Energy and AI 20 (2025) 100506

Contents lists available at ScienceDirect

Energy and AI

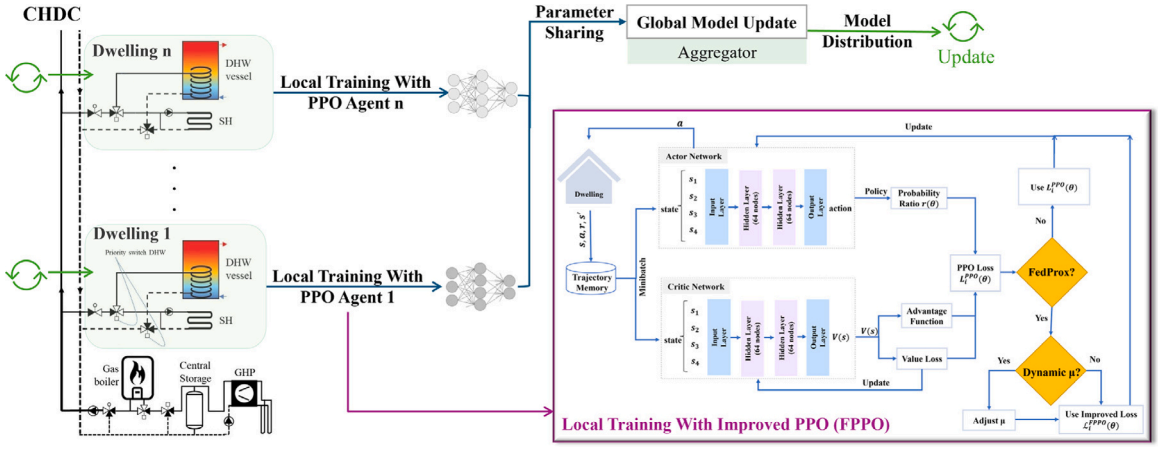
journal homepage: www.elsevier.com/locate/egyai

Federated proximal policy optimization with action masking: Application in collective heating systems

Sara Ghane [a,*], Stef Jacobs [b], Furkan Elmaz [a], Thomas Huybrechts [a], Ivan Verhaert [b], Siegfried Mercelis [a]

[a] University of Antwerp - imec, IDLab - Faculty of Applied Engineering, Sint-Pietersvliet 7, Antwerp, 2000, Belgium
[b] EMIB, Faculty of Applied Engineering - Electromechanics, University of Antwerp, Groenenborgerlaan 171, Antwerp, 2020, Belgium

GRAPHICAL ABSTRACT



HIGHLIGHTS

- Proposed FRL method, FPPO, combines PPO and FedProx for scalable, privacy-aware DHW tank control.
- Action masking ensures DHW comfort and simplifies reward design to single objective.
- Global reward design aligns agents for collective energy savings in heating systems.
- Dynamic proximal term coefficient (μ) adjustment enhances energy savings and user comfort in FPPO.
- FPPO aligns decentralized agents for coordinated energy saving actions in a privacy-aware manner.

ARTICLE INFO

Keywords:
Reinforcement learning
Federated reinforcement learning
Decentral DHW storage tank
Collective heating system

ABSTRACT

This paper introduces a novel privacy-aware Federated Proximal Policy Optimization (FPPO) method combined with action masking. As a Federated Reinforcement Learning (FRL) approach, the proposed method is used for optimizing the reloading of Domestic Hot Water (DHW) storage tanks, with a focus on energy savings and DHW thermal comfort in collective heating systems. The proposed approach combines FedProx as the Federated Learning (FL) method and Proximal Policy Optimization (PPO) as the Deep Reinforcement Learning

* Corresponding author.
E-mail addresses: sara.ghane@uantwerpen.be (S. Ghane), Stef.Jacobs@uantwerpen.be (S. Jacobs), Furkan.Elmaz@uantwerpen.be (F. Elmaz), Thomas.Huybrechts@uantwerpen.be (T. Huybrechts), Ivan.Verhaert@uantwerpen.be (I. Verhaert), Siegfried.Mercelis@uantwerpen.be (S. Mercelis).

https://doi.org/10.1016/j.egyai.2025.100506
Received 13 January 2025; Received in revised form 11 March 2025; Accepted 18 March 2025
Available online 27 March 2025

Thermal comfort
Energy saving

(DRL) technique to address the challenges of distributed control while ensuring data privacy. Key contributions include: (1) employing action masking to guarantee compliance with comfort level, (2) designing a global reward function to align agents actions toward collective energy savings, (3) implementing a privacy-aware design where only model parameters are shared with a global aggregator, avoiding raw data transmission, and (4) optimizing PPO's loss function for improved performance.

PPO was benchmarked using a common FL method (FedAvg) alongside two other DRL methods, where PPO outperformed both in scalability and energy savings, especially in larger systems. Then, PPO-based FRL was refined into FPPO by integrating a proximal term with coefficient μ into the loss function to enhance the performance. Experiments were conducted with both fixed and dynamically adjusted μ, with the latter demonstrating better energy savings and comfort. Results show that FPPO achieves up to 10.08% energy savings while maintaining DHW discomfort below 8.72% in systems with at least 20 dwellings. These findings highlight FPPO as a scalable, privacy-aware, and energy-efficient solution for distributed control in collective heating systems.

1. Introduction

The heat demand of European households accounts for 78% of the final energy used for Space Heating (SH) and Domestic Hot Water (DHW) production [1]. As the world deals with the consequences of excessive energy use on climate change and sustainable development, this research focuses on reducing the energy use in heating systems. Heating systems are one of the contributors to Greenhouse Gas (GHG) emissions, leading to climate change and its associated impacts. Reducing energy use in heating systems can help cities and societies to move toward a more sustainable energy infrastructure, potentially lowering energy bills and reducing GHG emissions.

The REPowerEU plan, driven by the European Commission, aims to provide affordable, secure, and sustainable energy for Europe. It recognizes that increasing buildings' energy efficiency gain and saving energy are the most cost and time-efficient measures to address the energy crisis [2,3]. Eurostat data reveals that DHW in Europe represents approximately 15% of the overall residential energy consumption [4]. The European Green Deal, Climate Goals by 2050, and the REPowerEU programs emphasize reducing GHG emissions and buildings' energy use [3,5,6]. Existing measures are expected to reduce GHG emissions in Belgium by 23% by 2030 compared to 2005 levels. However, to achieve the more ambitious target of a 46% reduction from 2005 levels, additional efforts are essential. These include enhancing energy efficiency in building heating systems and implementing energy-saving initiatives [2].

One of the existing measures is to deploy low-temperature emitters in newly-built and renovated dwellings [2]. Renovation reduces the heat demand, which allows to supply the heat at lower temperatures. These lower temperatures facilitate the use of Heat Pumps (HPs) and other Renewable Energy Sources (RESs) for heat production. In addition to renovating buildings, implementing collective thermal systems are promising to decarbonize thermal energy supply [7–9]. A common example are district heating systems, which connect various houses and industries through distribution pipes. This facilitates the integration of HPs and other RESs and provides flexibility to the electric grid. Also on a small-scale, these systems exist, i.e., the Combined Heat Distribution Circuit (CHDC) within apartment buildings. A CHDC consists of one supply pipe and one return pipe to distribute heat for both SH and DHW [10]. The heat is produced by a central renewable energy system, such as HPs or a connection to a larger district heating system. Since both SH and DHW are supplied to end-users with the same pipe, the supply temperature ($T_{sup,SP}$) is typically set at 65 °C in practice [11–13], which results in a lower coefficient of performance for the central HP. By using decentralized DHW storage tanks within a CHDC, a more sustainable and flexible approach to providing DHW can be achieved [10]. This can be considered as one of the additional measures to help reduce GHG emissions by 46% by 2030.

The central HP operates most efficiently at low supply temperatures. However, a higher temperature is necessary for storing DHW. Balancing the low temperature for SH with underfloor heating and the high temperature for DHW recharging is key to optimizing system efficiency while maintaining DHW comfort [10].

Achieving the aforementioned goal is challenging due to the intricate thermodynamic behavior of collective heating systems and the unpredictable variations introduced by factors such as weather conditions and diverse occupant behaviors. To address these challenges, advanced control strategies are required. Model Predictive Control (MPC) and Deep Reinforcement Learning (DRL) are two major types of more sophisticated control techniques that are becoming more prevalent in the Heating, Ventilation, and Air Conditioning (HVAC) sector and building's energy systems. They both offer intriguing potentials [14,15], yet DRL is renowned for its high adaptability and can cope with uncertainties that are associated with complex systems that are hard to model. Moreover, DRL does not need prior knowledge of environment. Although DRL can be computationally expensive to train, it is relatively efficient in deployment, requiring relatively simple calculations once trained. In contrast, MPC can be computationally demanding during real-time control because it requires solving an optimization problem at each control step. Given these characteristics, this research focuses on using DRL to control decentralized DHW storage tank loading. DRL's capacity to handle long-term (or infinite) prediction horizons and its low computational demands (after training and deployment) make it well-suited for application in real-life buildings [14–16].

To efficiently control DHW storage tanks in a CHDC system, training a DRL agent requires a substantial amount of data which is computationally expensive. This computational overhead can be reduced by using a centralized DRL-based system and training at a single central controller [17]. But, this centralized DRL approach causes three problems. First, the state space can be very big which makes the control problem very complex to solve [18]. Second, a centralized approach can be insufficient for dealing with a dynamic time-varying control problem that requires lots of data processing for training [17]. Third, collecting raw private data of the end-users at a central control agent increases the risk of privacy violation even though it can lead to energy savings [19]. Risking privacy violation is unacceptable, as it undermines trust, exposes personal data, and raise ethical and legal concerns. Protecting privacy is vital for responsible technology use and respecting individual rights. Therefore, in this research distributed training of DRL agents using Federated Learning (FL) [20] is employed to reduce the training costs, accelerate training, and to ensure better data privacy. In this approach which is known as Federated Reinforcement Learning (FRL), decentralized agents are collaboratively trained by sharing the learned experiences without direct data exchange to avoid privacy leakage [21].

1.1. State-of-the-art

Several studies have explored the use of FRL in energy management and control. Lee and Choi [17] used FRL to optimally schedule solar PV, energy storage systems, and home appliances without sharing energy usage data, while maintaining user comfort. The results demonstrated

successful energy management of appliances. Fujita et al. [22] combined Soft actor–critic (SAC) [23] with FL to control HVAC systems in buildings with solar panels, storage batteries, and air conditioners. This approach reduced training data and time while maintaining privacy. Lee et al. [24] proposed a FRL model for multi-residential energy scheduling under time-of-use and demand charge tariffs. It combined local energy management systems with a central server to optimize energy use across multiple residences without sharing sensitive data. Their method effectively balanced on-grid energy and battery use to reduce costs, showing advantages over conventional methods in performance, adaptability, and learning speed. Additionally, authors in [25] presented a FRL architecture to optimize the energy consumption of buildings with PV systems and shared energy storage systems, using a selective parameter sharing method to avoid privacy leakage. The authors employed an actor–critic RL method, which improved overall energy use. Gao et al. [26] introduced a privacy-preserving, cloud-free residential energy management system using personalized federated DRL. Their framework enables local model aggregation and collaborative training while enhancing model convergence through personalized neural network layers. It outperforms centralized approaches and traditional solutions in reducing energy use. In addressing HVAC control challenges in commercial buildings, a Federated Accelerated Multi-Agent DRL approach was proposed by Xia et al. [27]. By reformulating multi-zone HVAC control as an MDP and integrating a FL mechanism, the approach improves convergence, reduces energy use, and ensures thermal comfort.

