divergence and likelihood, arXiv preprint arXiv:1404.2000 (2014).

[37] F. Van Riet, Hydronic design of hybrid thermal production systems in buildings, PhD dissertation, University of Antwerp, 2019.

[38] S. Jacobs, F. Van Riet, I. Verhaert, A collective heat and cold distribution system with decentralized booster heat pumps: a sizing study., in: Building Simulation 2021: 17th Conference of IBPSA, volume 17, IBPSA, Bruges, Belgium, 1-3 September 2021, pp. 223–230. doi:10.26868/25222708.2021.30423.

[39] S. Jacobs, S. Van Minnebruggen, H. Matbouli, S. Ghane, P. Hellinckx, I. Verhaert, Evaluating innovative collective heating and cooling concepts by incorporating occupants' preferences for conflicting performance indicators, Energy and Buildings 314 (2024) 114264. doi:10.1016/j.enbuild.2024.114264.

[40] S. Jacobs, M. De Pauw, S. Van Minnebruggen, S. Ghane, T. Huybrechts, P. Hellinckx, I. Verhaert, Grouped charging of decentralised storage to efficiently control collective heating systems: Limitations and opportunities, Energies 16 (2023). doi:10.3390/en16083435.

[41] I. Meireles, V. Sousa, B. Bleys, B. Poncelet, Domestic hot water consumption pattern: Relation with total water consumption and air temper-

ature, Renewable and Sustainable Energy Reviews 157 (2022) 112035. URL: https://www.sciencedirect.com/science/article/pii/S1364032121012971. doi:10.1016/j.rser.2021.112035.

[42] E. Fuentes, L. Arce, J. Salom, A review of domestic hot water consumption profiles for application in systems and buildings energy performance analysis, Renewable and Sustainable Energy Reviews 81 (2018) 1530–1547. URL: https://www.sciencedirect.com/science/article/pii/S1364032117308614. doi:10.1016/j.rser.2017.05.229.

[43] ClimaWays BVBA, Collindi verwarmingssatellieten: geïndividualiseerde collectieve verwarmingssystemen (dutch), 2022. Datasheet.

[44] Buildwise, Average weather profiles for Belgium, 2023. URL: http://www.buildwise.be/nl/, the profiles were created by Buildwise as part of the IEA EBC Annex 80 project (https://annex80.iea-ebc.org/). Data are from the REMO15 model forced by MPI-M-MPI-ESM-LR. Accessed: 17-01-2023.

[45] Solar Energy Laboratory Univ. of Wisconsin-Madison (SELUWM), TRNSYS 17 volume 8 weather data, TRNSYS 17 8 (2014).

[46] Vlaams Energie – en Klimaatagentschap, Bijlage v – bepalingsmethode epw 2022 (dutch), https://www.energiesparen.be/bouwen-en-verbouwen/epb-pedia/epb-regelgeving/energiebesluit/bijlage-v/, 2022. [Online; accessed 9-February-2022].

[47] VLAIO, Instal 2020 project: Integraal ontwerp van installaties voor sanitair en verwarming (dutch), https://www.instal2020.be/, 2014-2018. VIS 135098.

[48] VLAIO, Productie en distributie van sanitair warm water: selectie en dimensionering (dutch), https://www.tetra-sww.be/, 2012-2014. TETRA 120145.

[49] J. De Schutter, I. Verhaert, M. De Pauw, A methodology to generate realistic random behavior profiles for space heating and domestic hot water simulations, in: The REHVA Annual Meeting Conference: Low Carbon Technologies in HVAC, REHVA, Brussels, Belgium, 2018, pp. 1–8.

[50] M. De Pauw, F. Van Riet, J. De Schutter, S. Binnemans, J. Van der Veken, I. Verhaert, 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, 2018, pp. 1–8.

[51] F. Jorissen, G. Reynders, R. Baetens, D. Picard, D. Saelens, L. Helsen, Implementation and verification of the ideas building energy simulation library, Journal of Building Performance Simulation 11 (2018) 669–688. URL: 10.1080/19401493.2018.1428361. doi:10.1080/19401493.2018.1428361.

[52] Solar Energy Laboratory Univ. of Wisconsin-Madison (SELUWM), TRNSYS 17 volume 4 mathematical reference: Type 60 (stratified fluid storage tank with internal heat exchangers), TRNSYS 17 4 (2009) 390–396.

[53] R. Zhang, D. Wang, Z. Yu, Y. Sun, H. Wan, Y. Liu, Q. Jiao, M. Gao, J. Fan, B. Lan, Dual-objective optimization of large-scale solar heating systems integrated with water-to-water heat pumps for improved techno-economic performance, Energy and Buildings 296 (2023) 113281. doi:10.1016/j.enbuild.2023.113281.

[54] F. Van Riet, G. Steenackers, I. Verhaert, A new approach to model transport delay in branched pipes, in: 10th International Conference on System Simulation in Buildings, Liège, Belgium, 2018, pp. 1–17.

[55] M. Kim, J.-S. Kim, M.-S. Choi, J.-H. Park, Adaptive discount factor for deep reinforcement learning in continuing tasks with uncertainty, Sensors 22 (2022). doi:10.3390/s22197266.

[56] Z. Zou, X. Yu, S. Ergan, Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network, Building and Environment 168 (2020) 106535. doi:https://doi.org/10.1016/j.buildenv.2019.106535.

[57] J. Xie, A. Ajagekar, F. You, Multi-agent attention-based deep reinforcement learning for demand response in grid-responsive buildings, Applied Energy 342 (2023) 121162. doi:10.1016/j.apenergy.2023.121162.