

---

This is an electronic reprint of the original article.  
This reprint may differ from the original in pagination and typographic detail.

Deng, Jifei; Eklund, Miro; Sierla, Seppo; Savolainen, Jouni; Niemistö, Hannu; Karhela, Tommi; Vyatkin, Valeriy

## Application of reinforcement learning for energy consumption optimization of district heating system

*Published in:*  
2023 IEEE 32nd International Symposium on Industrial Electronics (ISIE)

*DOI:*  
[10.1109/ISIE51358.2023.10228102](https://doi.org/10.1109/ISIE51358.2023.10228102)

Published: 31/08/2023

*Document Version*  
Peer reviewed version

*Please cite the original version:*  
Deng, J., Eklund, M., Sierla, S., Savolainen, J., Niemistö, H., Karhela, T., & Vyatkin, V. (2023). Application of reinforcement learning for energy consumption optimization of district heating system. In *2023 IEEE 32nd International Symposium on Industrial Electronics (ISIE)* IEEE.  
<https://doi.org/10.1109/ISIE51358.2023.10228102>

---

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

# Application of reinforcement learning for energy consumption optimization of district heating system

Jifei Deng<sup>1</sup>, Miro Eklund<sup>2,3</sup>, Seppo Sierla<sup>1</sup>, Jouni Savolainen<sup>3</sup>, Hannu Niemistö<sup>3</sup>, Tommi Karhela<sup>1,3</sup>, Valeriy Vyatkin<sup>1,4</sup>

<sup>1</sup>*Department of Electrical Engineering and Automation, Aalto University, Espoo, Finland*  
{jifei.deng, seppo.sierla, tommi.karhela, valeriy.vyatkin}@aalto.fi

<sup>2</sup>*Department of Information Technology, Abo Akademi University, Turku, Finland*  
miro.eklund@abo.fi

<sup>3</sup>*Semantum Ltd, Espoo, Finland*

{miro.eklund, jouni.savolainen, hannu.niemisto, tommi.karhela}@semantum.fi

<sup>4</sup>*Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, Luleå, Sweden*  
vyatkin@ieec.org

**Abstract**—Heating residential spaces consumed 64 percent of total household energy consumption in Finland. Considering the heat transfer and time delay in the district heating system, the calculation of setpoints of supply temperature requires a comprehensive understanding of the real system, and experienced operators need to manually determine the setpoints. To save energy, a more effective and accurate method is needed for setpoints calculation. In this paper, a reinforcement learning based method is proposed. Through interacting with an Apros-based simulation model, the agents learn to calculate supply temperature parallelly for lowering energy costs. Simulation results show that the proposed method outperforms the existing method and has the potential to address the problem in real factories.

**Keywords**—*district heating, energy consumption optimization, reinforcement learning*

## I. INTRODUCTION

According to Statistics Finland 2020 [1], the energy consumed in households amounted to close to 61 terawatt hours (TWh) in 2020, and 64 percent (39 TWh of energy) of total household consumption was used to heat residential spaces. The dwellings, dominated by blocks of flats, are mainly heated with a district heating system (DHS) [1]. In [2], various scenarios of the heating system development for a period of 20 years (2015–2035) in southern Finland were carried out to study the energy efficiency of buildings. And district heating is a significant topic when studying energy saving of the national energy consumption.

Based on prefabricated pipelines, substations, and thermal power plants, water was chosen to be the heat carrier of the Finnish district heating system [3]. Considering the energy consumption and heat distribution losses, a wide variation of supply temperature levels is used [2]. The control system consists of four different and independent control systems [3]. The heat demand and flow control systems are located in each customer heating system and substation, while the heat supplier is responsible for the centralized differential pressure and supply temperature control systems [4].

In Finland, a heuristic-based method was adopted in factories. Operators manually set and adjust the supply temperature of the DHS. In other words, supply temperature settings require the full knowledge of the real DHS and working experience. For heuristic-based methods, since the

principles of the process need to be studied, and engineers make decisions based on experience, which means precision and performance cannot be guaranteed. In the research field, the optimization-based method is popular [5], [6]. Based on the specific problems of district heating, variables and fitness functions are designed for evolutionary algorithms. Through iteration, the fitness function is minimized, and the best solution is the one that can make the fitness value close to 0 or the desired value. For optimization-based methods, computation is expensive, and an accurate model of the process is required. Moreover, the operators need to design specific cost functions for optimization.

