Build Simul (2022) 15: 233–248
https://doi.org/10.1007/s12273-021-0808-5

Differential pressure reset strategy based on reinforcement learning for chilled water systems

Xinfang Zhang[1], Zhenhai Li[1], Zhengwei Li[1,2] (✉), Shunian Qiu[1], Hai Wang[1]

1. School of Mechanical Engineering, Tongji University, Shanghai, China
2. Key Laboratory of Performance Evolution and Control for Engineering Structures of Ministry of Education, Tongji University, Shanghai, China

Abstract

Air conditioning water systems account for a large proportion of building energy consumption. In a pressure-controlled water system, one of the key measures to save energy is to adjust the differential pressure setpoints during operation. Typically, such adjustments are based either on certain rules, which rely on operator experience, or on complicated models that are not easy to calibrate. In this paper, a data-driven control method based on reinforcement learning is proposed. The main idea is to construct an agent model that adapts to the researched problem. Instead of directly being told how to react, the agent must rely on its own experiences to learn. Compared with traditional control strategies, reinforcement learning control (RLC) exhibits more accurate and steady performances while maintaining indoor air temperature within a limited range. A case study shows that the RLC strategy is able to save substantial amounts of energy.

Keywords

water system;
differential pressure reset;
reinforcement learning control;
energy saving

Article History

Received: 12 December 2020
Revised: 03 April 2021
Accepted: 13 April 2021

1 Introduction

Heating, ventilation and air-conditioning (HVAC) systems account for more than half of the total building energy consumption (Pérez-Lombard et al. 2008; Li et al. 2016; Hou et al. 2018). In a typical HVAC system, a chilled water system is an essential sub-loop whose operation significantly influences the entire system. Based on whether the water flow rate is adjustable, the chilled water systems can be divided into constant flow systems and variable flow systems. In variable flow systems, the normal operation strategies include control based on the difference in the supply and return water temperatures (delta-T control) (Liu et al. 2012; Gao et al. 2016) and control based on the difference in the supply and return water pressures (delta-P control) (Jin et al. 2007; Ji et al. 2009; Zhao et al. 2016). Of these, delta-P control is used more often; thus, the key control parameter is the differential pressure (DP) of water loops.

Many scholars have worked on optimal water system DP control methods. Jin et al. (2007) updated DP according to the AHU's valve openness, which was constrained to be close to 100% open under the control of two PID controllers. Zhao et al. (2016) also used valve position as a monitored signal to reset DP. They analyzed the difference between unfavorable thermodynamic loops and unfavorable hydraulic loops and chose the latter as a judgment reference. They also pointed out that the maximum valve position was not necessarily defined as completely open and redesigned the optimal reset valve position domain. Ji et al. (2009) proposed a feed-forward fuzzy immune (FFIM) control algorithm to improve PID control of water loops based on the rule that a valve control loop responds more quickly than does a pressure loop considering the pressure loop's time-varying, nonlinear characteristics.

In traditional delta-P control, the core rules are predefined, while the control process relies primarily on feedback information. This can cause two problems: (1) predefined rules usually do not directly fit a real system without online scrolling or other subsequent corrective operations; and (2) the feedback process takes time, causing changes in the control signal to lag behind changes in the controlled variable and making it difficult to achieve a good

E-mail: zhengwei_li@tongji.edu.cn

predictive effect (Ma and Wang 2009). In contrast, in an RLC strategy, an agent's gained experience relies on responses from the real environment; thus, it can fully reflect a real system situation. Meanwhile, intermediate units are implicit during the training process, and relationships between control signals and their consequent influences are learned from end to end, which avoids the time delay problem. An agent's action can be any desired decision, and the states are factors that contribute to decisions. Policy is a mapping from states to actions, and its quality is evaluated by a quantitative feedback reward. In summary, agents constantly improve the quality of their decisions by interacting with the environment. In this way, an RLC controller can quickly adapt and react to unknown environments without truly understanding the underlying mechanisms (Lewis et al. 2012).

The RLC strategy has been applied to building operation areas for various purposes (Han et al. 2019). Liu and Henze (2006a, b) used an RLC method to optimize the charge and discharge rates of ice tanks as well as indoor air temperature setpoints. As a result, the building operational cost was reduced by 9.9% compared with a benchmark building. Yang et al. (2015) optimized the energy system of a low-carbon building in Zurich using an RLC strategy. This system consisted of three water loops, including a solar hot water loop, a primary air source heat pump water loop, and a secondary ground source heat pump water loop. An RLC strategy was deployed to optimize the water flow rates of all three loops. After three years of learning, the energy savings reached 10%. In addition, Barrett and Linder (2015) developed an RLC framework to control a household air-conditioning system. In their design, the RLC controller managed the on/off status of the heating and cooling system by observing the indoor and outdoor air temperatures and household occupancy. Simulation results showed that this RLC strategy can achieve cost savings of 10%. Chen et al. (2018) employed an RLC strategy to determine when and how to utilize natural ventilation and coordinated its operation with an HVAC system. The RLC controller responded by making decisions that targeted both immediate and long-term goals at each time step. The algorithm was evaluated by numerical simulation, and the results showed that an energy savings of 13% compared to the heuristic control method. Meanwhile, even though the RLC strategy has attracted growing research interest, it is still in the research and development stage. Among 77 studies reviewed by Wang and Hong (2020), only nine controllers were implemented in real buildings: 3 domestic hot water controllers, 3 HVAC controllers, 2 lighting controllers and 1 window controller.

