Deep reinforcement learning for fuel cost optimization in district heating

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

Published in:
Sustainable Cities and Society

DOI:
10.1016/j.scs.2023.104955

Published: 01/12/2023

Document Version
Final published version

Sustainable Cities and Society 99 (2023) 104955

Contents lists available at ScienceDirect

Sustainable Cities and Society

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

Deep reinforcement learning for fuel cost optimization in district heating

Jifei Deng [a,*], Miro Eklund [b,c], Seppo Sierla [a], Jouni Savolainen [c], Hannu Niemistö [c], Tommi Karhela [a,c], Valeriy Vyatkin [a,d]

[a] Department of Electrical Engineering and Automation, Aalto University, Espoo, Finland
[b] Department of Information Technology, Åbo Akademi University, Turku, Finland
[c] Semantum Ltd, Espoo, Finland
[d] Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, Luleå, Sweden

ARTICLE INFO

Keywords:
Deep reinforcement learning
Digital twin
District heating
Setpoint optimization

ABSTRACT

This study delves into the application of deep reinforcement learning (DRL) frameworks for optimizing setpoints in district heating systems, which experience hourly fluctuations in air temperature, customer demand, and fuel prices. The potential for energy conservation and cost reduction through setpoint optimization, involving adjustments to supply temperature and thermal energy storage utilization, is significant. However, the inherent nonlinear complexities of the system render conventional manual methods ineffective. To address these challenges, we introduce a novel learning framework with an expert knowledge module tailored for DRL techniques. The framework leverages system status information to facilitate learning. The training is performed by employing model-free DRL methods and a refined digital twin of the Espoo district heating system. The expert module, accounting for power plant capacities, ensures actionable directives aligned with operational feasibility. Empirical validation through comprehensive simulations demonstrates the efficacy of the proposed approach. Comparative analyses against manual methods and evolutionary techniques highlight the approach's superior ability to curtail fuel costs. This study advances the understanding of DRL in district heating optimization, offering a promising avenue for enhanced energy efficiency and cost savings.

1. Introduction

A district heating system (DHS) consists of prefabricated pipelines, substations, and thermal power plants to provide customers with heat (Werner, 2017). To control a heating network, 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 (Frederiksen & Werner, 2013). In a DHS, water is chosen to be the heat carrier, and considering the energy consumption and heat distribution losses, a wide variation of supply temperature levels is used (Abdurafikov et al., 2017). Fuel costs can be minimized by adjusting the setpoints of supply temperature and the usage of thermal energy storage (Eklund et al., 2023). Since air temperature, customer demands, and fuel prices vary by the hour, timely control decisions are significant for energy saving and cost reduction.

A heuristic-based method is commonly adopted in DHS control, which means operators manually set and adjust the setpoints of the DHS (Eklund et al., 2023). This method requires full knowledge of the working principles of the DHS and rich working experience. Due to the complexity of the DHS, including slow thermohydraulic phenomenon and heat propagation, human-made decisions cannot guarantee the efficiency of energy saving and cost reduction. To address this limitation, optimization-based methods were studied by formulating the calculation of setpoints as an optimization problem (Bucking & Dermardiros, 2018; Żymełka & Szega, 2021). Optimization-based methods find out the optimal solutions by searching the solution space using smart optimization algorithms, such as genetic algorithm (Su et al., 2022), evolutionary algorithm (Fazlollahi et al., 2012), and mixed integer linear programming (Gonzalez-Salazar et al., 2023). In Eklund et al. (2023), the fitness function was designed by considering fuel cost and setpoints, and covariance matrix adaptation evolution strategy (CMA-ES) was adopted to optimize the setpoints in a digital twin (DT) of a DHS. Through iterations, fitness values are minimized, and the best solution is the one that can make the fitness value approach the desired value. However, to implement these methods, huge computational

* Corresponding author.
E-mail address: jifei.deng@aalto.fi (J. Deng).

https://doi.org/10.1016/j.scs.2023.104955
Received 17 June 2023; Received in revised form 24 August 2023; Accepted 20 September 2023
Available online 21 September 2023

resources are essential, lowering computational efficiency.

To address the above limitations, an intelligent optimization method that can understand the process dynamics of DHS and adaptively make decisions is significant. Owing to the digitalization construction, process data is recorded, and first principle models and DT are available. By learning from data, data-driven methods can provide insights and patterns that might not be apparent through traditional methods. In (Lumbreras et al., 2022), a data-driven prediction model was proposed for the characterization and prediction of heating demand in buildings. In (Wang et al., 2021), a data-driven adjustable robust unit commitment model was proposed for integrated electric-heat systems. Motivated by deep reinforcement learning (DRL), which is a machine learning method but able to learn a control policy from the data, DRL-based data-driven methods were studied to address control problems in district heating. In (Ren et al., 2022), a novel forecasting based reinforcement learning energy scheduling method is proposed to manage household energy using a dueling-double deep Q-learning neural network.

Motivated by data-driven methods and DRL control, this paper studies a DRL-based framework for setpoints optimization in a DHS. The DHS of Espoo in Finland is chosen as a case study. Apros simulation environment (Silvennoinen et al., 1989) was used to build the DT based on the first principles, and its effectiveness has been demonstrated in (Eklund et al., 2023). To make a fair comparison, state-of-the-art DRL methods are incorporated into the proposed framework to optimize the setpoints.

The rest of this paper is organized as follows. The applications of existing methods and state-of-the-art DRL methods in DHS are reviewed in Section 2. A case study, including the details of the simulation model and problem formulation, is presented in Section 3. Section 4 discusses the basic principles of DRL and the proposed method. Section 5 analyzes the results of the proposed method. Section 6 concludes the paper and describes future work.

2. Related work

As presented in Section 1, to save energy for a DHS, manual operation and evolutionary method (CMA-ES) were adopted and researched for optimizing the setpoints (Eklund et al., 2023). In (Solinas et al., 2021), the building's thermal response was modeled as a multi-armed bandit, while the end-user networks were modeled as an agent-based model, and a conventional RL method was adopted to address the peak-shaving problem. However, the accuracy of the end-user network cannot be guaranteed, making the real application less practical. In (Sun et al., 2022), a deterministic forecasting model with energy-saving consideration was adopted to learn heat load variation patterns, and a DRL method was adopted to determine optimal fusion weights. However, using the black-box based forecasting model as the environment generates a reality gap between the real system and the model, which means the learned method is not realistic. Moreover, the problem of heat losses of the twin- and triple-pipes was studied in (Alsagri et al., 2019), which adopted a multi-objective genetic algorithm to drive reliable correlations for estimating the rate of heat losses. In (Wei et al., 2022), a model predictive control based optimization framework for the heat pump system of a residential district was proposed, and a particle swarm optimization algorithm was adopted to determine the optimal operation strategy. In (Alsagri et al., 2019) and (Wei et al., 2022), either genetic algorithm or particle swarm optimization algorithm can address optimization problems, but due to the population size, these methods are computationally expensive. These methods are less adaptive to the changing environment than data-driven methods. In (Stepanovic et al., 2022), a DRL-based method was proposed to control the pipeline energy storage of a DHS for profit gain, and the method was evaluated using a simulator. In (Zhao et al., 2022), to better cope with the uncertainty introduced by the high penetration of renewable generation units, a DRL method was integrated into the energy management problem by considering time delay, and evaluated by a memory-augmented environment. In (Qin et al., 2022), a distributed DRL-based control strategy was proposed for building energy optimization, compared with model productive control and evolutionary algorithm, the proposed method is the most energy-efficient. In (Fang et al., 2021), to achieve distributed energy scheduling and strategy-making, a multi-agent DRL approach was proposed, and an optimal equilibrium selection mechanism was applied to improve the performance of DRL from benefit fairness, execution efficiency, and privacy protection. In (Stepanovic et al., 2022; Zhao et al., 2022; Qin et al., 2022; Fang et al., 2021), single-agent and multi-agent DRL were adopted to address district heating related problems, black-box models and simplified mathematical models were developed to train the DRL agents. Compared to evolutionary methods, DRL methods learn the policy from the interactions between the agent and its environment, which can be more adaptive and efficient in computation. However, the environment introduces the bias because of the simplification, approximation, and black-box models. To address the above limitations, high-fidelity models are significant, especially for safety-critical cases.

