Supply temperature control of a heating network with reinforcement learning

Sara Ghane
IDLab - Faculty of Applied Engineering
University of Antwerp - imec
Antwerp, Belgium
sara.ghane@uantwerpen.be

Stef Jacobs
EMIB – Faculty of Applied Engineering
University of Antwerp
Antwerp, Belgium
stef.jacobs@uantwerpen.be

Wim Casteels
IDLab – Faculty of Applied Engineering
University of Antwerp - imec
Antwerp, Belgium
wim.casteels@uantwerpen.be

Christian Brembilla
EMIB – Faculty of Applied Engineering
University of Antwerp
Antwerp, Belgium
christian.brembilla@uantwerpen.be

Siegfried Mercelis
IDLab – Faculty of Applied Engineering
University of Antwerp - imec
Antwerp, Belgium
siegfried.mercelis@uantwerpen.be

Steven Latré
IDLab – Department of Computer Science
University of Antwerp - imec
Antwerp, Belgium
steven.latre@uantwerpen.be

Ivan Verhaert
EMIB – Faculty of Applied Engineering
University of Antwerp
Antwerp, Belgium
ivan.verhaert@uantwerpen.be

Peter Hellinckx
IDLab – Faculty of Applied Engineering
University of Antwerp - imec
Antwerp, Belgium
peter.hellinckx@uantwerpen.be

Abstract—**Heating networks are typically controlled by a heating curve, which depends on the outdoor temperature. Currently, innovative heating networks connected to low heat demand dwellings ask for advanced control strategies. Therefore, the potentials of reinforcement learning are researched in a heating network connected to a central heat pump and four dwellings. The comparison between a discrete and continuous action space is made with respect to the weight factor of the reward function. The results indicate that in both cases the reinforcement learning-based controlling of the supply temperature can generally ensure energy savings while keeping the occupant's temperature requirements in comparison to the rule-based controller.**

Keywords—Reinforcement learning, heating network, geothermal heat pump, Proximal Policy Optimization, supply temperature control

Nomenclature

CA	Continuous action space
DA	Discrete action space
DRL	Deep reinforcement learning
HVAC	Heating, ventilation, and air conditioning
MDP	Markov decision process
ML	Machine learning
PID	Proportional-integral-derivative
PPO	Proximal Policy Optimization
RBC	Rule-based controller
RES	Renewable energy sources
RL	Reinforcement learning
RNN	Recurrent neural network

I. Introduction

Thermal networks are large distribution networks, connecting multiple end-users to a common production unit. This facilitates the use of more renewable energy sources (RES), such as geothermal heat pumps. This paper takes into consideration a thermal network for space heating connected to dwellings where the domestic hot water usage is not taken into account. They are hereinafter referred to as 'heating networks'.

Currently, the supply temperature of heating networks is controlled according to a heating curve [1] which defines a supply temperature based on the outdoor temperature and the indoor temperature setpoint. A higher supply temperature corresponds with a lower outdoor temperature because of increased transmission losses. However, the insulation rate of dwellings is increasing and the share of space heating is decreasing in the total energy demand of dwellings [2]. Therefore, the internal heat gains (from electrical appliances and occupants) and solar heat gains might even be sufficient to compensate the low heat losses. Moreover, a lower heat demand facilitates the use of low-temperature emitters, which means that lower supply temperatures are possible. These lower distribution temperatures are beneficial for the efficiency of the central production unit and for decreasing the distribution losses, which in turn leads to additional energy savings. It is clear that the current heating curve might be out of date, as the heat demands in new buildings do not solely depend on the outdoor temperature anymore.

In this paper, we explore if a model-free deep reinforcement learning (RL) agent can improve the control of the supply temperature setpoint of a heating network with a central geothermal heat pump. Thus the main contribution here is to provide a heating network with an agent-based controller, instead of current rule-based controller which controls according to a heating curve. An improvement is here defined as a decreased energy usage of the heat pump while fulfilling the room temperature requirements of the end-users. RL is a subfield of machine learning (ML) with the objective of optimizing the behaviour (or policy) of an agent in a specific environment, based on a reward function. The agent perceives the state of the environment and every timestep it can perform one or more possible actions in the environment. Depending on the result of these actions, the agent receives a reward which results in a feedback loop that allows the agent to optimize its behaviour in order to maximize the reward. It is named as model-free, because the agent does not possess over a model of the controlled environment. A more detailed explanation for RL is given in section II.