To address the above limitations, a novel solution that can adaptively make intelligent decisions without manual control but requires less computation resource is significant for the supply temperature optimization of DHS. Motivated by the learning ability of reinforcement learning (RL), which can learn from trial-and-error by maximizing the reward to push the learning policy towards a desired performance [7], this paper studied an RL-based method for supply temperature optimization framework.

Based on Apros platform [8], a simulation model of the studied DHS was built in [9]. In this paper, three state-of-the-art (SOTA) RL methods were adopted to calculate the setpoint. Considering the uncertainty of the process, three RL methods parallelly interact with the simulation model. The optimal policy was selected based on the total fuel cost for operating the DHS. The main intended contributions of this paper are as follows:

1. Given an Apros simulation model, which was built by fully analyzing the principles of the system, the reality gap between the environment and the real system of RL applications has been narrowed.
2. SOTA RL methods were combined to develop a framework to optimize the supply temperature by lowering the heating costs.
3. A reference plan from the factory was obtained to evaluate the proposed method, results showed the proposed method outperformed the existing method used at the case study factory.

The rest of this paper is organized as follows. The applications of existing methods and SOTA RL methods in DHS are reviewed in Section 2. A case study, including the

details of the simulation model, is presented in Section 3. Section 4 discusses the basic principles of RL and the proposed method. Section 5 analyzes the results of the proposed method. Section 6 concludes the paper and describes future work.

## II. RELATED WORK

To compute the supply temperature of the DHS, optimization-based methods and heuristic-based methods have been studied. Optimization-based methods use mathematical optimization algorithms to compute the setpoints that minimize a cost function, subject to system constraints and operational objectives. In [10], Multi-Objective Genetic Algorithm was used to drive reliable correlations for estimating the rate of heat losses from the twin- and triple-pipes. Su et al. proposed a Graph-Based Multi-Objective genetic algorithm approach for DHS to optimize pipe network layout and central plant positioning [11]. To address the problem in district energy systems operating at near-ambient temperatures, a particle swarm optimization approach was adopted to design the optimization framework for DHS [12]. Heuristic-based methods are mainly based on heuristics or rules of thumb. These methods are popular and adopted in factories because they are computationally efficient and easy to implement [13]. Using simplified mathematical equations, supply temperature can be approximated based on factors such as the ambient temperature, the heat loss from the buildings in the district, and the efficiency of the district heating system [14]. Heuristic-based methods are hard to provide optimal solutions and sensitive to the chosen heuristics. Optimization-based methods can provide global solutions to the optimization problem, but they may be computationally intensive and may require detailed models of the system.

To address the above problems, RL has been studied and adopted in various fields, including district heating [15] and energy management [16]. To address the peak-shaving problem in district heating, RL was combined with a thermodynamic model and agent-based model, and this

novel method achieved better results than the baseline [15]. Since wind power curtailment is inevitable in integrated electricity and DHS dispatch, a double-check RL was proposed to promote the integration of wind power by improving operational flexibility [17]. Qin et al. proposed a distributed RL-based control strategy for building energy optimization, compared with model productive control and evolutionary algorithm, the proposed method is the most energy-efficient [16]. To achieve distributed energy scheduling and strategy-making, a multi-agent RL approach was proposed, and an optimal equilibrium selection mechanism was applied to improve the performance of RL from benefit fairness, execution efficiency, and privacy protection [18]. Based on RL, a novel energy scheduling method was proposed to control and manage household energy [19]. In [20], the time delay caused by the heat transfer process in DHS was addressed by a memory-augmented RL method with a dueling network structure.

This paper studied the optimization of supply temperature setpoints using RL. The existing methods were developed based on the first principles and engineers' experience, which is inefficient. According to the above analysis of RL applications, RL which learns from the interaction between the agent and its environment to form a policy, can be a promising tool for the studied problem.

## III. PROBLEM FORMULATION

### A. Simulation model of DHS

The studied case is the DHS of Espoo in Finland. The power plant generates heat, which is transmitted to end users in an urban district through a network of hot water pipelines. Heat exchangers at the end user buildings extract the energy required for space heating and service water heating, and the cooled water returns to the power plant through separate return pipelines. The network can span the geographical area of the city, with multiple power plants and combined heat and power plants.