Tan et al. [28] propose a FRL approach for privacy-preserving energy management in residential microgrids. They used PPO with FedAvg to optimize energy scheduling decisions while preserving user data privacy. Clustering is employed to categorize families with different data distributions, enabling information sharing within the same class which led to improved learning efficiency. A cooperative multi-microgrid dispatch framework using personalized federated multi-agent RL with clustering was introduced by Yang et al. [29]. By leveraging Twin Delayed Deep Deterministic Policy Gradient (TD3)-based RL, and a personalized FL (FedAvg with clustering-based personalization), their approach optimized energy trading, carbon emissions, and dispatch precision while preserving data privacy. Rezazadeh and Bartzoudis [30] proposed a FRL for smart microgrid energy control using SAC. Their approach applies FedAvg to aggregate local models from smart homes while preserving privacy. By optimizing battery scheduling, energy trading, and load balancing, their method reduces energy costs and CO_2 emissions, outperforming traditional centralized RL approaches in multi-house distributed energy resource systems. In [31], a Federated Multi-Agent DRL (F-MADRL) approach was proposed for multi-microgrid energy management using PPO. Their method employs a FedAvg to aggregate RL policy parameters instead of raw model weights. By optimizing energy trading, battery storage, and decentralized scheduling, their framework enhances privacy, learning efficiency, and energy self-sufficiency across interconnected microgrids.

Table 1 demonstrates a clear comparison between previous studies and this work. While most prior research has focused on FedAvg-based FL, we employ FedProx, as it constrains how far each agent's local model parameters can deviate from the global model, helping to mitigate the effects of diverse data across agents. As it is evident from this table, the application of FRL to decentralized control, particularly in collective heating systems, remains a largely unexplored area of research. This gap highlights the innovative nature of our approach, which leverages privacy-aware, scalable learning to optimize decentralized control of DHW storage tanks. Our method not only ensures occupant comfort but also significantly enhances energy efficiency, offering a novel solution to a critical challenge in energy domain. Furthermore, this is the first work to employ FRL and action masking together, as no prior evidence of such an approach could be found in the literature. This is a complex decision-making problem due to many variables in system that may seriously affect its performance, such as notable thermal inertia of distribution pipes and reloading of storage tanks, and network time delays. Moreover, the delayed consequences of control actions adds to the complexity of control, as agents' actions directly impact the performance of the entire CHDC, including the central supply temperature. Balancing the conflicting objectives of minimizing global energy use of the collective heating system and maintaining thermal comfort for DHW in each dwelling is especially complicated when agents do not share data or coordinate DHW storage tanks reloading times. Our novel FRL approach addresses these challenges by using action masking to ensure comfort and defining a shared global reward for all agents to achieve overall energy reduction. Moreover, incorporating future demand information further enhances both comfort and energy savings in our proposed method.

Table 1
Comparison of FRL-based studies in energy management and control. Different RL methods were used in different studies, including Advantage Actor–Critic (A2C) [32], Soft Actor–Critic (SAC) [23], Deep Q-Network (DQN) [33], Multi Agent Deep Deterministic Policy Gradient (MADDPG) [34], Proximal Policy Optimization (PPO) [35], and Twin Delayed Deep Deterministic Policy Gradient (TD3) [36].

Study	District or Collective heating	RL method	FL method
[17]	✗	A2C	FedSGD [37]
[22]	✗	SAC	FedAvg [37]
[24]	✗	DQN	FedAvg
[25]	✗	Actor–Critic	FedAvg
[26]	✗	DQN	FedAvg
[27]	✗	MADDPG	FedAvg
[28]	✗	PPO	FedAvg
[29]	✗	TD3	FedAvg
[30]	✗	SAC	FedAvg
[31]	✗	PPO	FedAvg
This work	✓	PPO	FedProx [38]

1.2. Contributions

Controlling decentralized DHW storage tanks in a collective heating system with a central heat production, such as a CHDC connected to a HP, adds to the complication of the control task. A single supply pipe meets both SH and DHW demands, which requires a careful balance between low-temperature SH operation (a supply temperature of 30 °C to 45 °C for underfloor heating) and high-temperature DHW recharging (a supply temperature of 65 °C). Therefore, a good strategy is to ensure most storage tanks load simultaneously to improve overall energy savings by minimizing the frequency of high supply temperature requests.

Additionally, each dwelling exhibits unique dynamic characteristics influenced by factors such as occupant behavior, the number of occupants, window orientation, and other environmental factors. Consequently, a control strategy that is effective for one dwelling may not perform as well in another, making a single, uniform control policy across all dwellings insufficient [39]. To address this challenge, our study introduces a novel FRL approach, Federated Proximal Policy Optimization with action masking (FPPO). FPPO enables collaborative and privacy-preserving training among multiple DRL agents, where each agent is responsible for controlling a specific dwelling. This ensures an adaptable and scalable control strategy that improves energy efficiency while maintaining user comfort. After training, this control strategy is deployed and tested across all dwellings to confirm its effectiveness.

The primary contributions of this work are as follows:

- *First application of FRL to collective heating systems with privacy-aware and scalable DHW control*: While FRL has been explored in general energy management, this work represents its first application to decentralized control for collective heating systems, addressing an important gap in the literature. Our proposed approach (FPPO) enables collaborative learning across multiple dwellings without sharing raw data from dwellings, thereby preserving occupant privacy.

- *Enhanced RL optimization via customized loss function and action masking for enforcing comfort constraints*: We introduce a customized loss function for Proximal Policy Optimization (PPO) using FL method FedProx to improve optimization of our distributed control problem. Additionally, instead of multi-objective reward functions with conflicting objectives such as energy saving and comfort, we enforce comfort constraints directly through action masking, ensuring valid control actions at all times.
- *Implicit agent coordination via global reward function for energy-efficient control*: Our method eliminates explicit communication-based coordination by leveraging a shared global reward function and action masking. This enables natural alignment of DHW storage tank charging behaviors, reducing unnecessary high-temperature requests and optimizing overall energy efficiency.
- *Integration of future demand for improved energy savings and comfort*: By incorporating future demand predictions into the state space and action masking logic, our approach enhances both energy savings and occupant comfort, leading to a more effective and proactive control strategy.

This work provides a scalable, privacy-aware solution for decentralized DHW control in collective heating systems. By combining FRL, action masking, and defining a shared global objective for all agents, our approach addresses the challenge of decentralized control in collective heating systems while ensuring energy efficiency and occupant comfort.

1.3. Outline

The rest of the paper is organized as follows. Section 2 contains the background of current study. Section 3 describes the simulation environment, proposed FRL approach, and Key Performance Indicators (KPIs). Afterwards, experiments and results are discussed in Section 4. Finally, the research is concluded in Section 5.

2. Background

This section provides a brief overview of Deep Reinforcement Learning (DRL) and Federated Reinforcement Learning (FRL). For further details, refer to [37,40–43]. At the end, the baseline methods used for comparison in this study are also described.

2.1. Deep reinforcement learning

Reinforcement Learning (RL) is a subset of Machine Learning (ML) that provides a mathematical framework for sequential decision-making and autonomous control [43,44]. An RL agent interacts with its environment, taking actions based on observations and receiving rewards to learn an optimal policy that maximizes cumulative future rewards [45,46].

RL problems are modeled as Markov Decision Processes (MDPs), defined by a tuple (S, A, P, R, γ), where S and A are a finite set of states and actions, P represents state transition probabilities, R is the reward function, and $\gamma \in [0, 1]$ is the discount factor. A policy π maps states to actions, and the optimal policy π^* maximizes expected accumulated rewards [43–46].

Thus, the expected accumulated reward achieved by the agent measures the performance of a policy and is represented by state-value function. For any state $\forall s \in S$, the state-value function is given by Eq. (1).

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

Similarly, the action-value function for the policy π represents the expected future discounted rewards for an agent starting in state s, taking action a, and subsequently following policy π:

$$Q^{\pi}(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t), \Big| , \pi, s_0 = s, a_0 = a\right] \tag{2}$$

This allows the state-value function to be expressed in terms of the action-value function:

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

Moreover, the advantage function, $A^{\pi}(s, a) = Q^{\pi}(s, a) - V^{\pi}(s)$, is often used to evaluate how beneficial an action is compared to the policy's average action. In addition, an estimated advantage function, $\hat{A}$, can be used in the context of Generalized Advantage Estimation (GAE) to reduce the variance of policy gradient estimates by combining multiple-step returns.

RL in real-world applications faces challenges with large state and action spaces, making each state's value computation impractical. Deep Reinforcement Learning (DRL) overcomes this by integrating Neural Networks (NNs) to approximate the value function, policy, and/or model [47].

In this study, characteristics of the environment are unknown to agent and optimal policy should be obtained by interacting with environment. Therefore, we utilize model-free DRL algorithm, called Proximal Policy Optimization (PPO) [35] which is among the most widely used DRL methods due to its solid theoretical basis and well-established practical benefits. The details of PPO are explained later in Section 3.3.3 (Proximal Policy Optimization).

2.2. Federated reinforcement learning

Federated Reinforcement Learning (FRL) is an emerging approach that leverages the idea of FL in RL to enable decentralized training across multiple agents while preserving privacy in dynamic and unknown environments [21]. FRL enables agents to collaboratively learn optimal policies on decentralized data distributed across multiple environments, without exchanging raw private data, and thus upholding data privacy standards [20,21,48,49]. This implicit access to a larger amount of data through FL allows the agents to benefit from greater diversity in data than what any individual agent could access independently.

In addition to enabling agents to interact with their environments to explore and gain experience for optimal decision-making, FRL ensures that data obtained during each agent's exploration remains private and is not shared with others. Privacy is preserved by sharing only the parameters of locally trained models, rather than the data itself. Thus, FRL is regarded as a privacy-enhanced distributed DRL framework [20,21]. By exchanging model parameters during training rather than raw data, FRL significantly reduces the risk of compromising data privacy.

In FRL, a central aggregator collects the updates from each DRL agent and combines them to improve the global model [21]. Two common aggregation methods used in FL are Federated Averaging (FedAvg) [37] and FedProx [38], which differ in how they handle variations in local data distributions and the update process.