While many RL-based control methods have been proposed for HVAC components, only a few studies and projects are devoted to optimising the control of heating networks. Reference [3] developed a framework to train deep RL agents that control an office room connected to a district heating system. The RL-agents control the setpoint of the mixing valve before their zone. In this way, the hot water of the heating network, which is non controllable by the agent, was mixed by the returning cold water after the radiator. The

2021 IEEE International Smart Cities Conference (ISC2) | 978-1-6654-4919-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISC253183.2021.9562966

978-1-6654-4919-9/21/$31.00 ©2021 IEEE

RL-agent is implemented in a case study and led to heating demand savings of 16.7% during the 78 days of testing.

Reference [4] proposed a data-driven deep RL (DRL) approach to optimally control the supply water temperature of a district heating. They focused on delivering heat to every apartment on the feed line and trained a recurrent neural network (RNN) on the simulated data (generated by Dymola) to predict indoor temperatures and return temperature. Then, this model is used to train two DRL agents (one with expert guidance) to provide optimal control for the supply water temperature, using a single-objective reward function based on the target temperature. While this is closely related to the current work, there exist several differences. Here, we define a multi-objective reward function and incorporate a state space with as less variables as possible to reduce the complexity. Besides, we consider several different setpoint temperatures for different apartments in different times of the day. Moreover, as opposed to our approach, the dynamics of the production units and storage tanks are not taken into consideration. Thus, if the RL-agents defines a temperature of e.g. 40 °C, the supply temperature is exactly equal to 40 °C. This is not realistic, certainly in case of a heat pump, as the energy usage highly depends on the available temperatures.

Besides previous researches, other projects are determined to use RL in the energy systems. Firstly, VITO, a research institute in Belgium, recently developed the STORM-controller to improve the management of heat loads in an existing district heating. Three optimisation strategies are implemented in the STORM-controller [5, 6], namely peak shaving, cell balancing and market interaction. The agents manipulate the perceived inputs to the current available proportional-integral-derivative (PID) controllers of the district heating. This RL-based controller is developed to control existing heating networks, instead of finding new control strategies to replace the commonly used PID controllers. Secondly, the CityLearn framework [7] is an interesting initiative to develop RL-agents for balancing the production and demand of the electrical grid, by using RL-agent to control the storage tanks or batteries. However, the heat demands of the buildings are based on heat loads. As a result, the effects of temperatures and mass flows are not taken into account. As this paper focusses on controlling the supply temperature to the heating networks, the CityLearn framework is not suitable for our research.

In the next section, a model-free deep RL technique is described. Afterwards, the used models in the simulation environment representing the heating network are shortly described in the third section. In sections IV and V, the training methods and results are presented and discussed, respectively. Finally, this research is concluded in the sixth section.

II. Reinforcement Learning

Reinforcement learning (RL) is a branch of ML which provides a mathematical framework for solving control problems sequential decision making using autonomous agents [8, 9]. This is done through the repeated interaction of an agent with its surrounding environment. The agent observes the current state of the environment and takes an action (i.e. making a control decision), to maximize a reward [10, 11]. This sequential decision making process could be formulated as a Markov decision process (MDP). An MDP is formed by a tuple (S, A, P, R, γ), where:

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

Any change in the above mentioned elements could potentially lead to a different RL implementation and thus different control mechanisms.

A policy π is the solution of MDP that maps states to actions. The performance of a policy in a given state, is represented as the state value function of a state, which is the expected accumulated reward obtained by the agent. Given $\forall s \in S$, the value function is defined as follows:

$$V^{\pi}(s) = R\big(s, \pi(s)\big) + \gamma \sum_{s'} P\big(s' | s, \pi(s)\big) V^{\pi}(s') \qquad (1)$$

The goal of an RL agent is to find the optimal policy, which provides an optimal action for each state. Two main approaches of RL algorithms to achieve this are model-based and model-free approaches. Model-based RLs are usually used when the transition probability and reward function of the environment are known. But, in most real-world problems, the characteristics of the environment are unknown to the agent and the optimal policy should be obtained by the agent through the interaction with the environment and without knowing the environment's dynamics.

In this study we use Proximal Policy Optimization (PPO) [12] which is a well-known model-free deep policy gradient RL algorithm. It tries to find a balance between ease of implementation, sample complexity, and ease of tuning by computing an update that minimizes the objective function while maintaining a relatively small deviation from the previous policy [13].

PPO performs multiple minibatch gradient updates, instead of one update per sample. It avoids too large policy updates by alternating between data sampling from the environment and optimizing a surrogate objective function with clipped probability ratios. With θ being the policy parameter, the probability ratio is calculated as the following:

$$r_t(\theta) = \frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{old}}(a_t | s_t)} \qquad (2)$$

Then, the clipped surrogate objective function of PPO is defined as the following equation:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\big[min\big(r_t(\theta)\hat{A}_t, clip\big(r_t(\theta), 1 - \epsilon, 1 + \epsilon\big)\hat{A}_t\big)\big] \qquad (3)$$

Where ϵ is a hyperparameter for determining the range of clipping (usually 0.1 or 0.2) and $\hat{A}_t$ is the estimated advantage. According to the above equation, the minimum of nonclipped $(r_t(\theta)\hat{A}_t)$ and clipped $(clip(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)$ objective is the final objective of the PPO.