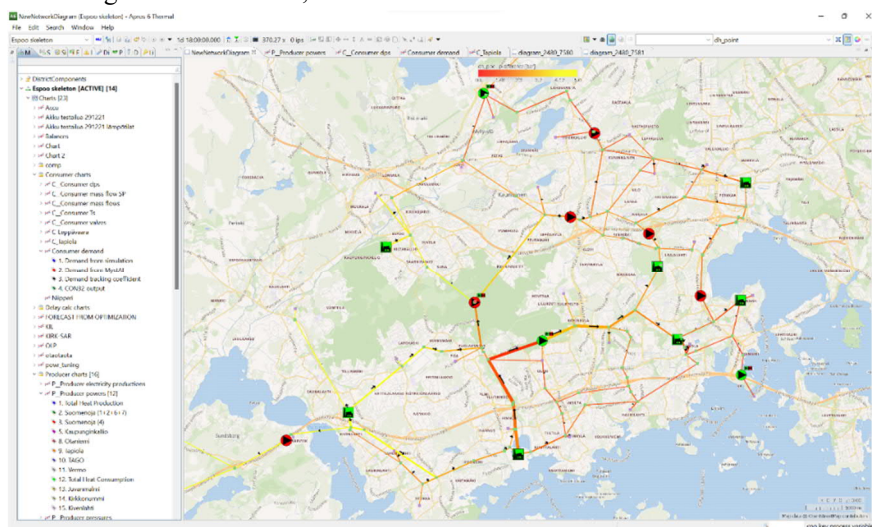


Fig. 1. The map district heating system in Espoo.

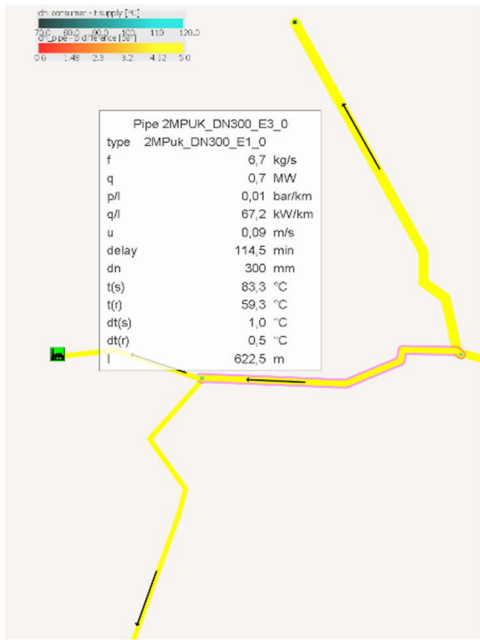


Fig. 2. An example of the network.

Apros [8] was adopted to build the simulation model of the district heating system. Figure 1 shows a snapshot from the Apros simulator. On the left is a tree view of the model configuration, charts, etc. The actual model can be viewed, for example, on a map user interface. In it, the studied district heating network is placed on top of a geographical map to show the real-world locations of heat producers, consumers, pumping stations, and pipes. In the network, several heat producers are used to heat pressurized water, which is then pumped to the customers.

Figure 2 shows a close-up of the network. There is a pipeline coming from the right (east, geographically), which

branches up and later down. These branches lead to customers. The remaining water flows to a producer (green symbol), which is here used to further heat the water, then returns to the network. Linewidths are used to visualize the pipe diameters whereas the lines' color indicates its pressure difference, a vital parameter for the network operation. The black arrows indicate the supply water flow direction and magnitude. Finally, by hovering the mouse cursor on top of any component, simulated values and other useful information are shown in pop-up boxes.

### B. Consumption optimization

The optimization is to adjust the temperature setpoint of the outgoing water at each plant (i.e., supply temperature) so that the fuel costs are minimized. In this paper, RL will not study heat storage tanks. Fuel price varies by the hour, and to allow timely business decisions. The optimization is not straightforward, due to the slow thermohydraulic phenomena and very large state space of the city-wide district heating network. As shown in Fig. 3, a 1-hour timestep is adopted, which means the fuel cost for operating the DHS for 1 hour is recorded as a reward. A simulator that captures these phenomena at a sufficient level of detail can be used to implement an environment for RL.