2.2.1. FedAvg

One of the most common methods in FL is FedAvg [37], which is a straightforward aggregation method in which the central aggregator computes the average of all local model updates. Suppose we have N agents, each with a local model parameter θ_i, where $i \in 1, 2, \ldots, N$. After each agent performs local updates, the aggregator calculates the global model parameter θ as follows:

$$\theta = \frac{1}{N} \sum_{i=1}^{N} \theta_i \tag{4}$$

This averaging process allows FedAvg to aggregate local knowledge from each agent, resulting in a shared global model that reflects information from all agents while keeping raw data private. FedAvg is employed as a baseline method for comparison in our research.

2.2.2. FedProx

FedProx is an enhanced version of FedAvg that adds a proximal term with an adjustable parameter μ to the local objective function, to effectively stabilize the training process by limiting the impact of local updates that deviate significantly from the global model. As a generalization and re-parameterization of FedAvg, FedProx becomes equivalent to FedAvg when the proximal term coefficient $\mu 0$ is set to 0. When $\mu > 0$, FedProx constrains how far each agent's local model parameters can deviate from the global model, helping to mitigate the effects of non-IID (non-Identically Independently Distributed) data across agents.

In FedProx, each agent i minimizes the following customized objective function $\mathcal{L}_i(\theta_i)$ during local updates:

$$\mathcal{L}_i(\theta_i) = L_i(\theta_i) + \frac{\mu}{2}|\theta_i - \theta|^2 \tag{5}$$

where $L_i(\theta_i)$ is the local objective function for agent i with local model parameter θ_i, θ is the global model parameter, and $\mu \geq 0$ is the proximal term coefficient that controls the regularization effect. The proximal term $\frac{\mu}{2}|\theta_i - \theta|^2$ penalizes large deviations from the global model to encourage each agent's local update to remain close to the global model, especially when local data distributions differ. This regularization helps improve stability and convergence in environments with heterogeneity across agents. Similar to FedAvg, the central aggregator combines the local parameters from each agent to update the global model, using Eq. (4).

Finally, FedProx allows flexibility in the choice of μ:

- *Fixed* μ: Using a constant μ provides steady regularization throughout training, but requires careful tuning. A large μ can slow convergence by keeping updates too close to the starting point, while a small μ may have minimal impact.
- *Dynamic* μ: μ can also be adjusted dynamically, providing adaptive regularization that responds to current performance of the model. This flexibility can further enhance convergence by applying stronger or weaker regularization as needed.

2.3. Baseline methods for comparison

The baseline methods that are used to evaluate the effectiveness of the proposed method are explained in the following subsections.

2.3.1. RBC baseline

The RBC baseline used here was introduced in [10], and is an optimized RBC which was designed for energy efficient control of decentralized DHW storage tanks. It follows the two-sensor control strategy proposed in [10], where it was shown that using decentralized storage tanks enables the lowering of central supply temperatures through grouped charging of the decentralized storage tanks. This approach ultimately resulted in a more energy efficient RBC. This is done by utilizing two temperature sensors in each decentralized storage tank to determine whether the central supply temperature should be raised to 65 °C (DHW mode). The central supply temperature alternates between 65 °C (DHW mode) and 35 °C (SH mode). The upper sensor in the DHW storage tank triggers a high central $T_{sup,SP}$ when its temperature is below the DHW set point (40 °C), while the lower sensor only charges the DHW tank when the central supply temperature is already high.

2.3.2. FRL baselines

The FRL baselines employ the FedAvg (Section 2.2.1) approach with three well-known DRL algorithms, namely Soft Actor–Critic (SAC), Deep Q-Network (DQN), and Proximal Policy Optimization (PPO).

- *SAC* [23] is a model-free, off-policy DRL algorithm that optimizes a stochastic policy while maximizing an entropy term to encourage exploration. This entropy regularization improves robustness and helps the agent avoid premature convergence to suboptimal policies. SAC is known for its stability, sample efficiency, and ability to perform well in both continuous and discrete action spaces, making it a good choice for complex environments.
- *DQN* [33] is a value-based, off-policy DRL algorithm that approximates the optimal action-value function using a NN. It employs experience replay memory to reuse past experiences and a target network to stabilize training, enabling effective learning in discrete action spaces. An enhanced version of DQN is used in our comparisons that combines Double Q-Learning to reduce overestimation bias with a dueling network architecture that separates state-value and advantage functions. This design improves its stability and performance in complex environments.
- *PPO* [35] is an on-policy DRL algorithm that improves stability and performance by limiting large policy updates. Its efficiency and simplicity make it a popular choice for various complex RL tasks. For more details, please refer to Section 3.3.3 (Proximal Policy Optimization).

First, FedAvg is used to aggregate local model parameters from individual dwellings to train a global model. Each DRL agent is trained locally based on interactions with its environment, sharing only model parameters with a central aggregator to avoid raw data exchange. FedAvg combines these local updates into a global model, which is distributed back to the agents. This privacy-preserving and collaborative training approach with FedAvg is used as a FRL baseline in our comparisons.

3. Methodology

This section presents the proposed Federated Proximal Policy Optimization (FPPO) framework for optimizing decentralized control of Domestic Hot Water (DHW) storage tanks in collective heating systems. The methodology follows a privacy-aware Federated Reinforcement Learning (FRL) approach, enabling multiple learning agents to collaboratively improve control strategies without sharing raw data.

3.1. Overview of the proposed approach

The key objective of this study is to develop a scalable and privacy-preserving learning-based control strategy for DHW storage tanks within a CHDC, called FPPO, which is explained later in Section 3.3.3. The proposed approach consists of the following core elements:

1. CHDC Simulation Environment: A CHDC simulator integrated with an OpenAI Gym environment is used to model the thermal behavior of a CHDC that delivers SH and DHW.
2. Proposed Federated Reinforcement Learning Framework: This consists of the following components:
    - MDP Formulation of Reinforcement Learning Agents: Defines the state and action space along with the reward function. A global reward function is designed to align agents' actions toward reducing energy use.
    - Action Masking for Comfort Assurance: To prevent undesired actions that could lead to DHW thermal discomfort, action masking is applied to restrict the agent's action space.
    - Proposed Federated Proximal Policy Optimization (FPPO): This section provides a brief overview of PPO, followed by the proposed loss function definition for FPPO and the dynamic adjustment of the proximal term coefficient.
    - Learning and Aggregation Process in Federated Reinforcement Learning: Agents train locally and share only model parameters, which are aggregated using FL.

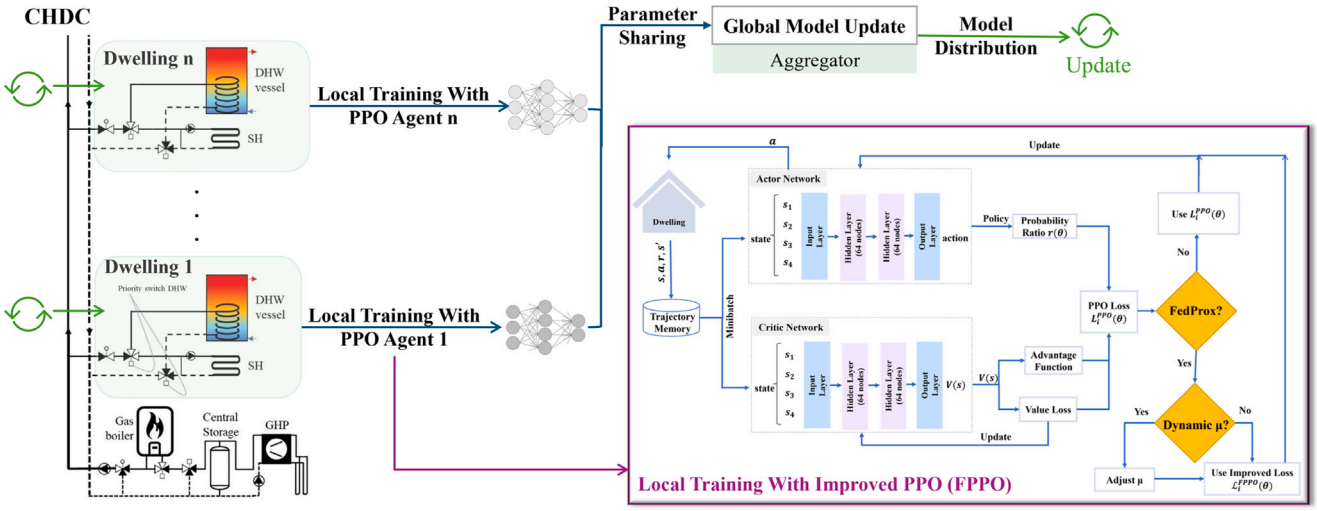


Fig. 1. High-level flow of the proposed approach, illustrating the interaction between local agents and the FL aggregator within a CHDC. The purple frame highlights the local training process with FPPO, demonstrating the flow with and without FedProx. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

A high-level overview of the proposed approach is illustrated in Fig. 1, depicting a CHDC that highlights the interaction between local agents and the federated learning aggregator. The local training process with FPPO is enclosed in a purple frame, illustrating the training flow with and without FedProx.

3.2. CHDC simulation environment

The simulator models the thermal behavior of a CHDC that delivers SH and DHW in an apartment building. In this study, the CHDC includes a supply and return pipe network supported by a central Geothermal HP (GHP), a storage tank, and a back-up boiler for peak loads. The number of connected dwellings varies throughout the experiments to investigate the scalability of the presented learning approach.

Each 100 m^2-dwelling has an underfloor heating system designed at 35 °C/ 30 °C for SH, and a stratified storage tank of 90 liters for DHW. The heat load of each dwelling in design conditions (21 °C indoor set point and −8 °C outdoor temperature) is between 1.4 kW and 2.1 kW, depending on its location (bottom, intermediate or top floors). A generic overview of the case study is given in Fig. 2.

The central GHP is designed to cover the SH demand at full-load, and also serves as a heat source for the boiler. The central storage tank has a storage volume equivalent to one hour of full-load operation of the GHP, as identified as the most efficient by Van Riet [50], at a temperature of 45 °C. The boiler is sized to increase the supply temperature from 45 °C to 65 °C when the decentralized storage tanks are being recharged.

The simulation time step is 10 s. In this way, the simulator reflects the behavior of both the components in the CHDC and the inhabitants in detail. The internal heat gains and occupancy profiles are based on a stochastic "profile generator" developed in the TETRA-SWW [51] and Instal2020 [52] projects with statistical occupancy data of different family types in Belgium. The family types consist of one to three persons and the average daily DHW consumption is 2.54 kW h per person.

The thermal models in the simulator are extensively described in [10,19,50]. The transient thermal behavior of all components in a CHDC are represented by first order, ordinary, linear, and non-homogeneous differential equations. Only storage tank models are partial differential equations along height and in temperature, based on type 60 of TRNSYS [53]. To take account of stratification in storage tanks, each decentralized DHW storage tank is divided in 4 equal water layers of 22.5 l and central storage tank has equal water layers of 180 liters (total volume depends on sizing GHP, and thus number of dwellings). Time delays of control valves are considered with a time constant (τ) of 32 s as in [50] and thermal behavior of mixing is according to mixing rule. The time delay in the pipes is modeled by applying plug-flow principle [54] and their thermal losses are characterized by a resistance model.