III. Simulation environment

To train the RL agent, a huge amount of data is required. This data is provided by a simulation environment (i.e. a simulator) in MATLAB. A schematic overview of the associated energy flows as well as the heating curve used in this research are given in Fig. 1. In this research, the RL-agent

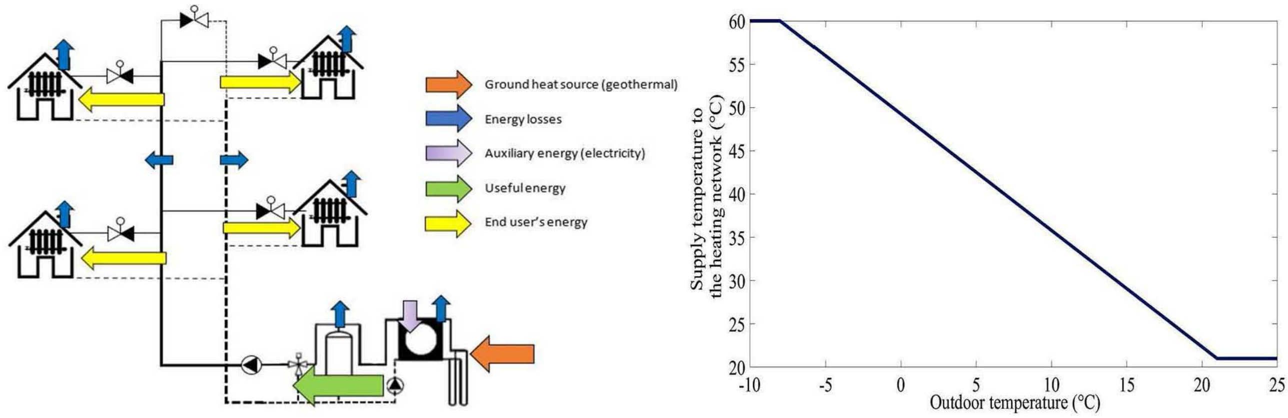
Ground heat source (geothermal)
Energy losses
Auxiliary energy (electricity)
Useful energy
End user's energy
Supply temperature to the heating network (°C)
Outdoor temperature (°C)
60
55
50
45
40
35
30
25
20
-10
-5
0
5
10
15
20
25

Fig. 1. On the left: "Overview of energy flows in the considered case study. An apartment building with four dwellings and a central geothermal heat pump." On the right: "The heating curve which is currently utilized by the rule-based controller. This is used as the baseline for the experiments."

controls the supply temperature of a heating network in an apartment building with four dwellings, inhabited by different families. The central heat production unit is a geothermal heat pump, which is connected to a storage tank. The energy losses of all components are taken into account and only the central heat pump uses energy (electricity). The heating network is assumed to be balanced correctly, so that the pressures and pressure drops are not taken into account in the simulation environment. The dwellings extract the needed energy out of the heating network by a throttle used to control the mass flow. The following part will shortly introduce the MATLAB simulation environment. More in-depth information about the used models can be found in [14-16].

The user behaviour is based on a profile generator [17] of INSTAL2020. In the latter study, a survey is held on 700 Flemish dwellings in Belgium. Based on the family type, a stochastic distribution of using different electrical appliances and the internal heat gains are randomly generated. The outdoor temperature and solar radiation profiles are extracted from a TRNSYS weather file for Uccle, Belgium [18].

The four dwellings of the apartment building are similar, but with different window orientations, and are modelled as a 3R2C-model. This model takes into account the ventilation losses (leakage and hygienic ventilation by a C-system), transmission losses (the dwellings' average U-value is 0.5 W/m²K), solar gains, internal heat gains and a capacity for the indoor air and the walls. The design heat loss of a single dwelling is 3.4 kW.

All apartments are equipped with hydronic radiators and thermostatic radiator valve (TRV) [19] where the supply temperature is equal to the distribution temperature of the heating network (controlled by the RL-agent). The temperature profile inside the hydronic radiator is modelled with three nodes as in previous research [14].