At each timestep, the Apros model provides the RL agent with fuel costs and measurement data, which will be used for computing supply temperature based on the policy. After the Apros model receives the supply temperature, new measurement data and fuel costs will be obtained. The policy will be updated using specific algorithms (e.g., PPO [21], SAC [22], TD3 [23]). Since the fuel costs are minimized during the training process, the supply temperature can be considered optimized for lowering the costs.

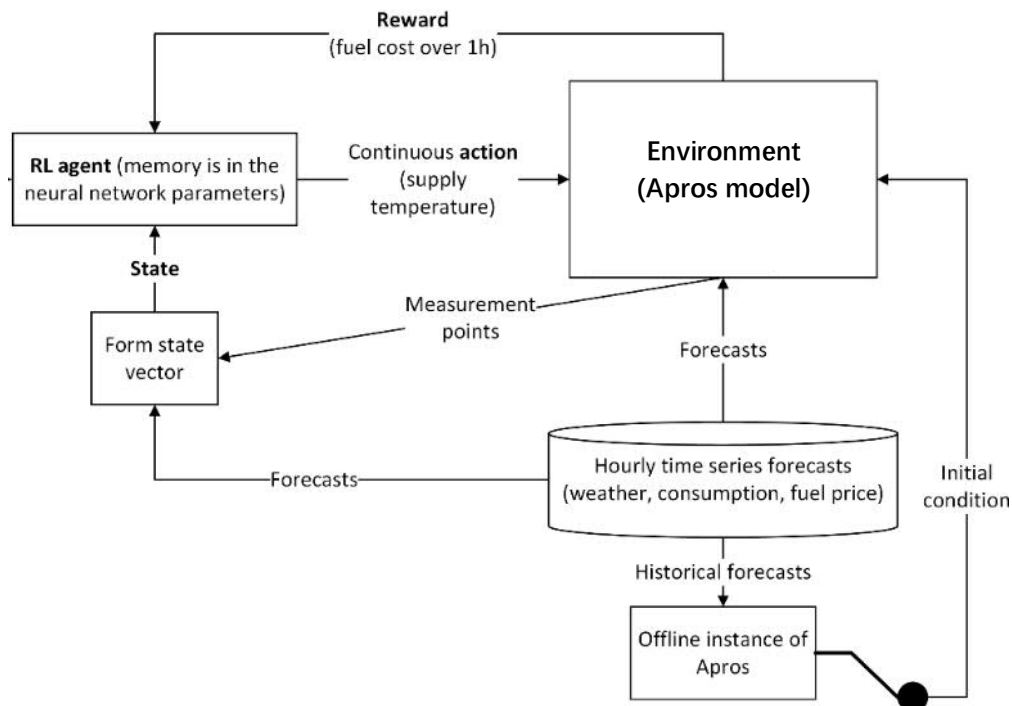


Fig. 3. Schematic of RL for supply temperature setpoints optimization.

## IV. REINFORCEMENT LEARNING

RL is a subfield of machine learning that trains agents to make decisions in complex environments rather than a static dataset. Based on the idea of trial-and-error, the RL agent learns to interact with its environment by maximizing a reward signal [7]. The RL process is composed of several key elements, including the environment and agent, states, actions, and rewards.

### A. Environment and agent

In RL, the environment is the dynamic system in which the agent operates. It provides the agent with observations and rewards for its actions. The agent, on the other hand, is the decision-maker that takes action for the environment. The environment can be represented by a Markov decision process (MDP), which is a mathematical framework for modeling sequential decision-making problems. An MDP is defined by states, actions, a transition function that describes the effect of actions on states, and a reward function that assigns a scalar reward to each state-action pair.

The agent's goal in RL is to learn a policy that maps states to actions in a way that maximizes the long-term reward. This is typically done using a value function, which assigns a scalar value to each state or state-action pair. The value function represents the expected long-term reward for the agent starting from that state or taking that action in that state.

In this paper, the environment was built based on the Apros model and using the framework of OpenAI Gym [24] and "HTTP" connection [9]. Since the RL agent was built and trained on Python, and the simulation model is on Apros, the "HTTP" connection is used to create a channel for RL and Apros model to exchange the data (state, action, reward) for each timestep.