In Espoo, the existing method (manual operation) for setpoint optimization is mainly based on expert operator experiences, which vary between individuals. Manual operation is easy to implement, but it is not efficient without any mathematical evaluation. For CMA-ES, through minimizing fitness values using the iteration method and parallel computation, the near-optimal could be obtained, but it is computationally expensive. DRL-based methods are popular and have been studied to address various problems in DHS, including energy storage (Stepanovic et al., 2022), energy management (Zhao et al., 2022), energy optimization (Qin et al., 2022), energy scheduling, and strategy-making (Fang et al., 2021). However, there is a lack of publications about the DRL optimization of setpoints in a DHS. Moreover, DRL applications introduced above were conducted based on simplified and approximated environment models. The bias between the real system and environment models affects the practicality of the DRL solutions.

To address the above limitations, this paper studies the optimization of supply temperature and thermal energy storage setpoints using DRL methods. Given a white-box-based DT of the DHS of Espoo, the obtained actions can be thought of as executable solutions to the real DHS. Compared to population-based evolutionary methods which have high computational demand and are difficult to learn from high-dimensional data (Majid et al., 2023), DRL is more sample-efficient and adaptive to the dynamic and changing system.

3. Problem formulation

3.1. Description of DHS

The study case is the DHS of Espoo in Finland. Power plants produce heat, which is distributed to urban district end users through a network of hot water pipelines. Within the end user buildings, heat exchangers extract the necessary energy for space heating and heating of service water. The cooled water then travels back to the power plant through dedicated return pipelines.

In the network, heat producers are used to heat pressurized water, which is then pumped to the customers. The temperature of the water leaving the producer (supply temperature) is given as a time-varying boundary condition to the producer component which forcibly sets this as the temperature of the relevant discretization nodes of the underlying equations. Mass flow and pressure out of the producer component are calculated with the pressure-flow solver.

The producer's pump and control valve components set suitable source terms for the solver to achieve the desired outlet pressure and thermal power. The thermal power comes from an external data source or is given by the user. The mass flow through an individual consumer is determined with a PID controller whose setpoint is calculated:

$$\dot{m}_{sp} = \frac{\dot{Q}_{demand}}{C_p(T_s - T_r)} \tag{1}$$

where $\dot{Q}_{demand}$ is the consumer's heat demand, kW. C_p is water heat capacity, kJ/kgK. T_s is the supply temperature, ℃. T_r is the return water temperature, ℃.

Accumulator's capacity refers to the percentage of the accumulator's effective volume that is filled with hot water. The effective volume takes into account the fact that the hot water phase typically is not allowed to fill the entire volume of that accumulator but rather is allowed to vary between a low and high limit which are configurable parameters for the model. The accumulator model gets a power setpoint value, P_{SP} as its input, with $P_{SP} > 0$ when discharging and $P_{SP} < 0$ when charging. The net power, P_{net}, is the same as P_{SP}, but with the following limitations taken into account: a fully charged accumulator cannot be charged further, a fully empty accumulator cannot be discharged further when pushing hot or cold water out of the accumulator into the network a too large network pressure will prevent this. In a heat accumulator, the bottom of the tank is filled with relatively cold water and the top with hot water. The interface of the two phases moves upwards or downwards when the accumulator is discharged or charged, respectively. The interface height is the vertical location, from the top of the accumulator, of this interface:

$$h_{IF} = h - \left[h_c(0) - \frac{\int \dot{m}_H dt}{r_c A}\right] \tag{2}$$

where h is the accumulator height, $h_c(0)$ is the height of the cold phase at the start of time, m. $\dot{m}_H$ is hot water flow into the accumulator, kg/s. r_c is cold phase density, kg/m^3. A is the accumulator's cross-sectional area, m^2. Three factors affect the hot and cold phase temperatures of the heat accumulator: incoming flow to the tank, heat mixed between the two phases, and heat loss to the environment. The energy that can be discharged from the accumulator is calculated as:

$$E = (T_H - T_C)C_p m_H \tag{3}$$

where T_H and T_C are the temperatures of hot and cold phases, ℃. C_p is the heat capacity of water, 4.19 kJ kg K. $\dot{m}_H$ is the hot phase mass, kg.

The pressure-flow solver calculates the pressures, flows, and temperatures in all parts of the model and it connects the producers to the consumers via pipelines and pumping stations. All the components described above manipulate the underlying pressure-flow solver's equations via boundary conditions or source terms. Temperatures are fed to the equations as boundary conditions whereas the effect of pumps and valves come in via source terms. The actual equations employed by the pressure-flow solver are partial differential equations for the conservation of mass, momentum, and energy. These equations are discretized w.r.t the spatial coordinate and then time-integrated as the simulation progresses (Patankar, 2018; Silvennoinen et al., 1989). Considering the ground temperature, pipe spacing, conductance, and heat loss from the pipe to the ground calculated according to formulas from Huovilainen and Koskelainen (1982).

The network in Fig. 1, which has over 800 km of piping, provides heating which is then used at the consumer sites to provide heating of the buildings as well as to produce domestic hot water. The idea of the network size and complexity can be found in this video: https://youtu.be/rNxlsSvoF70?t=57. There are a total of 9 pumping stations. Pipes are around 0.5 m under ground. Pipe diameters vary from DN1000 for the largest transfer pipes down to a few tens of millimeters for connections to individual buildings. The pipes are insulated and this is included in the simulation model. Insulation type can vary depending on each pipe type and age. A typical insulation material is polyurethane. Traditionally, in real life, the temperature of water leaving the producers follows the outdoor temperature according to the curve shown in Fig. 2. As the curve is conservative, the optimization is expected to strive for better performance.

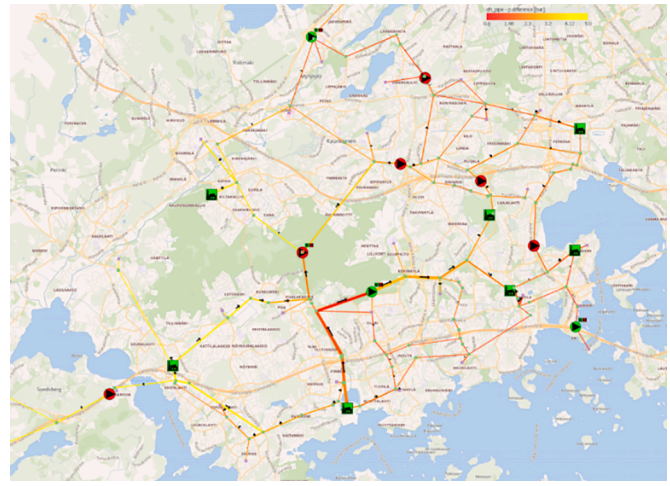

Fig. 1. The map of the district heating network in Espoo.