The central geothermal heat pump is sized to cover the total design heat load. The heat pump is a grey-box model of [14], based on the performance map of geothermal heat pumps of Viesmann for different source and sink temperatures. Equation (4) represents the dynamic behaviour of the heat pump at time t, based on Fig 2.

$$
\begin{aligned}
C_{HP}\frac{dT^{snk}_{out;t}}{dt} &= \dot{Q}_{eva;t} + \dot{Q}_{elec;t} \\
&\quad - UA_{con}\left(T^{snk}_{out;t} - T_{zone;t}\right) \\
&\quad + c_p\dot{m}_{snk;t}\left(T^{snk}_{in;t} - T^{snk}_{out;t}\right)
\end{aligned}
\qquad (4)
$$

Where C_{HP} (J/K) is the thermal capacity of the heat pump, $T^{snk}_{in;t}$ and $T^{snk}_{out;t}$ (°C) are respectively the ingoing and outgoing temperature at the condenser of the heat pump, $T_{zone;t}$ (°C) is the ambient temperature of the heat pump, which is fixed at 20°C, because it is placed in a central boiler room. $\dot{Q}_{eva;t}$ (W) is the extracted heat of the evaporator and $\dot{Q}_{elec;t}$ (W) is the used electricity, determined by the performance map based on the temperatures at both condenser and evaporator side. UA_{con} (W/K) determines the envelope losses and "$c_p\dot{m}_{snk;t}(T^{snk}_{in;t} - T^{snk}_{out;t})$" represents the heat transferred to the mass flow through the heat pump, where c_p is 4187 (J/kgK) and $\dot{m}_{snk;t}$ is the mass flow (kg/s).

The supply and return pipe are modelled according to the plug flow model [14], to simulate the time delay in the pipes as well as the heat losses.

IV. Experiments

To apply the RL based control strategy in this application domain, we used RLlib [20] which is an open-source library. As RLlib works with OpenAI Gym, we created an OpenAI Gym custom environment and integrated it with the heating network simulation environment such that the agent can take control of the supply temperature. An overview of the RL-based heating network simulation environment is depicted in Fig. 3.

The goal of the RL agent is to control the supply temperature while minimizing the energy usage of the heat pump and keeping the indoor temperature requirements for occupants. Therefore, a multi-objective reward function is defined as a weighted sum of these two criteria. The indoor temperature setpoints are assumed to be set by the end user,

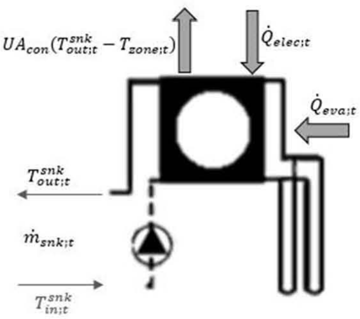

Fig. 2. Energy balance in the geothermal heat pump.

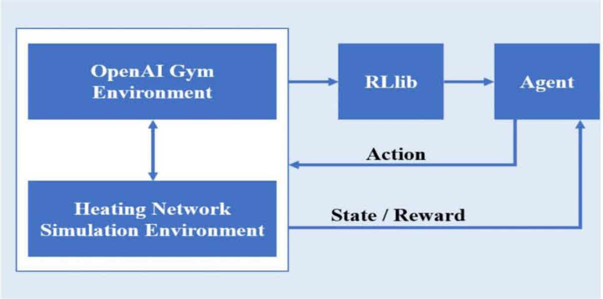

Fig. 3. Heating network simulation linked to RL-based control.

hence, the experiments contain different target temperatures for different dwellings at different times which makes the experiments closer to the real-world scenarios. The temperature deviation is considered as the Euclidean distance between the indoor temperature and the indoor temperature setpoint. The reward is calculated using the following equation:

$$reward_t = \left((1-\alpha)\cdot 1/\dot{Q}_{elec;t}\right) + (\alpha \cdot 1/\Delta T_t) \qquad (5)$$

Where $\dot{Q}_{elec;t}$ is the heat pump's electric power usage and ΔT_t is the temperature deviation at time t. The weight parameter is α with a range]0, 1[. With $\alpha = 0$ or $\alpha = 1$, the reward function turns into a single objective function, thus they will not be considered in this research. Both $\dot{Q}_{elec;t}$ and ΔT_t are normalized to the range $[0, 1]$. Considering the reward function, the state space variables are chosen in a way that reflect this behaviour in the observation space of the agent, thus we used only the most representative set of variables for state space which also helps in decreasing the computational complexity. Therefore, the outdoor temperature, indoor temperature of all dwellings, indoor temperature setpoint of all dwellings, and the supply temperature to the heating network which also effects the energy usage of the heat pump, are used as the set of state space variables.

The experiments were performed in the simulation environment for the winter season. The training of the agent is performed using 70 days of data simulation, and its performance is tested on the remaining 20 days of the season. During the test phase, the agent uses the knowledge obtained during the training and performs actions in the environment based on the learned policy. To investigate the effect of discrete and continuous action spaces on the learning procedure, two set of experiments were done based on them. In each set of experiment, different weights for the reward function are investigated to find a trade-off between energy usage of the heat pump and temperature deviation from the desired indoor temperature. The acceptable range for the temperature deviation is ± 0.5°C and the agent's reward is decreases outside this range.