### B. States, actions, and reward

In RL, the interactions between the agent and its environment are characterized by states, actions, and rewards. The state is the current situation of the environment which is determined by previous actions and the principles of the environment. The agent can take action in the environment, which can lead to a new situation. The reward is a scalar value that the agent receives after taking an action. The reward is determined by the reward function, indicating the quality of the action.

The states, actions, and reward together form the state space, action space, and reward space of the MDP. State space defines the set of possible states that the agent can encounter in the environment. The action space defines the set of possible actions that the agent can take in each state.

For DHS, the states are the temperature measurements at various stations in Espoo, the actions are the supply temperature for the heating plant. The reward function is defined as the inverse of the total fuel cost for operating the DHS ( $-1 \cdot \text{cost}$ ). During the training process, by maximizing the reward, the cost is minimized.

### C. Proposed method

This paper proposed a novel RL-based solution to lower the costs of the studied DHS. To address optimization

problems in large-scale industrial processes, based on three SOTA RL methods, an RL framework was developed to provide a reliable and efficient policy for the DHS. The pseudocode is shown in **Algorithm 1**.  $M$  and  $N$  are the number of episodes and the number of timesteps of each episode.  $IC$  is the initial condition.

In the RL framework, each method was trained separately to reduce uncertainty. For a real industrial process, the stability of the proposed method is as important as the performance. The proposed framework contains three SOTA RL methods, which can provide three available solutions. Depending on the actual situation, engineers could select the most suitable one.

---

#### Algorithm 1—RL framework for supply temperature optimization

---

```

Initialize RL agents (PPO, TD3, SAC) with the same setting for actor
and critic networks, import IC to Apros, create memory to store
information
For  $i$  in {PPO, TD3, SAC} do
  For  $k=1,2,3,\dots,M$  do
    Initialize Apros model using IC
    For  $j=1,2,3,\dots,N$  do
      Observe states ( $s$ ), and receive reward ( $r$ ) of Apros model
      Calculate actions ( $a$ ) for Apros model
      Update the actor and critic networks using ( $s,a,r$ ) pairs
    End for
    Record and send the information of the training process to memory
  End for
End for
Compare and choose the optimal solution from the results of the RL
agents
Save the optimal solution and compare it with the reference plan

```

---

Each RL method has the same IC for Apros model initialization. Since the adopted RL methods have a similar framework and the same hyperparameters, to guarantee a fair comparison, they will be fixed. The unique hyperparameters will be tuned using grid search.

## V. RESULTS AND DISCUSSION

In this paper, since the real costs cannot be disclosed due to confidentiality, the scaled cost is used as the indicator when comparing the RL methods with the reference plan. The formula is shown as follows:

$$C_{\text{scaled}} = C_{\text{actual}} / \mu \quad (1)$$

where  $C_{\text{scaled}}$  and  $C_{\text{actual}}$  are the scaled and actual costs,  $\mu$  is the scaling factor which is confidential and determined by the operators. The term 'reference plan' means human-decided setpoints for a period of one or more days, and it is the SOTA method of the studied case.

Table 1 Hyperparameters of RL methods.

Names	Values	Names	Values
	Shared		SAC and TD3
Total steps	2000	Start step	10000
Optimizer	Adam	Update after	1000
Actor learning rate	1e-3	Polyak	0.95
Critic learning rate	1e-3		TD3
Activation function	ReLU	Act noise	0.1
Update every	4	Target noise	0.2
Batch size	4	Noise clip	0.5
gamma	0.99	Policy delay	2
Hidden nodes of nets	128		SAC
	PPO	Entropy coefficient	0.2
Clip range ( $\epsilon$ )	0.2		
Lambda	0.98		

Each method was repeated 50 times with the same IC, and standard deviation (STD) and mean were used to draw the figures. Figures 4, 5, and 6 show the results of PPO, SAC, and TD3 (blue lines and areas), the lower and upper limits were computed by ‘mean-STD’ and ‘mean+STD’ (In statistics, the real values could go beyond the limits, see Fig. 7). The same reference plan (orange lines) was adopted for a fair comparison. The hyperparameters of the RL methods are shown in Table 1, including the shared parameters and unique ones.