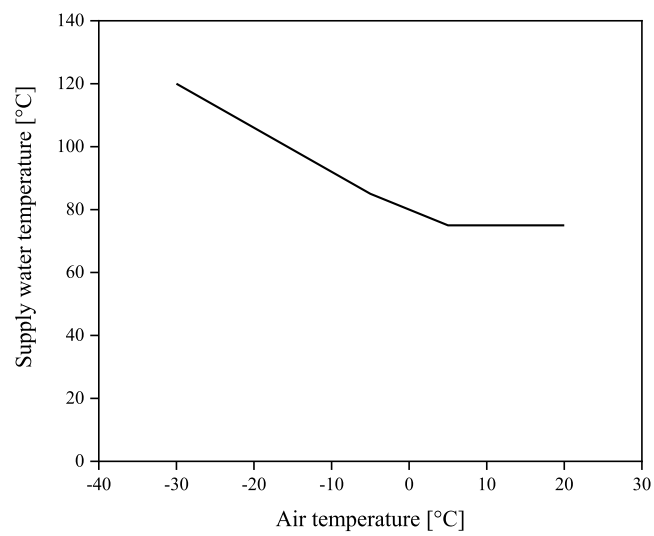

Fig. 2. The curve of air temperature and supply water temperature.

The network contains thousands of individual consumers. In the modeling, these are lumped according to their geographical locations into N ones. The network provided more than 800MW of heat during cold winter days, whereas during summer less than 100MW is provided. Both heating and hot water are supplied. The climate at the studied city can be classified as sub-arctic/humid continental climate according to the Köppen climate classification (Köppen, 2011). In the observation station, the air temperature is measured every hour, as shown in Fig. 3, 8761 data points were collected from the year 2022, and the air temperature varies between −20 °C and 30 °C.

3.2. Consumption optimization

In Table 1, 271 state variables are listed, such as the average demand, average consumption, and so on. Since the studied DHS covers the whole city, the values of different positions are considered separately. For example, each of the 29 positions has a separate average demand in the state. Setpoints include supply temperature (T_s) and the usage of thermal energy storage (Q_accum). As shown in Fig. 4, the optimization is to adjust the setpoints of the outgoing water and the usage of thermal energy storage for the network which is modeled as an Apros-based DT. Fuel cost, air temperature, and heat demand are provided as input to the

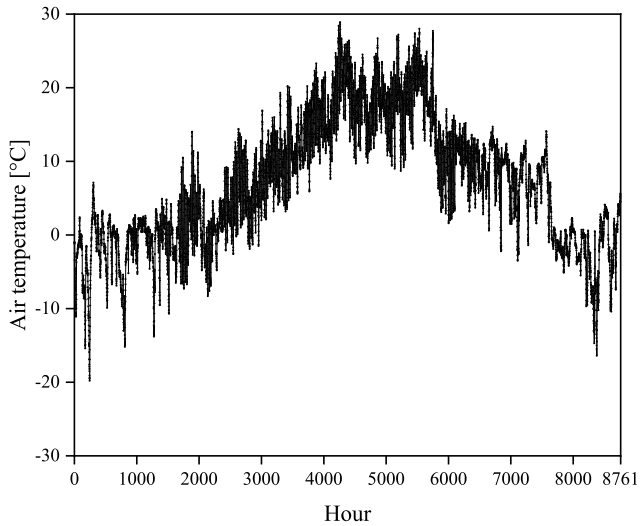

Fig. 3. Air temperature of Espoo in 2022.

DT. As shown in Fig. 5, historical data spanning 2 days (from 12:00 on Nov 2, 2022 to 12:00 on Nov 4, 2022) drives the DT for scientific exploration. In real-world applications, real-time optimization integrates forecasting models to supply inputs. Given climate-driven fluctuations in fuel cost, air temperature, and heat demand, an adaptable strategy is paramount. As depicted in Fig. 4, for example, if the DT receives forecasts for subsequent days (Tuesday/Wednesday) on the prior day (Monday) and performs optimization. Considering the fuel price, air temperature, and demand vary by the hour, an hourly setpoint optimization is adopted to minimize the fuel costs, yielding 48 pairs of setpoints for 2 days. When forecasts for Wednesday/Thursday arrive on Tuesday, further optimization can be performed again. This adaptive approach mitigates uncertainties arising from fuel cost, air temperature, and heat demand effectively, guaranteeing its reliability. This optimization problem can be formalized as:

$$\min C(T_s, Q\_accum),$$

$$\text{subject to}: T_s \in [T_s\_L, T_s\_U], |T_s\_C| \leq T_s\_limit \tag{4}$$

$$Q\_accum \in [Q\_accum\_L, Q\_accum\_U]$$

where $C(T_s, Q\_accum) = F(T_s, Q\_accum, \dot{Q}_{demand}, T_{air}, FP)$, C is the total cost, F is the DT which considers all the factors presented in Section 3.1, T_{air} and FP are the air temperature and fuel price which vary hourly. T_s_L and T_s_U are the lower and upper limits of T_s. T_s_C and T_s_limit are the actual change and the allowed maximum change between the steps. Q_accum_L and Q_accum_U are the lower limit and upper limit of Q_accum.

RL optimizes T_s and Q_accum with restriction to the limits. In this case, the modeled automation system in DT guarantees the fulfillment of consumer demands. Essentially, there exists no predetermined lower-cost limit. Because consumer requirements are assured by the DT, the optimization methods concentrate on minimizing fuel expenses to the greatest extent possible. At present, a manual operation method is adopted in the power plants in Espoo, which relies on the engineers' experience. However, as introduced in the above sections, the network is complicated, which means the DT of the DHS has characteristics of nonlinearity and high dimensionality. The manual operation method makes it difficult to lower the cost by addressing this optimization problem. Therefore, evolutionary methods such as CMA-ES have been studied, and compared to the manual operation method the superiority was demonstrated in (Eklund et al., 2023). In this paper, DRL methods are studied for optimization. The states of the DT and fuel price are sent to the DRL agent to calculate the setpoints (also called actions) for DT.

Table 1
Details of states.

Name	Number	Category	Calculation details
avg_demand	29	Consumer	1-hour time-averaged heat demand for each consumer
avg_consumption	29	Consumer	1-hour time-averaged heat consumption for each consumer
avg_supply_temperature	29	Consumer	1-hour time-averaged supply temperature for each consumer
avg_pressure_difference	29	Consumer	1-hour time-averaged pressure difference over each consumer's control valve
avg_heat_offset	29	Consumer	Consumer's heat demand – consumption, 1-hour average
max_valve_position	29	Consumer	Consumer's control valve's position's max value over 1-hour window
heating	24	Producer, boiler-wise	Produced power for each boiler. Note that some geographical sites have multiple boilers producing heat. That is the reason this is of length 24, and not 10 (as in *supply temperature*).
Pump_speed	10	Pumping station	Supply line pump speed for each pumping station, point measurement at the end of each 1-hour period
pump_mass_flow	10	Pumping station	Supply line pump mass flow for each pumping station, point measurement at the end of each 1-hour period
head_pump_actual	10	Pumping station	Maximum of supply line and return line pump heads for each pumping station, point measurement at the end of each 1-hour period
pressure_difference_discharge	10	Pumping station	Difference of pumping station supply pump's outlet and return pump's suction pressures, point measurement at the end of each 1-hour period
pressure_difference_charge	10	Pumping station	Difference of pumping station supply pump's suction and return pump's outlet pressures, point measurement at the end of each 1-hour period.