In the discrete action space case, the supply temperature is controlled by three actions, including increasing/decreasing temperature by 0.5°C or keeping it unchanged, while continuous action space involves increasing/decreasing temperature within the range of [-0.5,+0.5]°C.

V. RESULTS AND DISCUSSION

To find a trade-off between the energy usage of the heat pump and temperature deviation in the dwellings, six experiments are done using three different weight parameters ($\alpha \in \{0.3, 0.5, 0.7\}$) for the reward function. As can be seen in (5), a larger α gives the priority to meeting the indoor temperature requirements of the occupants, and with a smaller α the priority is given to decreasing the energy usage of the central heat pump. The results are compared against a rule-based controller (RBC), which uses a heating curve to set the supply temperature (see Fig. 1 for the used heating curve).

In Fig. 4 the total energy usage of the heat pump is plotted against the average temperature deviation from the setpoints, in the test phase. The energy usage at time t is equal to the power usage multiplied by the timestep and converted to kWh. The average temperature deviation is the absolute mean temperature difference between the indoor temperature setpoint and the "measured" indoor temperature. As mentioned before, a temperature difference of 0.5°C is considered as acceptable. As expected, increasing the α, generally provides a smaller deviation from the occupant's temperature requirements while it causes an increase in the energy usage of the heat pump. As Fig. 4 shows, in all cases of agent-based RL control, the heat pump used less energy in

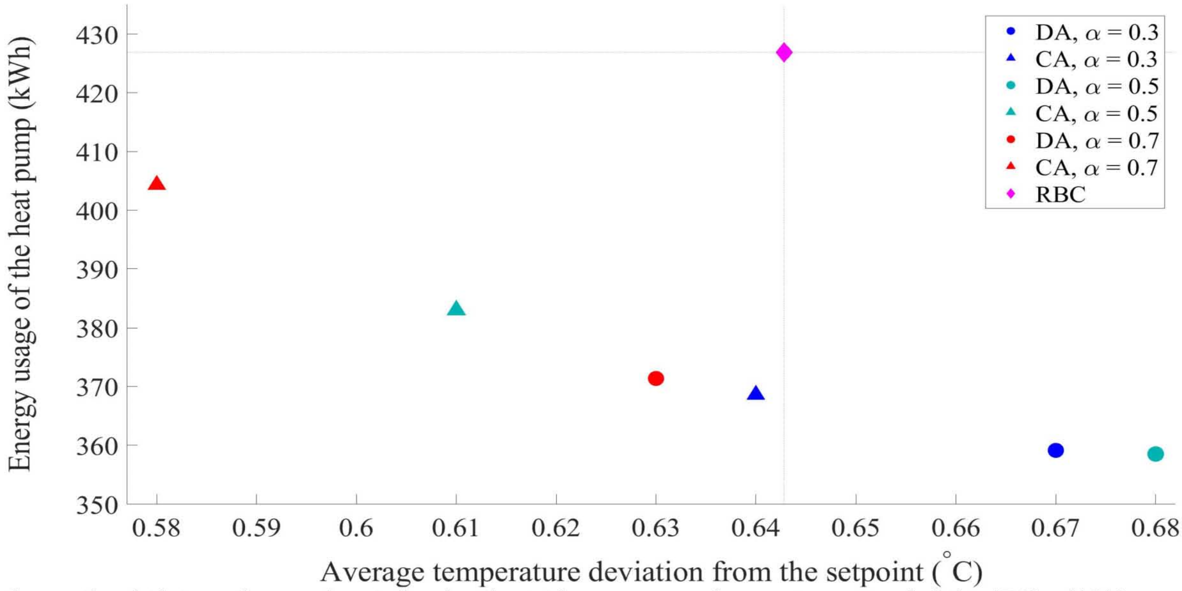

Fig. 4. Comparative plot between the research controllers, based on total energy usage and average temperature deviation. 'DA' and 'CA' represents the RL-agents with a discrete action space and a continuous action space, respectively. 'RBC' is the heating curve-based control.

comparison to the RBC. Besides, in 4 out of 6 experiments the RL-based controller performed better than RBC in terms of average temperature deviation from the indoor temperature requirements of the occupants. In all cases of continuous action space, the RL-based controller outperformed the RBC.

Table 1 compares the relative ratios of both average temperature deviations and energy usages of the heat pumps of the RL-agents to the RBC to assess their performances. When the RL-agent is trained with a continuous action space, the temperature deviations improves up to 9.78% as the weight for the reward function increases, compared to the RBC. The discrete action RL-agent outperforms the RBC only when the weight factor is 0.7 by 2%. Besides, the RL-based controls save 5.29% to 16.03% energy compared to the RBC.