As shown in Fig. 4, compared with PPO, lower costs were obtained by using the reference plan. The lowest cost of PPO is 0.72, while the highest cost of the reference plan is 0.74. On the contrary, in Fig. 5, SAC outperformed the reference at over 85% of the timesteps. Moreover, in Fig. 6, at all the timesteps, the costs of TD3 are lower than the reference. Figure 7 shows the overall results, the best solution of each RL method was chosen for comparison. For PPO and reference, neither one of them can completely outperform the other. Either SAC or TD3 has lower costs than reference and PPO, but SAC outperformed TD3 at 12 timesteps. SAC has the best performance.

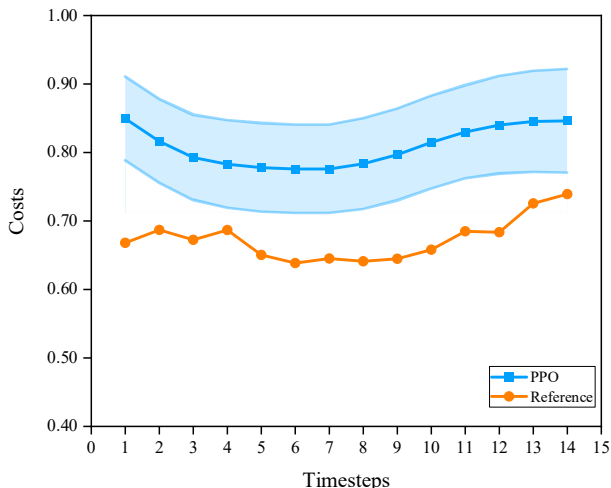


Fig. 4. Comparison of PPO and reference plan.

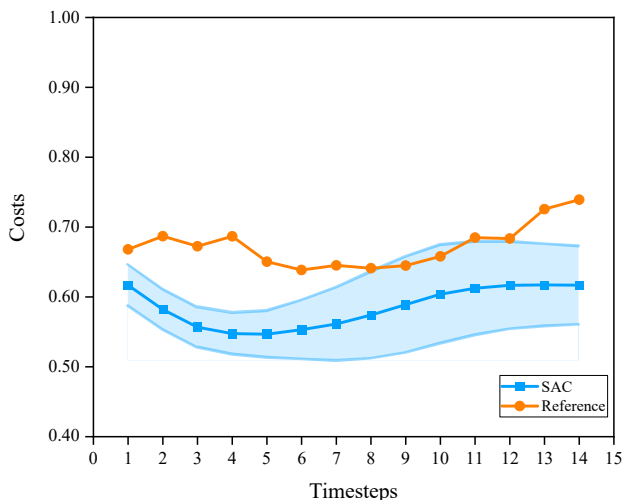


Fig. 5. Comparison of SAC and reference plan.

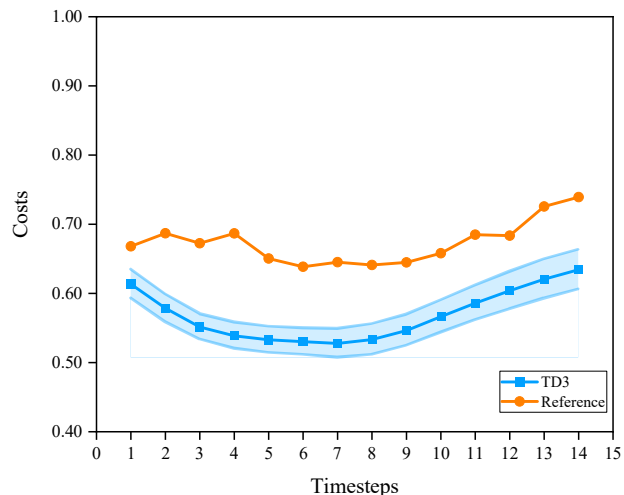


Fig. 6. Comparison of TD3 and reference plan.