(*continued on next page*)

Table 1 (*continued*)

Name	Number	Category	Calculation details
Supply_temperature	10	Producer, site-wise	Producer site supply temperature, point measurement at the end of each 1-hour period
network_accum_MWh	1	Network	Energy in supply lines + energy in return lines, summed over all pipes, point measurement at the end of each 1-hour period
accum_cold_phase_temperature	1	Heat accumulator	Temperature of the accumulator's cold phase, point measurement at the end of each 1-hour period
accum_energy	1	Heat accumulator	Energy that can be discharged from the hot phase of the accumulator, point measurement at the end of each 1-hour period
accum_capacity	1	Heat accumulator	Percentage of accumulator's effective volume that is filled with hot water, point measurement at the end of each 1-hour period
accum_hot_phase_temperature	1	Heat accumulator	Temperature of the accumulator's hot phase, point measurement at the end of each 1-hour period
accum_interface_height	1	Heat accumulator	Vertical location of the hot-cold phase interface from the top of the accumulator, point measurement at the end of each 1-hour period
accum_net_power_me	1	Heat accumulator	P_{net}, 1-hour time-averaged
tla_total_mw	1		Total power produced by the producers (excluding the accumulator), point measurement at the end of each 1-hour period
tla_forecast	1		Forecasted production, if forecast available. In this case forecast = sum of all boiler productions (excluding the accumulator), point measurement at the end of each 1-hour period.
Tla_main_dp_deviation	1		Control deviation (setpoint – measurement) of the main DP controller, point measurement at the end of each 1-hour period
tla_main_dp_calc	1		Largest of the DP deviations of the critical consumers, point measurement at the end of each 1-hour period
tla_forecast_correction	1		Output of the main DP controller. This is used to multiply the forecasted production in order to calculate how much heat will be produced, point measurement at the end of each 1-hour period
total_pipeline_heat_loss	1		Sum of all pipes' heat losses, point measurement at the end of each 1-hour period

Through interaction, state-action-reward pairs are collected to train the DRL agent. Repeating this process, DRL explores the state space to calculate the optimal actions that generate lower fuel costs.

4. Methodology

4.1. Preliminaries

DRL is a subfield of machine learning, which is based on the idea of learning from trial-and-error. The goal of a DRL agent is to maximize the total reward it receives over a trajectory of interaction with the environment (Sutton & Barto, 2018). The interaction is modeled as a Markov Decision Process (MDP), which is defined by a tuple (S,A,R,P,γ), where S denotes state space, A is action space, R is reward function, P is transition dynamics, and γ is a discount rate. Given values of the preceding state s and action a at the current timestep, the probability of state occurring at the next timestep is $P(s^{'}|s,a)$. The accumulation of the rewards $R(s,a)$ starting from time t until the end of the interaction is referred to as the return, which is defined as $G_t = \sum_{k=0}^{\infty}(\gamma^k R_{t+k+1}s,a)$. The discount rate determines the present value of future rewards: a reward received k time steps in the future is worth only γ^{k-1} times what it would be worth if it were received immediately. For policy π, the state-action value function is defined as: $Q^{\pi}(s,\ a) := \mathrm{E}_{\pi}[\sum_{t=0}^{\infty}\gamma^t R(s_t, a_t)|s_0 = s,\ a_0 = a]$, and the corresponding value function is $V^{\pi}(s) := \mathrm{E}_{a\sim\pi(\cdot|s)}[Q^{\pi}(s,\ a)]$. In DRL, an approximation to the Q-function is conducted based on the actor-critic paradigms. Q-function can be learned via temporal difference learning based on the Bellman equation $B^{\pi}Q^{\pi}(s,a)$, where B^{π} denotes the Bellman evaluation operator $B^{\pi}Q^{\pi}(s,a) := R(s,a) + \gamma E_{s^{'},a^{'}}[Q(s^{'},a^{'})]$, where $s^{'} \sim P(|s,a), a^{'} \sim \pi(|s)$.

4.2. DRL framework for setpoint optimization in DHS

As analyzed in Section 2, given an environment, model-free DRL methods, such as proximal policy optimization (PPO) (Schulman et al., 2017), twin delayed deep deterministic policy gradient (TD3) (Fujimoto et al., 2018), soft actor-critic (SAC) (Haarnoja et al., 2018) are promising tools to address real-world problems (Stepanovic et al., 2022; Zhao et al., 2022; Qin et al., 2022; Fang et al., 2021). However, these applications were conducted in simple and black-box environments rather than white-box and high-fidelity models. In this paper, given a DT of the DHS, the gap between the model and the real system has been narrowed, ensuring model-free DRL methods are promising for setpoint

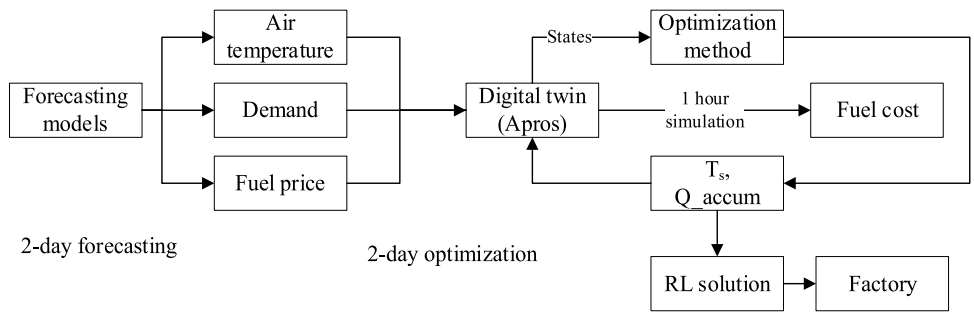

Fig. 4. Setpoint optimization for DHS.

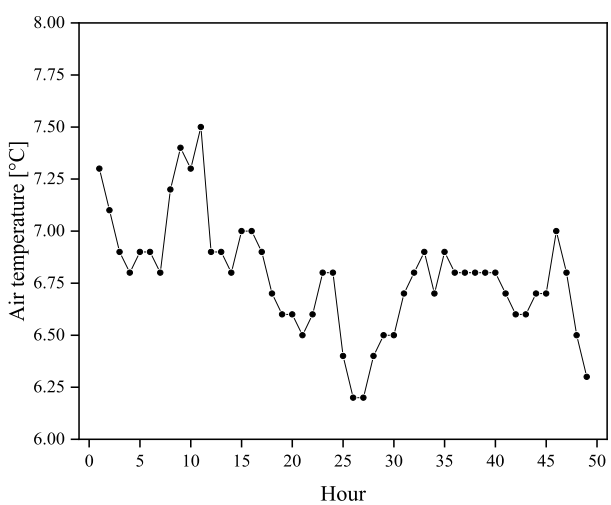

Fig. 5. Historical data from 12:00 on Nov 2, 2022 to 12:00 on Nov 4.

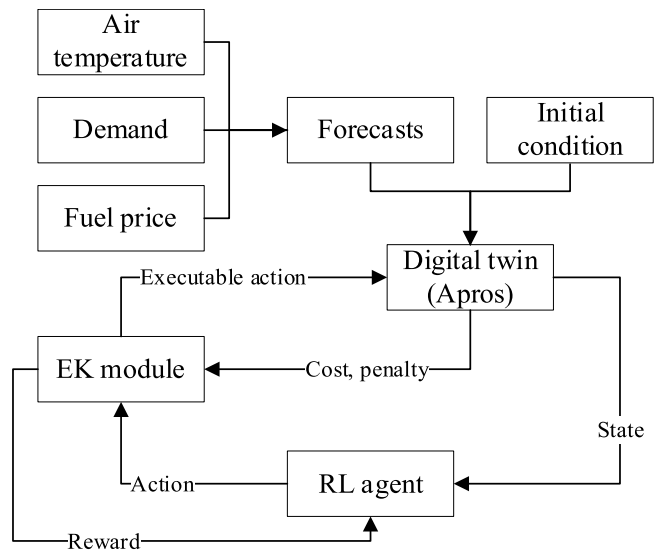

Fig. 6. Setpoint optimization using DRL with an EK module.

optimization.