The indoor temperature of a dwelling during the 20 days of testing is shown in Fig. 5 for the RL-based controls with continuous action spaces and the RBC. According to Fig. 4 and Fig. 5, RL-based control provides almost the same indoor temperature as the RBC, while the heat pump uses less energy.

Fig. 6 illustrates the supply temperature setpoint by the heating curve, discrete action agent and continuous action agent with both 0.7 as the weight α. It can be seen that the supply temperature is most of the time underneath the supply temperature of the heating curve. Sometimes, the supply temperature of the continuous RL-agent is higher than the heating curve's setpoint. The agent takes into account the indoor temperatures in addition to the outdoor temperature. A higher supply temperature setpoint by the agent compared to the heating curve, implies the effect of a low indoor temperature on the agent's decision. This is not taken into account in case of using a heating curve. By increasing the supply temperature, the agent minimizes the temperature deviation, to receive a higher reward. The RL-agent's supply temperature is fluctuating more throughout the day, which indicates that it is not solely based on the outdoor temperature (as it is the case with the heating curve), but it also takes into account other factors, such as indoor temperature, to control the supply temperature.

TABLE 1: RELATIVE TEMPERATURE DEVIATION AND RELATIVE ENERGY USAGE COMPARED TO THE HEATING CURVE (RBC).

Action space	**Weight factor of reward function**		
	0.3	***0.5***	***0.7***
Relative temperature deviation (%)			
Discrete	+4.23	+5.78	-2
Continuous	-0.44	-5.1	-9.78
Relative energy usage (%)			
Discrete	-15.88	-16.03	-13.01
Continuous	-13.66	-10.29	-5.29

The current study is based on a simplified version of a heating network. In order to make the simulation environment more representative of a real heating network, more dwellings should be considered. Besides the number of dwellings, also the domestic hot water demand should be taken into account. To minimize the gap between the simulation environment and a real-world case, and to have an RL-controller that could be applied to a real heating network, also the hydraulic behavior (i.e. the pressure drops, balancing valves, etc.) of a heating network should be taken into consideration. This in turn requires a more sophisticated state space and control space for the RL approach to provide optimal control decisions. Despite the fact that the proposed RL-based control approach achieves indoor temperatures close to the heating curve, but with less energy usage, an improved efficiency in this regard with providing more sophisticated control is anticipated. Above all, as the reward function has a considerable impact on RL agent control decisions, a reward function needs to be designed which is potentially able to derive the optimal policy for this complex decision making problem.

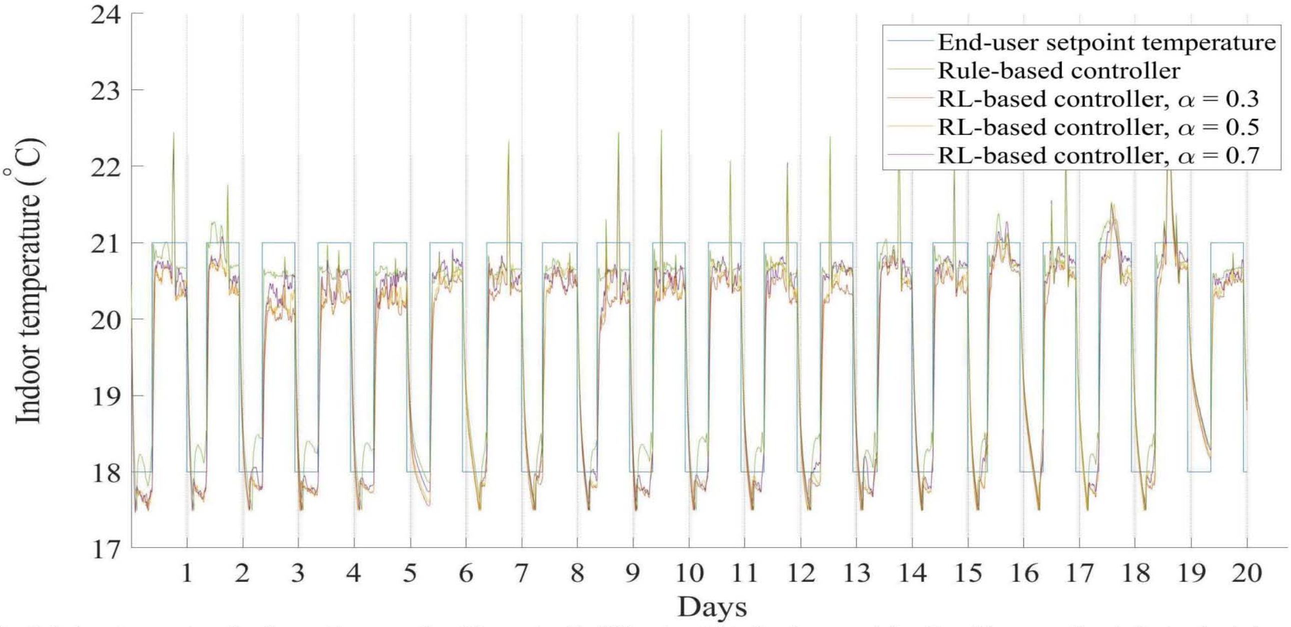

Fig. 5. Indoor temperature for the continuous action RL-agent with different weights for the reward function. The green line indicates the indoor temperature when using the heating curve for the supply temperature control.