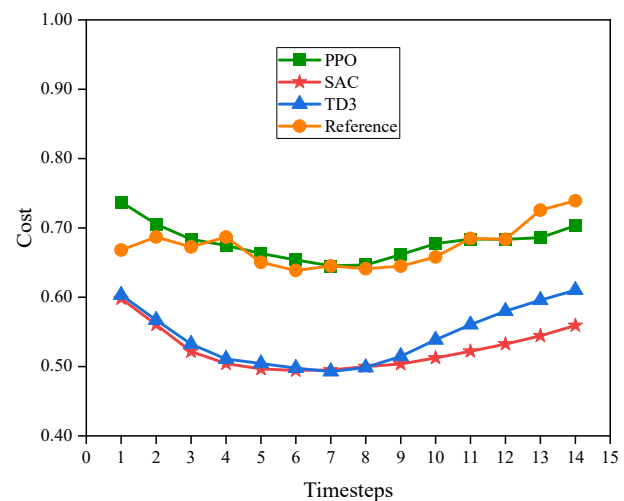


Fig. 7. The overall comparison of RL methods and reference plan.

Table 2 shows the mean and STD of the RL methods and reference plan. Compared with reference, PPO has a higher average cost, but the STD is lower. TD3 outperforms PPO in terms of mean, but it has the highest STD (0.041). Although the STD of SAC fails to be lower than that of the reference, it has the lowest cost (0.525). According to Fig. 4-7 and Table 2, the policy of SAC is the best solution for optimizing the supply temperature setpoints.

Table 2 Comparison of RL and reference in terms of mean and STD.

	PPO	SAC	TD3	Reference
Mean	0.679	0.525	0.543	0.673
STD	0.024	0.030	0.041	0.030

## VI. CONCLUSION

This paper proposed a novel intelligent method to optimize the setpoints of supply temperature in a district heating system. A simulation model which was built by fully analyzing the real process, was adopted to train the policy. To enhance the performance, a reinforcement learning based optimization framework, which consists of three state-of-the-art RL methods was proposed. Results showed that using the supply temperature calculated by the proposed method can reduce the fuel costs for the district heating system.

This paper is the first to study the reinforcement learning application for optimizing the supply temperature setpoints in a district heating system. This paper demonstrated that reinforcement learning is a promising tool to optimize the supply temperature setpoints. Next step, we will study RL to optimize both supply temperature and heat storage tank usage. Moreover, we will focus on developing new methods by considering more timesteps and improving computational efficiency.

#### ACKNOWLEDGMENT

This research was partly funded by Academy of Finland grant number 348415, and China Scholarship Council (202006080008). The authors would like to thank Timo Korvola from VTT Technical Research Centre of Finland, for expert advice on Apros and related code.