However, for industrial applications of DRL, although the action space is preset in the training process, the complexity of the real-world case could cause the generated actions to be inexecutable. To address this limitation, various settings have been studied. In (Ruelens et al., 2017) and (Ruelens et al., 2019), DRL methods were combined with an over-rule mechanism for the thermostatically controlled load to guarantee comfort and safety constraints. In (Patyn et al., 2018), to ensure adherence to constraints, a backup controller was designed for DRL to address the problem of residential demand response. Similarly, in (De Somer et al., 2017) and (Leurs et al., 2016), specific units were designed to avoid the failure of actions.

In this paper, to obtain an executable solution for the DHS, as shown in Fig. 6, an expert knowledge (EK) module is designed for DRL training to ensure the practicality of the learning procedure and actions. In the standard learning process, DRL can freely explore the action space, which means DRL will send any actions within the boundary to the environment. However, this is problematic in DHS, because, for example, although the T_s is within the boundary, it could also be inexecutable for its huge changes compared to the previous value. Moreover, considering the storage of the thermal energy tank, the penalty of *Q_accum* is sent to the EK module from DT. Since the tank cannot unlimitedly provide energy, the penalty of *Q_accum* is an indicator of thermal energy storage in the tank.

Fig. 7 shows the implementation of the abovementioned EK module. T_s and T_s_p are the supply temperature at the current timestep and the previous timestep. T_s_limit is the allowed maximum change between the timesteps. If the actual change is higher than the limit, the current T_s is inexecutable, because the huge changes in T_s could damage the power plants. If the current T_s is lower than T_s_p, the new T_s is obtained using T_s_p-T_s_limit. On the contrary, if the current T_s is higher, the new T_s is calculated using T_s_p+T_s_limit. *Q_accum_penalty* is the penalty of *Q_accum*, *Q_accum_penalty_limit* is the upper limit of *Q_accum_penalty*. In practice, this penalty should be lower than its limit, otherwise, the training of the current episode will be ended. For the DHS, a high *Q_accum_penalty* means the DRL agent discharges too much heat but there is a lack of thermal energy in the storage tank. These actions provided by DRL are inexecutable. In this paper, the reward for DRL agents is designed as the inverse of fuel cost, which is formulated as $R = -C$.

In other words, the EK module guarantees an executable action for the DT. In the standard DRL learning process, without the EK module, the DRL agent will provide actions within the boundaries, but unrealistic and inexecutable for the DHS. Because it is impossible to make a DRL agent study these practical factors by itself without any additional settings.

4.3. Experiment configuration

In this paper, using the DT of DHS, three DRL methods (PPO, TD3, SAC) are incorporated into the DRL learning framework. The goal is to learn a policy that maps states to actions in a way that maximizes the long-term costs. Since the DT was built using Apros, "HTTP" connection technology was adopted to create a channel between Python and Apros for data exchange (Eklund et al., 2023). OpenAI Gym framework was adopted to build an environment in Python, including step, reward, and reset functions (Brockman et al., 2016). Since the adopted DRL methods have a similar framework, to guarantee a fair comparison, the same

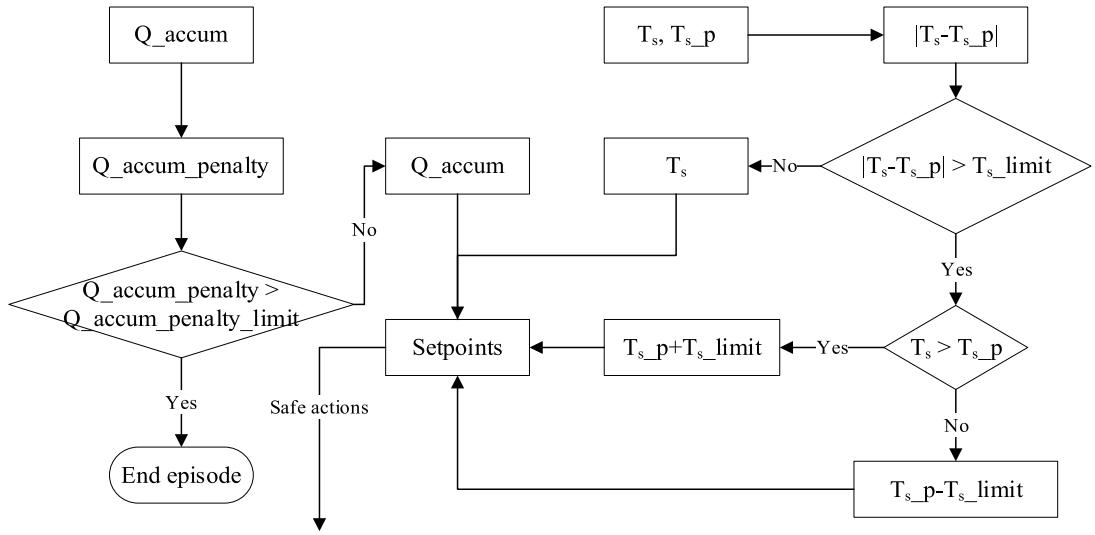

Fig. 7. EK module for DRL training.

hyperparameters will be fixed.

For DRL, each method was run for 30 episodes, each episode had 70 timesteps. To perform a fair comparison, each episode was run using the same initial condition (IC) and the same data source (air temperature, hourly heat demand, fuel price). In the simulation, the DRL agent employs a set of 16-hour setpoints generated through DRL. Following this, a subsequent 12-hour period is allocated to account for thermohydraulic delays inherent in the system, during which the agent operates without optimizing its actions. The 16+12 hour simulation run captures the fuel consumption data for all heat producers within the network, recorded on an hourly basis. By analyzing the fuel consumption throughout the 28-hour period, the total operating cost for this duration is derived. This paper studied 48 data points, and each data point means one hour between 12.00 2nd Nov. 2022 and 12.00 4th Nov. A CMA-ES method and a reference method (manual operation method) are adopted for comparison. The details of experiment settings and results of CMA-ES are presented in (Eklund et al., 2023).

5. Results

5.1. Comparison analysis

In this paper, since the actual costs cannot be disclosed due to confidentiality, the scaled cost is used as the indicator when comparing the DRL methods with CMA-ES and reference plan. The formula is shown as follows:

$$C_{\text{scaled}} = C_{\text{actual}} / \mu \tag{5}$$

where C_{scaled} and C_{actual} are the scaled and actual costs, μ is the scaling factor which is confidential and determined by the operators. The hyperparameters of the DRL methods are shown in Table 2, including the shared parameters and unique ones.

Table 2
Hyperparameters of DRL methods.

Hyperparameters	Values	Hyperparameters	Values
Total steps	2100	PPO—Clip range (ε)	0.2
Optimizer	Adam	PPO-Lambda	0.98
Actor learning rate	1e-3	(SAC, TD3)-Start step	4
Critic learning rate	1e-3	(SAC, TD3)-Update after	4
Activation function	ReLU	(SAC, TD3)-Polyak	0.95
Update every	4	TD3-Act noise	0.1
Batch size	4	TD3-Target noise	0.2
gamma	0.99	TD3-Noise clip	0.5
Hidden nodes of nets	256	TD3-Policy delay	2
		SAC-Entropy coefficient	0.2

Fig. 8 shows the solutions generated by PPO, TD3, and SAC in terms of costs. As illustrated in the last paragraph of Section 4, each line in Fig. 8 represents a solution. For each solution, the same initial condition, hourly heat demand, and fuel price were guaranteed to obtain fair comparison results. Considering the *Q_accum_penalty*, the number of potential solutions obtained by PPO is 10, while the TD3 obtains 4 potential solutions. Similarly, the SAC yields 7 potential solutions. Here, the term "potential solutions" refers to setpoints that are executable and can be deployed to the real DHS. Due to the trade-off between heating the water and charging/discharging the heat storage tank, all the solutions depicted in Fig. 8 exhibit a similar changing trend. However, it is important to note that the fuel price and demand vary on an hourly basis, leading to different cost outcomes for each solution.