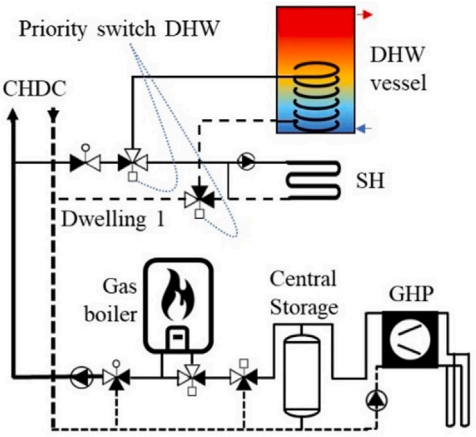


Fig. 2. Overview of the case study (one of the concepts considered in [10]).

3.3. Proposed federated reinforcement learning framework

The proposed FRL framework utilizes FedProx (Section 2.2.2) to train decentralized DRL agents collaboratively. Each agent optimizes the control of a DHW storage tank in a dwelling by interacting with the environment and only sharing model parameters while preserving privacy.

3.3.1. Markov Decision Process (MDP) formulation

The Markov Decision Process (MDP) formulation for each agent is defined by the following components:

State space. Each agent observes the following states:

- s_0: Current temperature of the top sensor in the DHW storage tank.
- s_1: Current central supply temperature (T_{sup}).
- s_2: Number of occupants in the dwelling.
- s_3: Future DHW demand indicator of the next hour (1 for expected demand, 0 otherwise).

Action space. Each agent selects a single action from the following three available options for loading of the DHW storage tank:

- a_0: Not loading.
- a_1: Guaranteed loading.
- a_2: Conditional loading when the central supply temperature (T_{sup}) is high.

Reward function. To align all agents with a common objective of optimizing global energy use, a shared reward function is designed. The reward function incentivizes energy savings based on the central supply temperature T_{sup}, where rewards are assigned to encourage efficient energy use. The reward $R \in [-1, 1]$ is defined as follows:

$$R = \begin{cases} 1, & T_{\text{supply}} \leq T_{\text{sup\_min}} \\ -1, & T_{\text{supply}} \geq T_{\text{sup\_max}} \\ 1 - 2 \cdot \frac{|T_{\text{supply}} - T_{\text{sup\_min}}|}{T_{\text{sup\_max}} - T_{\text{sup\_min}}}, & T_{\text{sup\_min}} \leq T_{\text{supply}} \leq T_{\text{sup\_max}} \end{cases} \quad (6)$$

where $T_{\text{sup\_min}} = 40$ °C and $T_{\text{sup\_max}} = 60$ °C define the temperature bounds, to ensure that rewards are maximized at lower temperatures and minimized at higher temperatures, with linear normalization in between.

3.3.2. Action masking logic

Action Masking is a technique used in reinforcement learning to improve action selection by dynamically disabling invalid or suboptimal actions at each decision step. Action masking reduces the variability in updates by focusing on valid actions, which can counterbalance potential instability caused by FL in the learning process [55].

By "masking out" invalid actions (assigning them near-zero probability), the model is restricted to sample only from the set of valid actions. This reduces exploration noise, speeds up learning, and ensures the agent does not learn to select infeasible or undesirable actions, which ultimately leads to more efficient and effective policies.

In this research, action masking is applied to restrict the action space based on observed state conditions, with the primary goals of ensuring DHW thermal comfort and encouraging simultaneous charging of DHW storage tanks which leads to saving energy [10]. This approach enables agents to select control actions that align with both the current state and anticipated future demand, without requiring knowledge of other agents' actions.

The masking is performed at each step based on the current DHW conditions (s_0) and next-hour DHW demand (s_3), allowing only valid actions to be selected.

- If the future demand indicator, s_3, is set (1) and the DHW tank temperature is $s_0 \leq 50$ °C, loading options (a_1 and a_2) are prioritized, to make sure water is warm for a future DHW demand in the next hour.
- If the DHW tank temperature, s_0, is above 50 °C, it means that the it is sufficiently warm. Therefore, guaranteed loading option (a_1) is masked out.
- If the DHW tank temperature, s_0, falls below 45 °C, only guaranteed loading option (a_1) is allowed, to ensure the temperature will stay above minimal comfort threshold of 40 °C.

By dynamically aligning agent actions with observed state variables, including real-time DHW conditions and anticipated demand, this approach is expected to result in simultaneous charging across multiple agents to decrease energy use and maintains DHW comfort levels. More importantly, this method eliminates the need for a multi-objective reward function to balance comfort and energy savings – typically conflicting objectives – which simplifies the learning process. Finally, by preventing the agents from sampling irrelevant actions, it accelerates training.

3.3.3. Proposed Federated Proximal Policy Optimization (FPPO)

This section introduces the proposed FRL method, Federated Proximal Policy Optimization (FPPO), for distributed training of control agents. We begin by detailing the core DRL algorithm, Proximal Policy Optimization (PPO), and its integration with the FedProx method (Section 2.2.2).

Proximal Policy Optimization (PPO). PPO [35] is a widely used model-free DRL algorithm from the on-policy policy gradient family. PPO optimizes a new control policy by minimizing an objective function that remains close to the original policy to ensure stable updates.

There are two main PPO variants designed to improve update stability and reliability:

1. *PPO with Kullback–Leibler (KL) Divergence*: This variant constrains policy updates by penalizing large deviations from the previous policy. The objective function, $L_i^{KL}(\theta)$, combines the estimated advantage ($\hat{A}$) with a KL penalty term to limit excessive updates. Given the probability ratio of $r_i(\theta) = \frac{\pi_\theta(a_i|s_i)}{\pi_{\theta_{old}}(a_i|s_i)}$, the KL-penalized objective ($L_i^{KL}(\theta)$) is:

$$L_i^{KL}(\theta) = \mathbb{E}\left[r_i(\theta)\hat{A}_i - c_{KL} KL[\pi_{\theta_{old}}(.|s_i), \pi_\theta(.|s_i)]\right] \quad (7)$$

   where c_{KL} adjusts the influence of the KL divergence penalty.
2. *PPO with Clipping*: This variant directly restricts policy updates by using a clipping mechanism to control deviations from the previous policy. With a clipping parameter ϵ, the objective function ($L_i^{CLIP}(\theta)$) is:

$$L_i^{CLIP}(\theta) = \mathbb{E}\left[min\left(r_i(\theta)\hat{A}_i, clip(r_i(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i\right)\right] \quad (8)$$

   When neural networks are used to approximate the policy and value functions, the loss function is further modified to combine $L_i^{CLIP}(\theta)$ with a value function error term ($L_i^{VF}(\theta) = (V_\theta(s_i) - V_i^{target})^2$). The resulting objective is:

$$L_i^{CLIP+VF+\mathbb{S}}(\theta) = \mathbb{E}\left[L_i^{CLIP}(\theta) - c_{VF} L_i^{VF}(\theta) + c_{ent}\mathbb{S}[\pi_\theta](s_i)\right] \quad (9)$$

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

To achieve stable and effective learning in this study, we use a combined PPO loss function ($L_i^{PPO}(\theta)$), shown in Eq. (10).

$$L_i^{PPO}(\theta) = \mathbb{E}\left[L_i^{CLIP}(\theta) - c_{VF} L_i^{VF}(\theta) + c_{ent}\mathbb{S}[\pi_\theta](s_i) - c_{KL} KL[\pi_{\theta_{old}}(.|s_i), \pi_\theta(.|s_i)]\right] \quad (10)$$

In this research, PPO operates within an actor–critic framework, which combines a policy network (actor) with a value function network (critic) for better stability and efficient learning. For readability, we refer to this actor–critic PPO variant as PPO throughout the paper.

Proposed PPO loss function. The combined PPO loss function $L_i^{PPO}(\theta)$ (Eq. (10)) balances policy optimization and exploration by incorporating both a clipping mechanism and a KL divergence penalty. The clipping function ensures that policy updates remain close to the previous policies, while the KL divergence term penalizes excessive deviation

from the original policy, providing additional stability. However, in federated settings, the diverse conditions across local environments can cause substantial variation in agent policies, potentially destabilizing the learning process.

To address this, we introduced a FedProx-based proximal term into the combined PPO objective. This addition regularizes each agent's local updates in relation to the global model parameters, improving convergence, and stabilizing learning by controlling the degree of deviation in non-IID environments.

The customized loss function $\mathcal{L}_i^{\text{FPPO}}(\theta_i)$ for each agent i is given by:

$$\mathcal{L}_i^{\text{FPPO}}(\theta_i) = L_i^{\text{PPO}}(\theta_i) + \frac{\mu}{2}|\theta_i - \theta|^2 \tag{11}$$

where $L_i^{\text{PPO}}(\theta_i)$ is the standard PPO loss for agent i with local model parameter θ_i, θ is the aggregated global model parameter, and $\mu \geq 0$ is the proximal term coefficient that is either a fixed number or it can be dynamically adjusted based on training performance. The proximal term $\frac{\mu}{2}|\theta_i - \theta|^2$ penalizes large deviations from the global model to help agents converge more consistently despite differing local conditions.

Dynamic adjustment of proximal term coefficient μ. The FPPO framework incorporates FedProx with a dynamic μ, which is adjusted based on training performance to ensure stable updates. If the training loss increases, μ is increased by 0.1, imposing tighter regularization to stabilize updates, i.e., forcing the updates to be close to the starting point. Conversely, if the loss decreases, μ is reduced by 0.1, allowing more flexibility for faster convergence. This adjustment value of 0.1 is based on the best practices established in the original FedProx paper [38], where it was demonstrated to effectively balance stability and convergence across a range of scenarios. The parameter μ is initially set to 1, and the idea behind its dynamic tuning during training is to ensure adaptability to evolving conditions, and enhances the framework's overall performance.

3.3.4. Learning and aggregation process in FRL

Each agent, deployed within an individual dwelling, learns a policy based on local interactions with the environment, which varies due to differences in user profiles. Due to these variations, an agent trained in one dwelling may struggle to adapt and perform effectively in another. While independent learning for each agent is possible, it is neither cost-effective nor efficient due to time constraints. Instead, the FRL approach allows agents to share only their policy model parameters, such as weights, which a central aggregator combines to avoid raw data sharing and thus preserve privacy.

The FRL approach follows a FL process that allows DRL agents to collaboratively optimize their policies while preserving data privacy:

1. *Local training*: Each agent independently trains a local model by interacting with its environment to learn an optimal policy.
2. *Parameter sharing*: After each episode, agents send their updated model parameters to a central aggregator.
3. *Global model update*: The aggregator combines the local updates into a global model.
4. *Model distribution*: The updated global model is sent back to each agent, which then refines its local policy in the next training iteration.

An overview of this process within the CHDC simulation environment is illustrated in Fig. 3.