#### REFERENCE

- [1] Statistics Finland, "Statistics Finland - Energy consumption in households 2020." [https://stat.fi/til/asen/2020/asen\\_2020\\_2021-12-16\\_tie\\_001\\_en.html](https://stat.fi/til/asen/2020/asen_2020_2021-12-16_tie_001_en.html) (accessed Sep. 21, 2022).
- [2] R. Abdurafikov *et al.*, "An analysis of heating energy scenarios of a Finnish case district," *Sustainable Cities and Society*, vol. 32, pp. 56–66, Jul. 2017, doi: 10.1016/j.scs.2017.03.015.
- [3] S. Werner, "International review of district heating and cooling," *Energy*, vol. 137, pp. 617–631, Oct. 2017, doi: 10.1016/j.energy.2017.04.045.
- [4] S. Frederiksen and S. Werner, *District Heating and Cooling*. Professional Publishing Svc., 2013.
- [5] S. Bucking and V. Dermardiros, "Distributed evolutionary algorithm for co-optimization of building and district systems for early community energy masterplanning," *Applied Soft Computing*, vol. 63, pp. 14–22, Feb. 2018, doi: 10.1016/j.asoc.2017.10.044.
- [6] P. Żymelka and M. Szega, "Short-term scheduling of gas-fired CHP plant with thermal storage using optimization algorithm and forecasting models," *Energy Conversion and Management*, vol. 231, p. 113860, Mar. 2021, doi: 10.1016/j.enconman.2021.113860.
- [7] R. S. Sutton and A. G. Barto, *Reinforcement learning: an introduction*, Second edition. in Adaptive computation and machine learning series. Cambridge, Massachusetts: The MIT Press, 2018.
- [8] Apros, "Dynamic Process Simulation Software for Nuclear and Thermal Power Plant Applications." <https://www.apros.fi/> (accessed Oct. 28, 2022).
- [9] "Using a Digital Twin as the Objective Function for Evolutionary Algorithm Applications in Large Scale Industrial Processes | IEEE Journals & Magazine | IEEE Xplore." <https://ieeexplore.ieee.org/document/10064259> (accessed Apr. 22, 2023).
- [10] A. S. Alsagri, A. Arabkoohsar, M. Khosravi, and A. A. Alrobaian, "Efficient and cost-effective district heating system with decentralized heat storage units, and triple-pipes," *Energy*, vol. 188, p. 116035, Dec. 2019, doi: 10.1016/j.energy.2019.116035.
- [11] L. Su *et al.*, "Optimizing pipe network design and central plant positioning of district heating and cooling System: A Graph-Based Multi-Objective genetic algorithm approach," *Applied Energy*, vol. 325, p. 119844, Nov. 2022, doi: 10.1016/j.apenergy.2022.119844.
- [12] A. Allen, G. Henze, K. Baker, G. Pavlak, and M. Murphy, "An optimization framework for the network design of advanced district thermal energy systems," *Energy Conversion and Management*, vol. 266, p. 115839, Aug. 2022, doi: 10.1016/j.enconman.2022.115839.
- [13] I. Sarbu, M. Mirza, and E. Crasmareanu, "A review of modelling and optimisation techniques for district heating systems," *International Journal of Energy Research*, vol. 43, no. 13, pp. 6572–6598, 2019, doi: 10.1002/er.4600.
- [14] B. Talebi, P. A. Mirzaei, A. Bastani, and F. Haghighat, "A Review of District Heating Systems: Modeling and Optimization," *Frontiers in Built Environment*, vol. 2, 2016, Accessed: Dec. 15, 2022. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fbuil.2016.00022>
- [15] F. M. Solinas, L. Bottaccioli, E. Guelpa, V. Verda, and E. Patti, "Peak shaving in district heating exploiting reinforcement learning and agent-based modelling," *Engineering Applications of Artificial Intelligence*, vol. 102, p. 104235, Jun. 2021, doi: 10.1016/j.engappai.2021.104235.
- [16] Y. Qin, J. Ke, B. Wang, and G. F. Filaretov, "Energy optimization for regional buildings based on distributed reinforcement learning," *Sustainable Cities and Society*, vol. 78, p. 103625, Mar. 2022, doi: 10.1016/j.scs.2021.103625.
- [17] B. Zhang, A. M. Y. M. Ghias, and Z. Chen, "A double-deck deep reinforcement learning-based energy dispatch strategy for an integrated electricity and district heating system embedded with thermal inertial and operational flexibility," *Energy Reports*, vol. 8, pp. 15067–15080, Nov. 2022, doi: 10.1016/j.egy.2022.11.028.
- [18] X. Fang, Q. Zhao, J. Wang, Y. Han, and Y. Li, "Multi-agent Deep Reinforcement Learning for Distributed Energy Management and Strategy Optimization of Microgrid Market," *Sustainable Cities and Society*, vol. 74, p. 103163, Nov. 2021, doi: 10.1016/j.scs.2021.103163.
- [19] M. Ren, X. Liu, Z. Yang, J. Zhang, Y. Guo, and Y. Jia, "A novel forecasting based scheduling method for household energy management system based on deep reinforcement learning," *Sustainable Cities and Society*, vol. 76, p. 103207, Jan. 2022, doi: 10.1016/j.scs.2021.103207.
- [20] H. Zhao, B. Wang, H. Liu, H. Sun, Z. Pan, and Q. Guo, "Exploiting the Flexibility Inside Park-Level Commercial Buildings Considering Heat Transfer Time Delay: A Memory-Augmented Deep Reinforcement Learning Approach," *IEEE Transactions on Sustainable Energy*, vol. 13, no. 1, pp. 207–219, Jan. 2022, doi: 10.1109/TSTE.2021.3107439.
- [21] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv*, pp. 1–12, 2017.
- [22] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *ICML*, vol. 5, pp. 2976–2989, 2018.
- [23] S. Fujimoto, H. Van Hoof, and D. Meger, "Addressing Function Approximation Error in Actor-Critic Methods," *ICML*, vol. 4, pp. 2587–2601, 2018.
- [24] G. Brockman *et al.*, "OpenAI Gym." *arXiv*, Jun. 05, 2016. Accessed: Oct. 29, 2022. [Online]. Available: <http://arxiv.org/abs/1606.01540>