Fig. 9 illustrates the optimal solution obtained by each DRL method in comparison to the reference plan and CMA-ES. The term "optimal solutions" refers to setpoints selected from potential solutions of each DRL method that result in the lowest cost for the DHS. Both the reference plan and CMA-ES exhibit a similar changing trend. Specifically, there is an increase in cost from the 1st to the 17th hour, followed by a decrease from the 18th to the 26th hour. After that, the cost begins to increase again. In contrast to the reference and CMA-ES methods, the DRL methods exhibit a distinct strategy. During the first 17 h, the DRL methods experienced an increase in cost; however, this increase is lower compared to the reference and CMA-ES methods. While the cost of the reference and CMA-ES methods begins to decrease at the 18th hour, the DRL methods maintain a stable cost without significant fluctuations. This difference in cost behavior highlights the unique strategy employed by DRL methods in optimizing setpoints for the DHS. Finally, the costs increase again.

Table 3 provides a summary of the scaled costs for all the methods, including the sum, mean, and standard deviation (STD). The sum represents the total cost accumulated over the 49-hour period, and the objective of our work is to minimize this total cost. The lowest total cost of 35.26 is achieved by the SAC, indicating its superior performance in cost reduction. In contrast, the reference has the highest total cost of 37.29. The CMA-ES yields a total cost of 36.27, which is lower than that of the reference and PPO methods but higher than the TD3 and SAC. Furthermore, the lowest mean value of 0.72 is achieved by the SAC, while the reference and CMA-ES methods have mean values of 0.76 and 0.74, respectively. The STD reflects the smoothness of the cost profile. SAC and TD3 exhibit STD values of 0.08 and 0.07, respectively, indicating smoother changes in cost compared to the reference and CMA-ES

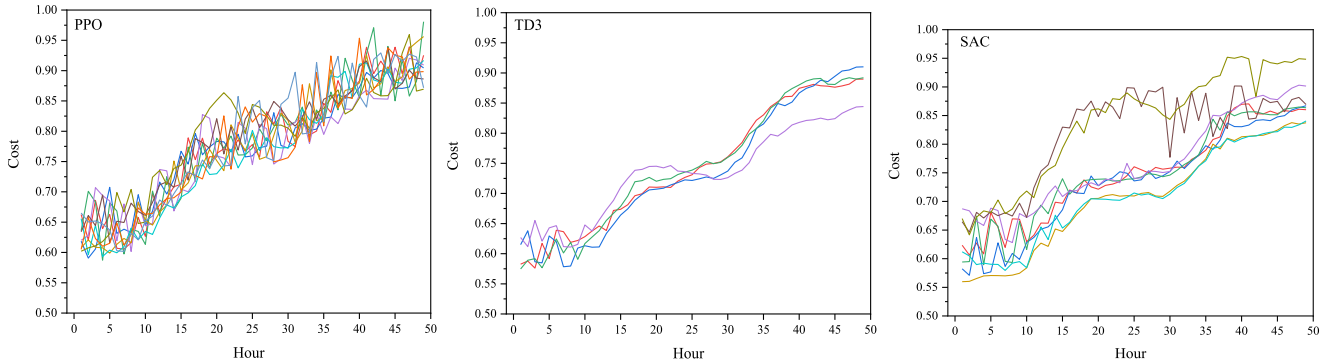

Fig. 8. The costs of PPO, TD3, and SAC.

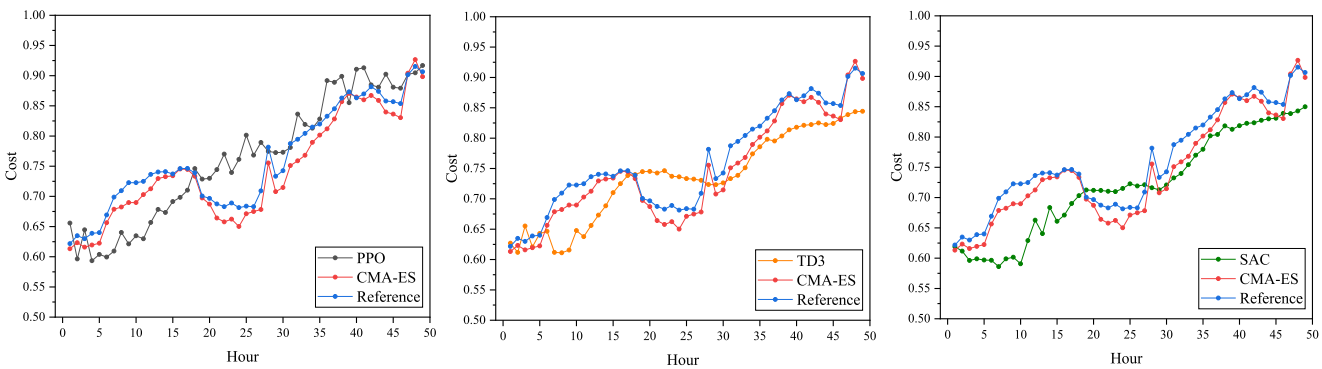

Fig. 9. The best solution of each DRL method.

Table 3
Comparison of costs in terms of sum, mean, and std.

	Reference	CMA-ES	PPO	TD3	SAC
Sum	37.29	36.47	37.42	35.97	35.26
Mean	0.76	0.74	0.76	0.73	0.72
STD	0.08	0.09	0.10	0.07	0.08

methods, which have STD values of 0.08 and 0.09, respectively. This implies that SAC and TD3 methods generate costs with more consistent and stable patterns of change over time.

As indicated in Table 3, SAC obtained the lowest total cost, the total cost reductions achieved by SAC over the reference, CMA-ES, PPO, and TD3 are 5.44 %, 3.32 %, 5.79 %, and 1.97 %, respectively. Fig. 10 depicts the cost reduction of the SAC method compared to the reference (blue), CMA-ES (green), PPO (orange), and TD3 (black) at each hour. However, it should be noted that the SAC method does not consistently yield lower costs at every hour, as indicated by negative cost reductions observed in each figure of Fig. 10. Compared to reference, it can be observed that from the 19th hour to the 27th hour, the costs generated by the SAC are higher, because the cost reductions are negative, ranging from 0 % to −5 %. Similarly, the same phenomenon is observed when comparing SAC with CMA-ES, PPO, and TD3, although SAC obtained the lowest total cost, it did not obtain the lowest cost at every hour.

Overall, according to Figs. 8–10, and Table 3, SAC is the best DRL method for lowering total fuel cost. The reduction over the reference, CMA-ES, PPO, and TD3 are 5.44 %, 3.32 %, 5.79 %, and 1.97 %; respectively. In contrast to the CMA-ES method, which consistently outperforms the reference method at every hour in terms of cost reduction, SAC adopts a different strategy. SAC does not aim to lower costs for every individual hour but instead focuses on achieving an overall lower total cost. As a result, there are specific hours where the SAC method may not exhibit cost reductions compared to CMA-ES. This distinct strategy employed by SAC highlights its approach of prioritizing overall cost optimization rather than hour-by-hour cost reduction.