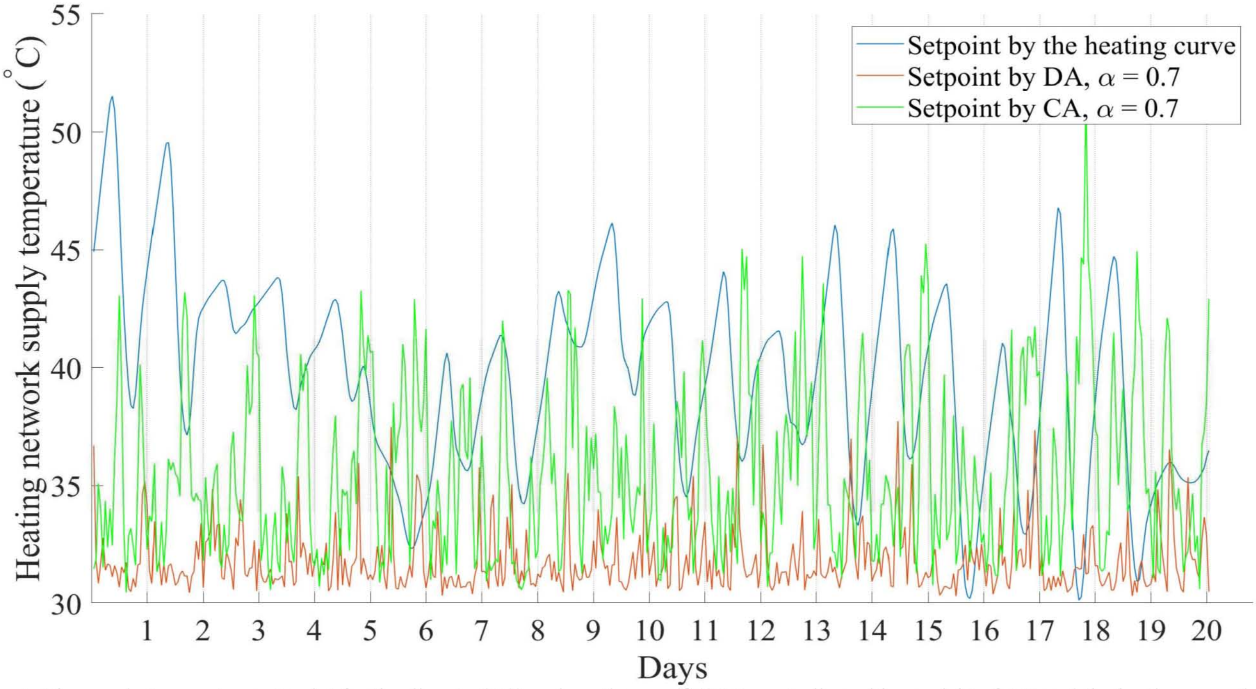

Fig. 6. The supply temperature set point for the discrete (DA) and continuous (CA) RL-controller, with a weight of 0.7, and the heating curve-based controller.

VI. Conclusion

The current supply temperature control for heating networks is rule-based, i.e. according to a heating curve. However, the innovative heating networks with more renewable energy sources and low heat demand dwellings, require advanced controlling systems.

In this work, we utilized model-free deep RL to control the supply temperature of a heating network with a central geothermal heat pump. The two objectives for the reward function were the electric energy usage of the geothermal heat pump and the indoor temperature deviation from the indoor setpoint temperature. A continuous action space, a discrete action space and three weight parameters of the reward function, including 0.3, 0.5 and 0.7, were examined.

A custom OpenAI Gym environment was developed and a MATLAB simulation environment, which simulates the thermodynamic behavior of the heating network, was integrated into it. This integration enabled us to use PPO method which is provided by the RLlib library and train an agent over a period of 70 days for controlling the supply temperature of the heating network.