3.4. Key performance indicators

Several Key Performance Indicators (KPIs) are used for evaluation. The DHW comfort is indicated using average duration of discomfort ($t_{DHW,dc}$) [%]. To make it comprehensible, $t_{DHW,dc}$ is demonstrated in Fig. 4 where the average of all tapping periods for a storage tank is

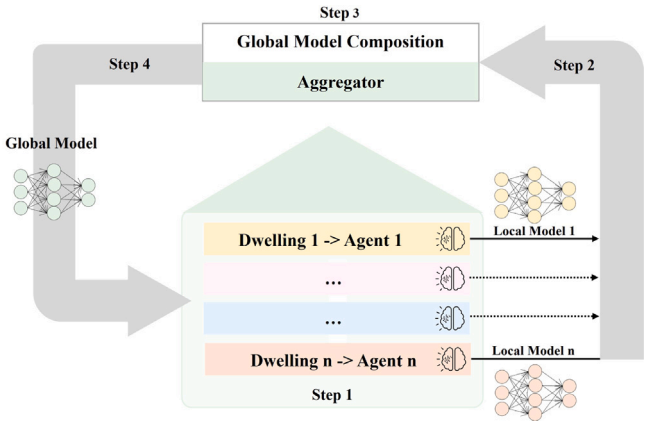


Fig. 3. An overview of FRL-based approach for control strategy optimization in the CHDC simulation environment.

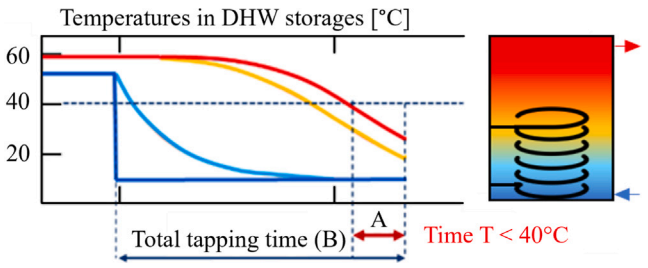


Fig. 4. DHW discomfort duration KPIs. The colors refer to temperatures in different layers of the storage tanks. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

displayed. The used temperature for DHW (with a setpoint of 58 °C) is the top temperature in storage tank which is shown in red. The temperatures of other storage tank layers are shown on the graph in orange, light blue, and dark blue. Given an ideal mixing valve at tapping points, DHW temperature should always be 40 °C, hence a top temperature below 40 °C is regarded as DHW discomfort. Given total discomfort time (A) and total tapping time (B), relative discomfort duration is A/B[%]. A lower percentage indicates a low duration of experiencing discomfort by occupants [10,56].

The energy performance of the system is evaluated using the total Primary Energy use (PE_{use}) [kWh] of central production [50,56] where a conversion factor of 2.5 is used for converting electricity to primary energy (PE) [57]. This KPI is used to calculate the energy savings [%].

Finally, as the grouped charging of the storage tanks is the desired pattern of reloading them, the degree of coordination is quantified using a novel Coordination Score (C_{score}). This score is defined as the proportion of dwellings taking coordinated actions (a_1 or a_2) at each time step, given that at least one dwelling has initiated charging (a_1).

4. Experiments and discussion

The effectiveness of the proposed approach was evaluated through a series of experiments in a simulation environment (based on the data from Belgium) using distinct training and testing periods. Training was performed in January, February, and March, while testing took place in October, November, and December (high-demand periods for heating in Belgium). Each DRL agent utilized a neural network with two hidden layers of 64 units, a discount factor (γ) of 0.997, and a learning rate schedule starting at 0.005 and decaying to 0.000001. First, Section 4.1 presents the experiments and their corresponding results. Then, Section 4.2 provides a discussion of the findings.

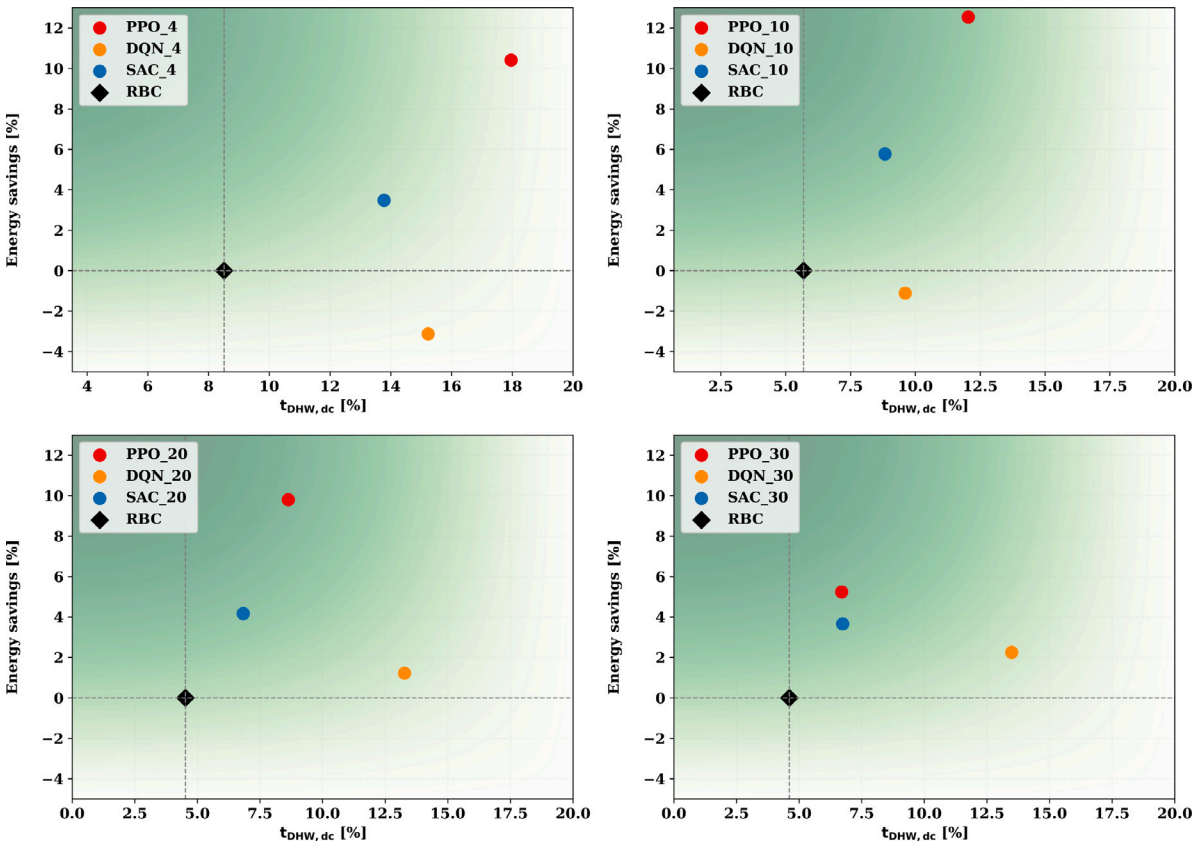


Fig. 5. Performance comparison of FRL methods (FedAvg combined with PPO, DQN, and SAC) and RBC in balancing energy savings [%] and $t_{DHW,dc}$ across varying dwelling scales. The green-highlighted top-left region represents the optimal balance of higher energy savings and lower thermal discomfort. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.1. Experiments

Key experiments included comparisons between FRL approaches using FedProx and FedAvg, assessments across different types and numbers of agents, and performance benchmarking against a RBC method [10] (Section 2.3.1), which uses a two-sensor control strategy to adjust central supply temperatures ($T_{sup,SP}$) based on DHW demand. These experiments aimed to analyze scalability, efficiency, and the impact of FedProx's proximal term coefficient μ.

4.1.1. Evaluation of FedAvg with PPO, DQN, and SAC

The initial set of experiments employed the FedAvg algorithm (i.e., FedProx with $\mu = 0$) in combination with three well-known DRL algorithms: PPO, DQN, and SAC. These experiments were conducted using varying numbers of dwellings (4, 10, 20, and 30) to assess performance across different scales. To analyze the results across different number of dwellings, Fig. 5 illustrates the performance of FRL methods and RBC in balancing the energy savings [%] and $t_{DHW,dc}$. The results reveal the differences in performance when combining FedAvg with PPO, DQN, and SAC across varying scales of apartment buildings. The goal is to be closer to the top-left region of the graphs, highlighted in green, which represents higher energy savings and lower discomfort. PPO excels at larger scales (PPO_20 and PPO_30) by achieving the best energy savings while maintaining low discomfort levels. Its scalability makes it a top performer by positioning it closer to the optimal green zone. SAC performs consistently well across all scales, maintaining good energy savings and low discomfort levels, positioning it reliably in the green zone. DQN improves a bit at larger scales (DQN_20 and DQN_30), but remains farther from the optimal region compared to PPO and SAC. RBC achieves minimal discomfort. The key observations are as follows:

- Energy savings: FRL methods, specifically those utilizing PPO and SAC, demonstrated notable energy savings compared to the RBC. PPO achieved the highest savings in smaller-scale systems, such as 4 and 10 dwellings, with percentages reaching 10.41% and 12.54%, respectively. SAC provided more consistent but slightly lower savings, particularly in larger dwelling scales. DQN's performance was less reliable, showing negative savings in smaller scale scenarios and slight improvements for larger dwelling numbers.
- DHW discomfort: It is measured via $t_{DHW,dc}$, varied across methods, with lower values indicating reduced discomfort.
  - PPO showed higher discomfort in smaller dwelling scenarios, with $t_{DHW,dc}$ at 17.96% for 4 dwellings, but improved significantly as the number of dwellings increased, achieving a discomfort level of 6.7% for 30 dwellings. This indicates that PPO becomes more effective at reducing discomfort in larger systems.
  - SAC delivered relatively low discomfort values across all scales, starting at 13.78% (4 dwellings) and improving slightly to 6.74% (30 dwellings), which shows its ability to maintain low discomfort levels consistently for more than 10 dwellings.
  - DQN displayed discomfort levels between 9.61% and 15.23%. Therefore, it is less effective compared to PPO and SAC in achieving lower discomfort overall.
  - RBC achieved the lowest discomfort levels, ranging from 4.61% (30 dwellings) to 8.51% (4 dwellings). However, achieving these levels comes at the expense of increased energy use and reduced savings.

Table 2
Experiments with FedProx using the PPO algorithm (FPPO). The blue section of the table corresponds to FPPO experiments with a fixed $\mu = 1$, while the remaining section presents results for FPPO with a dynamically adjusted μ where column *Rounds* is the rounds to adjust μ.

Experiment name	Number of dwellings	Rounds	Energy savings [%]	$t_{DHW,dc}$ [%]	$t^{Max}_{DHW,dc}$ [%]
FPPO_20	20	0	9.87	8.8	15.37
FPPO_30	30	0	6.34	6.71	12.29
FPPO_20_1	20	1	9.74	8.66	14.97
FPPO_30_1	30	1	8.34	6.7	14.2
FPPO_20_2	20	2	10.05	8.49	14.13
FPPO_30_2	30	2	4.93	6.63	11.68
FPPO_20_3	20	3	10.08	8.72	15.06
FPPO_30_3	30	3	8.14	8.46	14.75

These results indicate that while PPO and SAC achieved relatively low discomfort and notable energy savings, DQN lagged behind in both mitigating discomfort and achieving energy savings, particularly for scenarios with 4 and 10 dwellings where it failed to save energy. Additionally, the maximum DHW discomfort ($t^{Max}_{DHW,dc}$) was generally higher for PPO in smaller systems, reaching 20.67% for 4 dwellings, but improved significantly to 11.09% for 30 dwellings which reflects better performance in larger-scale applications. SAC maintained a more stable maximum discomfort profile across all scales, ranging between 11.82% (30 dwellings) to 15.14% (4 dwellings). DQN, however, exhibited inconsistent maximum discomfort levels (ranging from 15.49% to 20.48%), which remained higher than those of PPO and SAC, particularly in larger systems.