6. Discussion

According to the results presented in Section 5.1, the proposed DRL framework outperformed the existing method and evolutionary algorithm based method in lowering the fuel cost. However, limitations regarding the convexity and model formulation need further discussion.

To perform optimization, a “skeleton” model combines pipes and consumers in such a way that the Apros model does not have to simulate every individual consumer and pipe in the system. The level of simplification determines the simulation speed and the accuracy of the results. With more detailed model, DRL will provide more accurate results but require more time for simulation. The studied case was formulated as a non-convex optimization problem, which means it is difficult for either DRL or evolutionary algorithms to find a global optimal solution. With the simplification of the model, the accuracy of the proposed method drops. The proposed framework can be flexibly adapted to different applications with different models (e.g., Modelica, Simulink, etc.). However, the principle of DRL requires a specific definition of the states, actions, and reward functions. Considering the safety, specific strategies are necessary to ensure that the generated results are safe and realistic.

7. Conclusion

In this paper, we have presented a DRL framework for setpoint optimization in district heating systems. The aim is to save energy and reduce fuel costs by effectively adjusting the supply temperature and utilization of thermal energy storage, considering variations in air temperature, customer demand, and fuel price. The limitations of manual operation methods, attributed to the nonlinearity and

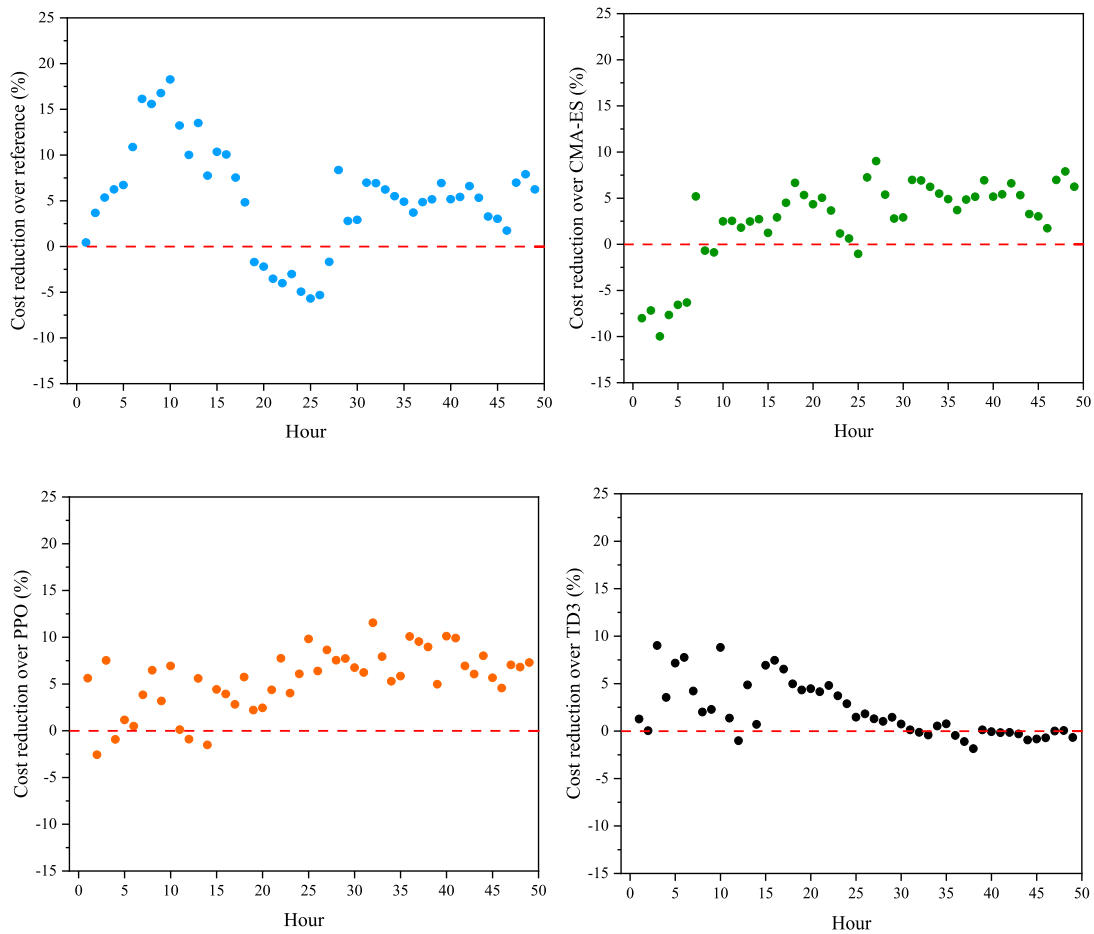

Fig. 10. Cost reduction over reference (blue), CMA-ES (green), PPO (orange), and TD3 (black).

complexity of the network, have been addressed through the proposed learning framework. The integration of an expert knowledge module ensures that executable actions are generated. The main intended contributions of this paper are as follows:

1. An Apros-based DT is provided, which ensures the training strategy is practical, and white-box models inside improve the practicality of the DRL solutions.
2. A novel learning framework for DRL with an EK module is designed to generate executable actions to avoid damage to the network.
3. State-of-the-art DRL methods are incorporated into the framework to optimize the setpoints. Simulation results show that the proposed framework outperforms the existing manual operation method and evolutionary method.

By incorporating state-of-the-art DRL methods into the framework, we have achieved significant advancements in setpoint optimization. The simulation results demonstrate the superiority of our proposed method over existing manual operation method and evolutionary techniques. The framework outperforms these traditional approaches, yielding substantial reductions in fuel costs. Furthermore, the provision of an Apros-based digital twin has enhanced the practicality of our training strategy. The inclusion of white-box models within the digital twin has improved the practicality of our DRL solutions, allowing for seamless application in real-world district heating systems.

This paper demonstrated that DRL is a promising tool for optimizing setpoints. The fundamental problems of the studied case are inefficient operation strategy and low computational efficiency. The goal is to lower the fuel costs. In future research, we intend to explore additional enhancements to the DRL methods that can fully extract underlying information of the industrial data to form better operation strategies. Currently, due to the low computational efficiency, only short-term (2-day) optimization was conducted. In the future, a long-term (e.g., 7-day) optimization will be performed by adopting parallel computation and increasing the sample efficiency of DRL methods. These extensions will provide the control system with more effective strategies for lowering fuel costs in a longer time frame, which can further improve the competitiveness of the district heating factory.

Declaration of Competing Interest

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

Data availability

The data that has been used is confidential.

Acknowledgment

This research was partly funded by Academy of Finland Grant No. 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.

References

Abdurafikov, R., et al. (2017). An analysis of heating energy scenarios of a Finnish case district. *Sustainable Cities and Society, 32*, 56–66. https://doi.org/10.1016/j.scs.2017.03.015

Alsagri, A. S., Arabkoohsar, A., Khosravi, M., & Alrobaian, A. A. (2019). Efficient and cost-effective district heating system with decentralized heat storage units, and triple-pipes. *Energy, 188*, Article 116035. https://doi.org/10.1016/j.energy.2019.116035

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*, 1–4.

Bucking, S., & Dermardiros, V. (2018). Distributed evolutionary algorithm for co-optimization of building and district systems for early community energy masterplanning. *Applied Soft Computing, 63*, 14–22. https://doi.org/10.1016/j.asoc.2017.10.044

De Somer, O., Soares, A., Vanthournout, K., Spiessens, F., Kuijpers, T., & Vossen, K. (2017). Using reinforcement learning for demand response of domestic hot water buffers: A real-life demonstration. In *Proceedings of the IEEE PES innovative smart grid technologies conference Europe* (pp. 1–7). ISGT-Europe. https://doi.org/10.1109/ISGTEurope.2017.8260152.