The results over 20 days of testing suggest that utilizing a PPO agent with a multi-objective reward function over our simulation environment, can help in energy savings while maintaining the indoor temperature requirements for occupants. It also outperforms the heating curve control (i.e. RBC) as our baseline using a continuous action space. Regardless of the chosen weight for the reward function and of the type of action space, the results indicate that using RL-based control reduces the electric energy usage of the geothermal heat pump in comparison to the RBC. In this regard, the energy savings were up to 16.03% for the discrete action space and 13.66% for the continuous action space.

A next step is to increase the state space and involve more control actions which gives the agent more information and more degrees of freedom to optimize the policy. The heating network simulation environment could also be made more representative by taking into account the hydraulic behavior of a heating network as well as the domestic hot water demand.

Acknowledgment

This research was funded by the IOF-SBO project *Smart Thermal Grids* of the University of Antwerp, Belgium.

References

[1] Lauenburg. CH11 - Temperature optimization in district heating systems. In: Advanced District Heating and Cooling (DHC) Systems. Woodhead Publishing, 2016, pp. 223–40 (bookchapter).

[2] S. Buffa, M. Cozzini, M. D'Antoni, M. Baratieri and R. Fedrizzi, "5th generation district heating and cooling systems: A review of existing cases in Europe," Renewable and Sustainable Energy Reviews, Vol. 104, pp. 504–22, April 2019.

[3] Z. Ziang, A. Chong, Y. Pan, C. Zhang and K. P. Lam, "Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning," Energy and Buildings, Vol 199, pp. 472 – 490, September 2019.

[4] A. L. Coz, T. Nabil, and F. Courtot. "Towards Optimal District Heating Temperature Control in China with Deep Reinforcement Learning." arXiv preprint arXiv:2012.09508, unpublished.

[5] C. Johansson, D. Vanhoudt, J. Brage and D. Geysen, "Real-time grid optimisation through digitalisation – results of the STORM project," Energy Procedia, Vol 149, pp. 246-255, September 2018.

[6] T. Van Oevelen, D Vanhoudt, C. Johansson and E. Smulders, "Testing and performance evaluation of the STORM controller in two demonstration sites," Energy, Vol. 197, p. 117177, April 2020.

[7] J.R. Vázquez-Canteli, S. Dey, G. Henze and Z. Nagy. "CityLearn: Standardizing Research in Multi-agent Reinforcement Learning for Demand Response and Urban Energy Management," ArXiv:2012.10504v1, unpublished.

[8] R. S. Sutton, and A. G. Barto. "Reinforcement learning: An introduction 2nd ed." MIT press Cambridge 1, no. 2, 2018.

[9] Z. Wang, and T. Hong. "Reinforcement learning for building controls: The opportunities and challenges." Applied Energy, vol. 269, 2020.

[10] L. Lei, Y. Tan, K. Zheng, S. Liu, K. Zhang, and X. Shen. "Deep reinforcement learning for autonomous internet of things: Model, applications and challenges." IEEE Communications Surveys & Tutorials, vol. 22, no. 3, pp. 1722-1760, 2020.

[11] S. Padakandla. "A survey of reinforcement learning algorithms for dynamically varying environments." arXiv preprint arXiv:2005.10619, unpublished.

[12] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. "Proximal policy optimization algorithms." arXiv preprint arXiv:1707.06347, unpublished.

[13] Y. Zhan, P. Li, and S. Guo. "Experience-driven computational resource allocation of federated learning by deep reinforcement learning." In IEEE International Parallel and Distributed Processing Symposium (IPDPS), 2020, pp. 234-243.

[14] F. Van Riet., "Hydronic design of hybrid thermal production systems in buildings," Doctoral thesis for degree of Doctor in Applied Engineering, 2019.

[15] S. Jacobs, F. Van Riet and I. Verhaert, "A collective heat and cold distribution system with decentralised booster heat pumps: a sizing study," In Building Simulation 2021 Conference: 17th Conference of IBPSA. Bruges (Belgium), September 2021, in press.

[16] S. Jacobs, F. Van Riet, I. Verhaert, "Method to evaluate a heat and cold distribution in collective buildings with decentralised booster heat pumps," In Young Energy Researchers Conference (YERC) at the WSED, online, June 2021, in press.

[17] J. De Schutter, I. Verhaert and 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. Brussels (Belgium), 23 April 2018.

[18] Solar Energy Laboratory Univ. of Wisconsin-Madison (SELUWM), "TRNSYS 17 Volume 18 Weather data". 2017.

[19] D. P. Muniak, "Sizing the Radiator Control Valve Taking Account of Inner Authority," Procedia Engineering, Vol. 157, pp. 98-105, 2016.

[20] E. Liang, R Liaw, R Nishihara, P. Moritz, R. Fox, J. Gonzalez, K. Goldberg, and I. Stoica. "Ray rllib: A composable and scalable reinforcement learning library." arXiv preprint arXiv:1712.09381, unpublished.