The poor performance of DQN in the FRL setup can be attributed to several factors. First, the difference across local environments (i.e., dwellings) made it difficult for DQN to converge reliably. FedAvg's focus on model parameter aggregation disrupted the stability of DQN's value-function updates, that led to suboptimal policies. Second, although both SAC and DQN rely on replay buffers, SAC's actor–critic structure and use of entropy regularization offers some advantage. SAC's replay buffer supports simultaneous updates of the actor (policy) and critic (value function), which helps stabilize learning and encourages exploration. DQN, on the other hand, depends solely on its replay buffer to update Q-values, lacking an actor to guide the exploration. This reliance makes DQN more vulnerable to the isolated and dynamic environments which are typical in federated settings. While parameter aggregation in FedAvg disrupts DQN's ability to maintain consistent Q-values, SAC's coupling of actor and critic updates allows it to generalize better across diverse agents.

Based on these findings, it is observed that the number of dwellings should be at least 20 for consistent and meaningful application of FL. Among the tested algorithms, PPO appeared as the best performer, since it effectively balanced energy savings and comfort. While discomfort remained slightly higher than RBC, the independent operation of agents in FRL explains this trade-off, as agents lacked knowledge of other dwellings' states and actions. Subsequent experiments focused on PPO with 20 and 30 dwellings to further evaluate its performance with FedProx in larger-scale settings.

4.1.2. Evaluation of FPPO with fixed μ

The next phase involved using the FedProx algorithm with PPO (FPPO), with a fixed μ (set to 1). This FRL approach was evaluated using 20 and 30 dwellings, referred to as FPPO_20 and FPPO_30, respectively. As shown in the blue part of the Table 2, the results for FPPO shows similarities with FedAvg (PPO_20 and PPO_30 in Fig. 5). Both approaches achieved closely comparable energy savings and discomfort levels across different scales of dwellings. For instance, in the FedAvg experiments, PPO_20 achieved 9.8% energy savings with 8.63% average DHW discomfort, while FPPO_20 in the FedProx experiments achieved 9.87% energy savings with 8.8% DHW discomfort. Similarly, PPO_30 in FedAvg had 5.24% energy savings with 6.7% discomfort, and FPPO_30 in FedProx had 6.34% energy savings with 6.71% discomfort. Thus, with the same level of discomfort, FPPO_30 saved 1.1% more energy than PPO_30.

The primary benefit of using FedProx with a fixed $\mu > 0$ over FedAvg (FedProx with fixed $\mu = 0$) lies in its ability to provide slightly better performance, particularly in larger systems. This slight improvement can be notable in real-world applications where maintaining a balance between energy savings and user comfort is essential. However, the differences are subtle, and both methods appear to be effective in their respective evaluations. This indicates the potential to further enhance the FedProx-based FRL approach by adapting the value of μ, which is addressed in the next subsection.

4.1.3. Evaluation of FPPO with dynamically adjusted μ

To further improve the performance of FPPO approach, dynamic adjustment of μ is introduced. Starting with $\mu = 1$, the value was adjusted based on loss trends across rounds of loss calculation (column *Rounds* in Table 2). Experiments tested adjustments every round, every 2 rounds, and every 3 rounds for 20 and 30 dwellings. Dynamic adjustment of μ and the frequency of adjustments impacted performance in several ways.

Dynamic adjustment of μ allowed the algorithm to adjust more effectively to varying conditions (Section 3.3.3). Higher μ encouraged alignment with the global model, enhancing stability, while lower μ enabled local adaptation, to capture dwelling-specific subtle distinctions or variations. This adaptability enabled FPPO to further optimize the balance between energy savings and user comfort.

The results indicate that while dynamic μ adjustment enhances FPPO performance compared to a fixed μ, the exact timing of these adjustments (i.e., change rounds) has a limited impact for 20 dwellings but shows a more noticeable effect for 30 dwellings. Fig. 6 illustrates the impact of dynamic μ adjustment on energy savings and DHW discomfort. The key observation is that while the performance differences across different change rounds are minimal for 20 dwellings, for 30 dwellings, the impact is more pronounced. For instance, for 20 dwellings:

- Adjusting μ every round resulted in 9.74% energy savings, 8.66% DHW discomfort, and maximum discomfort of 14.97%.
- Adjusting μ every 2 rounds led to 10.05% energy savings, 8.49% DHW discomfort, and maximum discomfort of 14.13%.
- Adjusting μ every 3 rounds achieved 10.08% energy savings, 8.72% DHW discomfort, and maximum discomfort of 15.06%.

However, for 30 dwellings:

- Adjusting μ every round provided 8.34% energy savings, 6.7% DHW discomfort, and maximum discomfort of 14.2%.
- Adjusting μ every 2 rounds resulted in 4.93% energy savings, 6.63% DHW discomfort, and maximum discomfort of 11.68%.
- Adjusting μ every 3 rounds led to 8.14% energy savings, 8.46% DHW discomfort, and maximum discomfort of 14.75%.

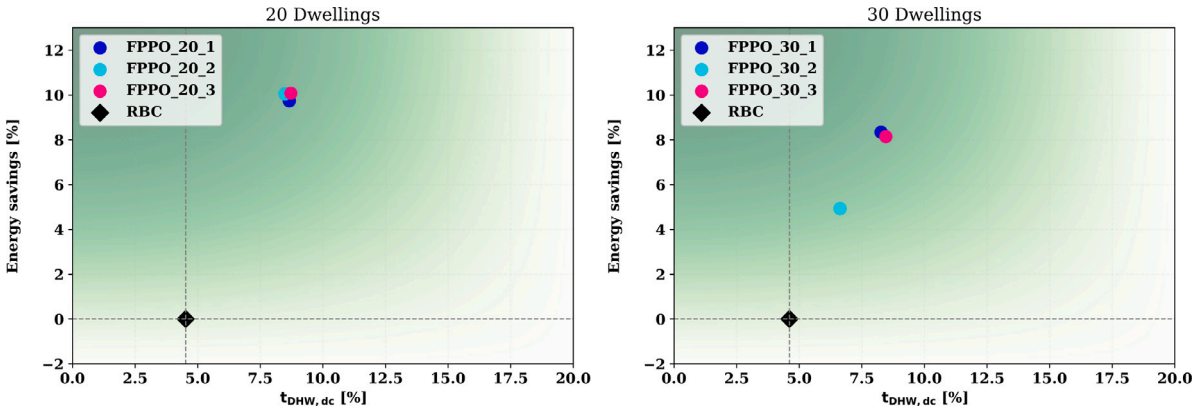


Fig. 6. Impact of dynamic μ adjustment on energy savings and $t_{DHW,dc}$. The green-highlighted top-left region represents the optimal balance of higher energy savings and lower thermal discomfort. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

These results show that for 20 dwellings, the choice of change rounds has only minor variations, whereas for 30 dwellings, adjusting every 2 rounds led to significantly lower energy savings, suggesting that change rounds can influence performance in larger systems. This suggests that the frequency of μ adjustments can influence performance in larger scales, as it may introduce greater variance in local updates due to increased heterogeneity among dwellings.

Averaging the results across 20 and 30 dwellings, FPPO with μ adjustment *Round* of one achieves the best balance, with 9.04% energy savings and 7.68% DHW discomfort, making it the most effective configuration. It is worth noting that FPPO with a fixed μ ($Rounds = 0$) across 20 and 30 dwellings, achieved an average of 8.11% energy savings and 7.76% DHW discomfort.

Extreme discomfort levels were better managed compared to fixed μ, with maximum discomfort capped at 14.97% for 20 dwellings and 14.2% for 30 dwellings when adjusting every round. Adjusting μ every 2 rounds for 30 dwellings showed a notable decrease in maximum discomfort to 11.68%, though this came at the expense of lower energy savings.

When comparing FPPO with dynamic μ to FedAvg and FPPO with fixed μ, several benefits emerge. Dynamic μ allows for better adaptability to changing conditions, improving both energy savings and comfort levels. The dynamic adjustment of μ helps to maintain consistent performance across different scales, addressing the limitations observed with fixed μ. The dynamic tuning approach achieves higher energy savings and lower discomfort levels compared to both FedAvg and FedProx with fixed μ, particularly in larger systems, i.e., 30 dwellings.

This is also reflected in Fig. 6, where the results cluster in the green region. FPPO consistently save more energy than RBC and demonstrates strong performance at large scales, where achieving a balance between energy efficiency and comfort is typically more challenging due to several factors:

- Larger systems involve more dwellings, each with unique usage patterns, thermal properties, and user behaviors. This adds complexity to the optimization process.
- With more dwellings, there is increased variability in demand and supply, as well as fluctuations in external factors (e.g., occupancy). This variability makes it harder to maintain consistent performance.
- At larger scales, the diversity and variability in energy demand, dwelling characteristics, and user preferences amplify the trade-offs between minimizing energy use and maintaining user comfort. Sophisticated algorithms, such as FRL, are necessary to address these challenges, as they must effectively optimize competing objectives across a larger system.

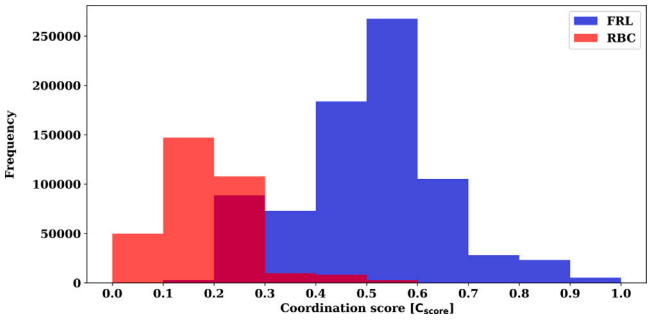


Fig. 7. Histogram showing the distribution of coordination scores (C_{score}) for FRL (FPPO_20_1) and RBC. FRL achieves higher C_{score} more frequently, indicating better alignment of actions among dwellings compared to the RBC, without compromising the privacy.

Overall, the dynamic tuning of μ in FPPO demonstrates noteworthy potential for optimizing FRL in CHDC, and provides a more balanced and adaptable approach compared to the (traditional) methods which often overlook the privacy of user's data.

4.1.4. Analysis of implicitly learned coordination among dwellings

FL aggregates knowledge from multiple dwellings (i.e., environments), while action masking restricts the action space based on observed state conditions to ensure DHW thermal comfort and encourage simultaneous charging of DHW storage tanks. This simultaneous charging leverages the already requested high supply temperature when at least one agent is charging, leading to significant energy savings. In FPPO, FL and action masking together create a privacy-aware synergy, enabling the development of a global policy that balances energy efficiency, comfort, and implicitly learned coordination among dwellings, without requiring DRL agents to know the actions of others.