Eklund, M., Sierla, S. A., Niemistö, H., Korvola, T., Savolainen, J., & Karhela, T. A. (2023). Using a digital twin as the objective function for evolutionary algorithm applications in large scale industrial processes. *IEEE Access Practical Innovations Open Solutions, 11*, 24185–24202. https://doi.org/10.1109/ACCESS.2023.3254896

Fang, X., Zhao, Q., Wang, J., Han, Y., & Li, Y. (2021). Multi-agent deep reinforcement learning for distributed energy management and strategy optimization of microgrid market. *Sustainable Cities and Society, 74*, Article 103163. https://doi.org/10.1016/j.scs.2021.103163

Fazlollahi, S., Bungener, S. L., Becker, G., Maréchal, F., Bogle, I. D. L., & Fairweather, M. (2012). Multi-objectives, multi-period optimization of district heating networks using evolutionary algorithms and mixed integer linear programming (MILP)". in *Computer Aided Chemical Engineering, 30*, 262–266. https://doi.org/10.1016/B978-0-444-59519-5.50053-8. Eds., in 22 European Symposium on Computer Aided Process Engineering.

Frederiksen, S., & Werner, S. (2013). *District heating and cooling*. Professional Publishing Svc.

Fujimoto, S., Van Hoof, H., & Meger, D. (2018). Addressing function approximation error in actor-critic methods. *ICML, 4*, 2587–2601.

Gonzalez-Salazar, M., Klossek, J., Dubucq, P., & Punde, T. (2023). Portfolio optimization in district heating: Merit order or mixed integer linear programming? *Energy, 265*, Article 126277. https://doi.org/10.1016/j.energy.2022.126277

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *ICML, 5*, 2976–2989.

Huovilainen, R. T., & Koskelainen, L. (1982). *Kaukolämmitys. [Lappeenrannan teknillinen korkeakoulu]*.

Köppen, W. (2011). The thermal zones of the Earth according to the duration of hot, moderate and cold periods and to the impact of heat on the organic world. *Meteorologische Zeitschrift*, 351–360. https://doi.org/10.1127/0941-2948/2011/105

T. Leurs, B.J. Claessens, F. Ruelens, S. Weckx, and G. Deconinck, "Beyond theory: Experimental results of a self-learning air conditioning unit", in Proceedings of the IEEE international energy conference (ENERGYCON), Apr. 2016, pp. 1–6. doi: 10.1109/ENERGYCON.2016.7513916.

Lumbreras, M., et al. (2022). Data driven model for heat load prediction in buildings connected to district heating by using smart heat meters. *Energy, 239*, Article 122318. https://doi.org/10.1016/j.energy.2021.122318

Majid, A. Y., Saaybi, S., Francois-Lavet, V., Prasad, R. V., & Verhoeven, C. (2023). Deep reinforcement learning versus evolution strategies: a comparative survey. *IEEE Transactions on Neural Networks and Learning Systems*, 1–19. https://doi.org/10.1109/TNNLS.2023.3264540

Patankar, S. (2018). *Numerical heat transfer and fluid flow*. Boca Raton: CRC Press. https://doi.org/10.1201/9781482234213

Patyn, C., Ruelens, F., & Deconinck, G. (2018). Comparing neural architectures for demand response through model-free reinforcement learning for heat pump control. In *Proceedings of the IEEE international energy conference (ENERGYCON)* (pp. 1–6). https://doi.org/10.1109/ENERGYCON.2018.8398836

Qin, Y., Ke, J., Wang, B., & Filaretov, G. F. (2022). Energy optimization for regional buildings based on distributed reinforcement learning. *Sustainable Cities and Society, 78*, Article 103625. https://doi.org/10.1016/j.scs.2021.103625

Ren, M., Liu, X., Yang, Z., Zhang, J., Guo, Y., & Jia, Y. (2022). A novel forecasting based scheduling method for household energy management system based on deep reinforcement learning. *Sustainable Cities and Society, 76*, Article 103207. https://doi.org/10.1016/j.scs.2021.103207

Ruelens, F., Claessens, B. J., Vandael, S., De Schutter, B., Babuška, R., & Belmans, R. (2017). Residential demand response of thermostatically controlled loads using batch reinforcement learning. *IEEE Transactions on Smart Grid, 8*(5), 2149–2159. https://doi.org/10.1109/TSG.2016.2517211

Ruelens, F., Claessens, B. J., Vrancx, P., Spiessens, F., & Deconinck, G. (2019). Direct load control of thermostatically controlled loads based on sparse observations using deep reinforcement learning. *CSEE Journal of Power and Energy Systems, 5*(4), 423–432. https://doi.org/10.17775/CSEEJPES.2019.00590

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 1–12.

Silvennoinen, E., Juslin, K., Hänninen, M., Tiihonen, O., Kurki, J., & Porkholm, K. (1989). The APROS software for process simulation and model development. *Valtion teknillinen tutkimuskeskus. tutkimuksia - Research Reports*. Espoo: VTT Technical Research Centre of Finland.

Solinas, F. M., Bottaccioli, L., Guelpa, E., Verda, V., & Patti, E. (2021). Peak shaving in district heating exploiting reinforcement learning and agent-based modelling. *Engineering Applications of Artificial Intelligence, 102*, Article 104235. https://doi.org/10.1016/j.engappai.2021.104235

Stepanovic, K., Wu, J., Everhardt, R., & de Weerdt, M. (2022). Unlocking the flexibility of district heating pipeline energy storage with reinforcement learning. *Energies, 15*(9). https://doi.org/10.3390/en15093290. Art. no. 9.

Su, L., et al. (2022). Optimizing pipe network design and central plant positioning of district heating and cooling System: A Graph-Based Multi-Objective genetic algorithm approach. *Applied Energy, 325*, Article 119844. https://doi.org/10.1016/j.apenergy.2022.119844

Sun, J., Gong, M., Zhao, Y., Han, C., Jing, L., & Yang, P. (2022). A hybrid deep reinforcement learning ensemble optimization model for heat load energy-saving prediction. *Journal of Building Engineering, 58*, Article 105031. https://doi.org/10.1016/j.jobe.2022.105031

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction, second edition. in adaptive computation and machine learning series*. Cambridge, Massachusetts: The MIT Press.

Wang, C., Gong, Z., He, C., Gao, H., & Bi, T. (2021). Data-driven adjustable robust unit commitment of integrated electric-heat systems. *IEEE Transactions on Power Systems, 36*(2), 1385–1398. https://doi.org/10.1109/TPWRS.2020.3019412

Wei, Z., et al. (2022). Data-driven application on the optimization of a heat pump system for district heating load supply: A validation based on onsite test. *Energy Conversion and Management, 266*, Article 115851. https://doi.org/10.1016/j.enconman.2022.115851

Werner, S. (2017). International review of district heating and cooling. *Energy, 137*, 617–631. https://doi.org/10.1016/j.energy.2017.04.045

Zhao, H., Wang, B., Liu, H., Sun, H., Pan, Z., & Guo, Q. (2022). 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, 13*(1), 207–219. https://doi.org/10.1109/TSTE.2021.3107439

Żymełka, P., & Szega, M. (2021). Short-term scheduling of gas-fired CHP plant with thermal storage using optimization algorithm and forecasting models. *Energy Conversion and Management, 231*, Article 113860. https://doi.org/10.1016/j.enconman.2021.113860