To demonstrate the implicitly learned coordination among dwellings, the alignment of actions across dwellings is compared between FRL (FPPO_20_1) and RBC, and is visualized in Fig. 7, where x-axis shows the coordination score C_{score} and y-axis displays the frequency. This figure illustrates the frequency of coordinated actions which reflects the extent to which dwellings align with the desired group charging pattern that contributes to energy savings. Specifically, the desired pattern is achieved when, if one dwelling initiates charging (option a_1), it is optimal for the others to also charge (option a_1 or a_0). As it is shown in Fig. 7, FRL demonstrates significantly higher coordination which indicates better alignment of actions among dwellings and more frequent group charging behavior. In contrast, RBC is concentrated in the lower C_{score} range. FRL achieves coordination scores above 0.4 far more frequently which shows its superior capacity

for learning to coordinate the actions in decentralized systems without relying on knowledge of other dwellings' actions. These results highlight FRL's effectiveness in managing multi-dwelling coordination and its advantage over RBC in encouraging optimal group behavior.

The histogram clearly shows that FRL (FPPO_20_1) strategy ensures that when guaranteed loading option (a_1) is picked by one agent, the other agents often do not pick option a_0, which aligns perfectly with the intended group charging behavior, as described in Section 3.3.2. Interestingly, FRL appears to request storage tank reloading far more frequently, which increases the supply temperature to the highest level (65 °C). Despite this, it achieves lower energy use compared to RBC.

This behavior could be attributed to the agents learning to strategically reload tanks at optimal times. For instance, reloading when the supply temperature is already relatively high (such as when another dwelling's storage tank is being reloaded, triggering a high $T_{sup,SP}$ request), requires less energy to reach the maximum supply temperature (65 °C) compared to reloading at lower temperatures. Over time, the agents have learned that it is more energy-efficient in the long term to charge the DHW storage tank when the supply temperature is high, even if the water is still not cold. This capability is enhanced by the agents' access to future demand information from the observation space and the guidance of the global reward function.

4.2. Discussion

In this section, the experimental results are discussed to gain insights into the associated trade-offs. Additionally, a potential direction for further improvements is provided.

4.2.1. Understanding the trade-offs

FRL methods inherently prioritize privacy by design. This privacy-preserving approach ensures that sensitive data from individual dwellings is not shared directly, but rather through aggregated model updates. This contrasts with conventional RBC systems, which may not have the same level of privacy considerations. The slight increase in discomfort observed in FRL compared to RBC can be explained by the following points:

- Privacy preservation: FRL methods ensure that individual data, such as usage patterns (s_3), number of occupants present at the house (s_2) and other sensitive data, do not leave the dwelling and remain private. This is achieved by only sharing model parameters rather than raw data, which can lead to less precise adjustments and slightly higher discomfort levels.
- Decentralized learning: In FRL, each dwelling's local model is trained independently before being aggregated into a global model. This decentralized approach can introduce variability and less optimal coordination compared to a RBC or a centralized system, which has a holistic view of all dwellings.
- Scalability: While FRL methods are designed to adapt to diverse and dynamic environments, the trade-off is that they may not always achieve the same level of comfort as RBC systems, specially when the number of participating agents is low. However, FRL methods perform very well as the number of participating agents increases. This is because the decentralized nature of FRL allows it to leverage the diverse experiences and data from a larger pool of agents which leads to more generalized models. In larger systems, FRL can effectively manage the complexity and variability, and ensure better overall performance in terms of energy savings and maintaining acceptable comfort levels. This scalability makes FRL particularly suitable for applications involving a high number of agents, where RBC and traditional centralized methods might struggle to maintain efficiency and adaptability.
- Energy and comfort trade-offs: While RBC ensures the lowest discomfort levels, it does so at the expense of significantly higher energy use. As a FRL method, FPPO finds a balance to achieve notable energy savings (using the reward function) while keeping discomfort at manageable levels (using action masking).

Table 3
Incorporating a penalty for discomfort to the reward function in FPPO_20_1.

Penalty	Energy savings [%]	$t_{DHW,dc}$ [%]	$t^{Max}_{DHW,dc}$ [%]
0	9.74	8.66	14.97
−0.2	6.13	6.61	11.93
−1	6.22	6.41	11.49

Overall, the slight increase in discomfort with FRL is a trade-off for enhanced privacy and scalability. While some degree of suboptimality in DHW thermal comfort may occur, action masking mitigates it as much as possible, and the global reward function promotes collective energy efficiency. As a result, the learned global policy remains scalable, adaptive, and energy-efficient. As FRL methods continue to evolve, it is expected that these discomfort levels can be further minimized while maintaining the privacy and efficiency benefits.

It is worth mentioning that, based on the experimental results, while additional comparisons with other FL methods may offer further insights, the consistency of our findings suggests they are unlikely to significantly alter the core conclusions of this study. Nevertheless, we acknowledge the potential value of exploring alternative FL approaches and consider this a potential direction for future research.

4.2.2. Incorporating a penalty for discomfort

In the previous setup, the objective of the reward function R (Eq. (6)) was optimizing global energy use, with action masking as the primary mechanism for ensuring DHW thermal comfort. However, FRL exhibited slightly higher discomfort levels compared to the RBC method. To address this, a penalty term to the reward function is introduced which is based on local observations (s_0), aiming to reduce discomfort more effectively. A penalty was applied when s_0 fell below 40 °C, i.e., the DHW discomfort threshold. Two configurations were tested on one of the best-performing methods, FPPO_20_1, by adding penalties of -0.2 and -1 to R.

As shown in Table 3, the addition of the penalty term effectively reduced both average and maximum discomfort levels, with the higher penalty (-1) showing slightly more improvement. For a lower penalty of -0.2, the energy savings were 6.13%, with an average DHW discomfort of 6.61% and a maximum discomfort of 11.93%. In comparison, the -1 penalty achieved slightly better comfort metrics, reducing average DHW discomfort to 6.41% and maximum discomfort to 11.49%, while slightly increasing energy savings to 6.22%.

Although energy savings remained stable across the two configurations with penalty, higher penalties improved overall comfort, suggesting that local penalties can effectively mitigate the discomfort associated with FRL with action masking compared to RBC. These findings highlight the potential of fine-tuning reward functions using local penalties to balance occupant comfort and energy savings. Therefore, while action masking shows strong results, the combination of local and global rewards is recommended for scenarios where satisfying constraints is critical.

Additionally, this suggests that an important area for future work is incorporating a local component in the reward design to further ensure comfort while exploring fair global reward distribution among individual agents. For instance, in a collective heating system, if one agent has significantly higher demand, a global reward structure that does not account for individual differences could either unfairly penalize or overly reward it.

5. Conclusion and future works

This study introduced Federated Proximal Policy Optimization (FPPO), which is a FRL approach to optimize decentralized control of DHW storage tanks in collective heating systems, leveraging Proximal Policy Optimization (PPO) within a privacy-aware design. The integration of FL with action masking proved effective in optimizing energy

savings and DHW thermal comfort. Using FedAvg, PPO and SAC consistently outperformed DQN across all experiments, with PPO excelling in scalability and SAC offering stable performance across varying dwelling scales and saving energy compared to the RBC. Moreover, the results highlight the importance of scale. With at least 20 dwellings, the FRL approach becomes more consistent and meaningful. This suggests that employing FL can be more beneficial in larger communities.

The proposed FPPO method with fixed μ, which combines PPO with FedProx, demonstrated superior performance compared to FedAvg-based FRL, particularly in larger systems (20 and 30 dwellings). FPPO achieved 9.87% energy savings with 8.8% discomfort for 20 dwellings and 6.34% energy savings with 6.71% discomfort for 30 dwellings. The dynamic adjustment of the μ further enhanced FPPO's adaptability in the system with 30 dwellings, resulting in a better balance between energy savings and comfort. For instance, adjusting μ every round improved energy savings to 8.34% and with a discomfort of 6.7% for 30 dwellings.

Building on these findings, while FRL methods managed to notably save energy, their discomfort levels do not match the very low discomfort levels of the RBC. This trade-off suggests that higher energy savings are possible if occupants are willing to tolerate slight discomfort. Moreover, this trade-off suggests that while FPPO can improve energy savings, there might be a need for further optimization to enhance user comfort. For example, FPPO with a penalty for discomfort reduced average discomfort to 6.41% while maintaining energy savings of 6.22%.

The combination of action masking and a global reward function enabled implicit coordination among dwellings, encouraging simultaneous charging of DHW storage tanks. This approach reduced energy use by leveraging high supply temperatures when at least one agent initiated charging, without requiring direct communication between agents. FPPO demonstrated higher coordination scores compared to RBC, indicating better alignment of actions among dwellings and more frequent group charging behavior.

An important area for future work is exploring fair reward distribution among individual agents, rather than providing the same reward to all. Moreover, in countries with volatile electricity prices, the reward function could be further optimized by incorporating day-ahead electricity market prices into DRL-based solutions, which is a factor typically not utilized by RBC systems. Future research should also explore real-world deployment using a two-phase approach: initial model training in simulation, followed by online adaptation, to enable dynamic policy refinement for improved adaptability and performance. Finally, pre-training on historical data (i.e., offline RL) could enhance the performance of FRL agents in accelerating learning and improving overall efficiency.

CRediT authorship contribution statement

Sara Ghane: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Stef Jacobs:** Writing – review & editing, Software, Investigation, Conceptualization. **Furkan Elmaz:** Writing – review & editing, Conceptualization. **Thomas Huybrechts:** Writing – review & editing, Project administration. **Ivan Verhaert:** Writing – review & editing, Supervision. **Siegfried Mercelis:** Writing – review & editing, Supervision, Project administration.

Funding

This work was partly funded by a PhD fellowship of the Research Foundation Flanders (FWO) [1S08624N].

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.

Appendix

FRL-based training algorithms for DHW storage tanks loading of multiple dwellings connected to a residential collective heating system, i.e., CHDC.

Algorithm 1. Local model training at each dwelling by a DRL agent.

1. time step t = 0
2. Initiate environment, hyperparameters, and state S_0
3. for episodes in range(max_episodes):
4. performs local RL training (takes action A_t, receives state S_{t+1} and reward R_{t+1})
5. if episode_is_finished:
6. sends local_model_weights to aggregator
7. receives global_model_weights from aggregator
8. updates its local_model_weights using global_model_weights
9. end
10. end

Algorithm 2. Global model generation by aggregator using Federated Learning.

1. for episodes in range(max_episodes):
2. receives local_model_weights of each agent
3. applies Federated Learning on agents local_model_weights
4. updates the global_model_weights
5. sends back the updated global_model_weights to all agents
6. end

Data availability

Data will be made available on request.

References

[1] European Environment Agency. Decarbonising heating and cooling — a climate imperative. 2023, https://www.eea.europa.eu/publications/decarbonisation-heating-and-cooling. [Online Accessed 05 December 2023].

[2] European Environment Agency. Greenhouse gas emissions from energy use in buildings in europe. 2022, https://www.eea.europa.eu/ims/greenhouse-gas-emissions-from-energy. [Online Accessed 30 July 2023].

[3] European Commission. Repowereu: affordable, secure and sustainable energy for Europe, https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal/repowereu-affordable-secure-and-sustainable-energy-europe_en. [Online Accessed 27 July 2023].

[4] Eurostat. Energy consumption in households. 2023, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Energy_consumption_in_households. [Online Accessed 30 May 2023].

[5] European Commission. A European green deal. 2021, https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en. [Online Accessed 27 July 2023].

[6] European Commission. Renovation wave, https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/renovation-wave_en. [Online Accessed 27 July 2023].

[7] Zhao B, Jin Y, Li W, Zheng H. Analysis on the technical situation and applied difficulties of district heating load forecasting. Therm Eng 2022;69:464–72.

[8] Lichtenwoehrer P, Erker S, Zach F, Stoeglehner G. Future compatibility of district heating in urban areas — a case study analysis in the context of integrated spatial and energy planning. Energy Sustain Soc 2019;9.

[9] Lund H, Werner S, Wiltshire R, Svendsen S, Thorsen JE, Hvelplund F, Mathiesen BV. 4Th generation district heating (4gdh). Energy 2014;68:1–11.

[10] Jacobs S, De Pauw M, Van Minnebruggen S, Ghane S, Huybrechts T, Hellinckx P, Verhaert I. Grouped charging of decentralised storage to efficiently control collective heating systems: Limitations and opportunities. Energies 2023;16(3435).

[11] Vaillant Rebollar J, Himpe E, Laverge J, Janssens A. Influence of recirculation strategies in collective heat distribution system on the performance of dwelling heating substations. In: VIII international conference for renewable energy, energy saving and energy education. 2015.

[12] Vaillant Rebollar JE, Himpe E, Laverge J, Janssens A. Sensitivity analysis of heat losses in collective heat distribution systems using an improved method of EPBD calculations. Energy 2017;140:850–60.

[13] EMIB-BiRD University of Antwerp. Infofiche: Resultaten in-situ meetcampagnes – deel 1 en 2. 2021, https://www.warmtenet.info/publicaties.html. [Online Accessed 27 July 2023].

[14] Arroyo J, Manna C, Spiessens F, Helsen L. Reinforced model predictive control (RL-MPC) for building energy management. Appl Energy 2022;309:118346.

[15] Raman NS, Devraj AM, Barooah P, Meyn SP. Reinforcement learning for control of building hvac systems. In: 2020 American control conference. ACC, IEEE; 2020, p. 2326–32.

[16] Nagy Z, Henze G, Dey S, Arroyo J, Helsen L, Zhang X, Chen B, Amasyali K, Kurte K, Zamzam A, Zandi H, Drgoňa M, McCullogh S, Park JY, Li H, Hong T, Brandi S, Pinto G, Capozzoli A, Vrabie D, Bergés K, Marzullo T, Bernstein A. Ten questions concerning reinforcement learning for building energy management. Build Environ 2023;241:110435.

[17] Lee S, Choi DH. Federated reinforcement learning for energy management of multiple smart homes with distributed energy resources. IEEE Trans Ind Inform 2022;18:488–97.

[18] Gupta S, Bhambri S, Dhingra K, Buduru AB, Kumaraguru P. Multi-objective reinforcement learning based approach for user-centric power optimization in smart home environments. In: 2020 IEEE international conference on smart data services. SMDS, IEEE; 2020, p. 89–96.

[19] Ghane S, Jacobs S, Casteels W, Brembilla C, Mercelis S, Latré S, Verhaert I, Hellinckx P. Supply temperature control of a heating network with reinforcement learning. In: 2021 IEEE international smart cities conference (ISC2). IEEE; 2021, p. 1–7.

[20] Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: Concept and applications. ACM Trans Intell Syst Technol 2019;10:1–19.

[21] Qi J, Zhou Q, Lei L, Zheng K. Federated reinforcement learning: Techniques, applications, and open challenges. 2021, arXiv preprint arXiv:2108.11887.

[22] Fujita K, Fujimura S, Sun Y, Esaki H, Ochiai H. Federated reinforcement learning for the building facilities. In: 2022 IEEE international conference on omni-layer intelligent systems. COINS, IEEE; 2022, p. 1–6.

[23] Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: International conference on machine learning. PMLR; 2018, p. 1861–70.

[24] Lee JH, Park JY, Sim HS, Lee HS. Multi-residential energy scheduling under time-of-use and demand charge tariffs with federated reinforcement learning. IEEE Trans Smart Grid 2023;14:4360–72.

[25] Lee S, Xie L, Choi DH. Privacy-preserving energy management of a shared energy storage system for smart buildings: A federated deep reinforcement learning approach. Sensors 2021;21(4898).

[26] Gao J, Wang W, Nikseresht F, Govinda Rajan V, Campbell B. Pfdrl: Personalized federated deep reinforcement learning for residential energy management. In: Proceedings of the 52nd international conference on parallel processing. 2023, p. 402–11.

[27] Xia Y, Wang X, Yin X, Bo W, Wang L, Li S, Li K. Federated accelerated deep reinforcement learning for multi-zone hvac control in commercial buildings. IEEE Trans Smart Grid 2025.

[28] Tan M, Zhao J, Liu X, Su Y, Wang L, Wang R, Dai Z. Federated reinforcement learning for smart and privacy-preserving energy management of residential microgrids clusters. Eng Appl Artif Intell 2025;139:109579. http://dx.doi.org/10.1016/j.engappai.2024.109579, URL: https://www.sciencedirect.com/science/article/pii/S0952197624017378.

[29] Yang T, Xu Z, Ji S, Liu G, Li X, Kong H. Cooperative optimal dispatch of multi-microgrids for low carbon economy based on personalized federated reinforcement learning. Appl Energy 2025;378:124641. http://dx.doi.org/10.1016/j.apenergy.2024.124641, URL: https://www.sciencedirect.com/science/article/pii/S0306261924020245.

[30] Rezazadeh F, Bartzoudis N. A federated drl approach for smart micro-grid energy control with distributed energy resources. In: 2022 IEEE 27th international workshop on computer aided modeling and design of communication links and networks. CAMAD, 2022, p. 108–14. http://dx.doi.org/10.1109/CAMAD55695.2022.9966919.

[31] Li Y, He S, Li Y, Shi Y, Zeng Z. Federated multiagent deep reinforcement learning approach via physics-informed reward for multimicrogrid energy management. IEEE Trans Neural Netw Learn Syst 2024;35:5902–14. http://dx.doi.org/10.1109/TNNLS.2022.3232630.

[32] Mnih V, Badia AP, Mirza M, Graves A, Lillicrap TP, Harley T, Silver D, Kavukcuoglu K. Asynchronous methods for deep reinforcement learning. 2016, URL: https://arxiv.org/abs/1602.01783, arXiv:1602.01783.

[33] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M. Playing atari with deep reinforcement learning. 2013, arXiv preprint arXiv:1312.5602.

[34] Lowe R, Wu Y, Tamar A, Harb J, Abbeel P, Mordatch I. Multi-agent actor-critic for mixed cooperative-competitive environments. 2020, URL: https://arxiv.org/abs/1706.02275, arXiv:1706.02275.

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

[36] Fujimoto S, van Hoof H, Meger D. Addressing function approximation error in actor-critic methods. 2018, URL: https://arxiv.org/abs/1802.09477, arXiv:1802.09477.

[37] McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. PMLR; 2017, p. 1273–82.

[38] Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated optimization in heterogeneous networks. Proc Mach Learn Syst 2020;2:429–50.

[39] Fu Y, Li C, Yu FR, Luan TH, Zhang Y. A selective federated reinforcement learning strategy for autonomous driving. IEEE Trans Intell Transp Syst 2022;1–14.

[40] Konečnỳ J, McMahan HB, Ramage D, Richtárik P. Federated optimization: Distributed machine learning for on-device intelligence. 2016, arXiv preprint arXiv:1610.02527.

[41] Konečnỳ J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated learning: Strategies for improving communication efficiency. 2016, arXiv preprint arXiv:1610.05492.

[42] McMahan HB, Moore E, Ramage D, y Arcas BA. Federated learning of deep networks using model averaging. 2016, arXiv preprint arXiv:1602.05629. 2, 2.

[43] Sutton RS, Barto AG. Reinforcement learning: An introduction. In: Adaptive computation and machine learning ser. 2nd ed.. MIT Press; 2018.

[44] Wang Z, Hong T. Reinforcement learning for building controls: The opportunities and challenges. Appl Energy 2020;269:115036.

[45] Lei L, Tan Y, Zheng K, Liu S, Zhang K, Shen X. Deep reinforcement learning for autonomous internet of things: Model, applications and challenges. IEEE Commun Surv Tutor 2020;22:1722–60.

[46] Padakandla S. A survey of reinforcement learning algorithms for dynamically varying environments. ACM Comput Surv 2021;54:1–25.

[47] Li Y. Deep reinforcement learning: An overview. 2017, arXiv preprint arXiv:1701.07274.

[48] Shi Y, Yang K, Jiang T, Zhang J, Letaief KB. Communication-efficient edge AI: Algorithms and systems. IEEE Commun Surv Tutor 2020;22:2167–91.

[49] Yang Z, Chen M, Saad W, Hong CS, Shikh-Bahaei M. Energy efficient federated learning over wireless communication networks. IEEE Trans Wirel Commun 2021;20:1935–49.

[50] Van Riet F. Hydronic design of hybrid thermal production systems in buildings. Antwerp (Belgium: University of Antwerp; 2019.

[51] VLAIO b. Productie en distributie van Sanitair warm water: selectie en dimensionering (Dutch), TETRA 120145 (2012–2014), https://www.tetra-sww.be/.

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

[53] Solar Energy Laboratory Univof Wisconsin-Madison (SELUWM). Mathematical referenc - type 60: Stratified fluid storage tank with internal heat exchangers. volume 4. 2009, p. 390–6.

[54] Van Riet F, Steenackers G, Verhaert I. A new approach to model transport delay in branched pipes. In: Proceedings of the 10th international conference on system simulation in buildings. Liege, Belgium; 2018, p. 10–2.

[55] Sun Z, Niu X, Wei E. Understanding generalization of federated learning via stability: Heterogeneity matters. In: International conference on artificial intelligence and statistics. PMLR; 2024, p. 676–84.

[56] Jacobs S, De Pauw M, Hellinckx P, Verhaert I. Decentralized storage in combined heat distribution circuits: how to control? In: CLIMA 2022: the 14th REHVA HVAC world congress, 22nd-25th May, 2022. Rotterdam, The Netherlands; 2022, p. 1–8.

[57] International Organization for Standardization (ISO). Energy performance of buildings—overarching epb assessment—part. 1: General framework and procedures (iso 52000-1: 2017). 2017.