Considering the performance of the RLC strategy in the abovementioned studies, it is natural to choose RLC as the control strategy for the DP setpoint reset problem of a chilled water system. The remainder of this paper is organized as follows. Section 2 introduces the general methodology of RLC. Section 3 presents the detailed algorithm and control framework of the chilled water DP setpoint problem. Section 4 shows the simulation model and control performance of traditional and proposed strategies. Finally, Section 5 discusses the findings and provides concluding remarks.

2 General RLC methodology

Reinforcement learning (RL) involves learning how to map states to actions by maximizing cumulative rewards (Dayan and Niv 2008). RL is based on the Markov decision process (MDP), which is a sequential decision mathematical model used to simulate random strategies. MDP $<S, A, P, R, \gamma>$ contains two interacting objects: agents and the environment, and four key elements: state, action, policy and reward (Recht 2019). During the training process, the agent collects information from the environment and selects the action that maximizes the cumulative reward. This basic principle of RL is shown in Figure 1. In a specific application scenario, the most important step is to define the boundary between the agent control and environmental influence. Modules that agents cannot change are considered to belong to the environment. After that boundary has been determined, a particular decision task forms naturally.

There are two major categories of RL algorithms: model-based RLs and model-free RLs. According to the research, 77% of existing studies used value-based RL algorithms, among which Q-learning is the most popular. Actor-critic and policy gradient approaches have become more frequently used since 2017 due to their ability to facilitate transfer learning (Wang and Hong 2020). In fact, almost all reinforcement learning problems can be described as a generalized strategy iteration process, which can be further divided into two processes, strategy evaluation and strategy improvement (Sutton and Barto 2018). Strategy evaluation involves updating the agent's state value function or state-action value function based on feedback from the environment, while strategy improvement focuses on how to optimize the decision-making strategy according to the value function. Commonly used methods in strategy evaluation can be divided into the dynamic programming, Monte Carlo, and time difference

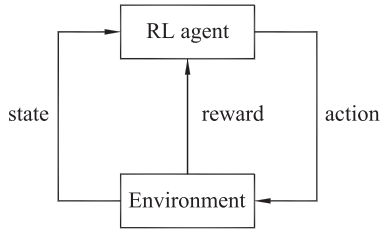


Fig. 1 Basic principle of RL

methods. Specifically, in the dynamic programming method, the current state value is updated based on other state values; then, an accurate probabilistic transition model is needed to predict all possible next states in an MDP process. In Monte Carlo method, multiple trajectories are obtained by sampling to replace the transition model, and the value of each state is calculated by mathematical averaging. Time difference method utilizes both sampling and bootstrap operations. To accelerate the convergence process, state value calculations do not need information from the entire trajectory; they require only the values of the subsequent state. The strategy evaluation and strategy improvement processes cooperate with each other; thus, the optimization results gradually approach a final overall goal.

Based on whether the strategy for generating the sampling data sequence is the same as the strategy that actually plays a decision-making role, TD can be further divided into Sarsa, Q-learning and other methods. The updated definition of Q-learning is:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right] \quad (1)$$

where $Q(S_t, A_t)$ is state-action pair's value at time t; α is a constant step size parameter when updating is in augmented form; γ is the discount rate, which determines the present value of future rewards; $\max_a Q(S_{t+1}, a)$ is the optimal action value of the next moment, which means that the strategy that generates the sampling data sequence is greedy and differs from the strategy that generates the trajectory of the agent's decision sequence. Thus, the Q-learning method is offline learning.

The corresponding pseudocode for the updating process is shown below.

Initialize *Q-Table*
Repeat:
  Interact *with environment (retrieve data from database or software model output)*
  Read *current state(s) of agent*
  Choose *action (a) corresponding to state(s) from Q-table according to policy π*
  Calculate *reward (r) using environment feedback data*
  Update *Q-learning table following rule: Q(s,a) ← Q(s,a) + α[R + γ max$_{a'}$Q(s',a') − Q(s,a)]*
  Transmit *action information to environment*
  Cover *s with s'*
Until *s is terminal*

3 RL-based water loop control

3.1 Typical delta-P control water system

As shown in Figure 2, in a typical delta-P control water system, the chilled water is supplied to the AHUs by a set of variable speed pumps. Each water valve in the AHU is controlled by a PID controller that attempts to maintain the supply air temperature at a setpoint. The speeds of the variable speed pumps are controlled by a DP controller, which attempts to maintain the pressure differential of the end-user loops at a setpoint.

Figure 3 shows the energy saving potential of the system by adjusting the DP value. The three types of curves are pump curves, control curves and pipe characteristic curves. Pump curves are a cluster of curves with similar shapes that show the operation characteristics of pumps at different frequencies. According to Bernoulli's principle, the pump head is equal to the pressure drop of the most unfavorable loop, which consists of main pipes and end-user pipes. The pressure–flow relationship of the former is approximately quadratic, while the pressure drop of the latter is fixed; thus, the control curves are a set of quadratic curves that intersect the y-axis at the DP setpoints. Pipe characteristic curves are mainly affected by the valve opening degree: the smaller the opening degree is, the steeper the curve is. The intersection points of these three curves are the actual operating points of the pumps. When DP is set to $DP_{set,1}$, the valve openness is small, and pumps run at high frequency. Keeping the end users' demand flow rate unchanged and reducing the DP setpoint to $DP_{set,2}$, when the system is stable again, the pump frequency will decrease and the valve openness will simultaneously increase. The corresponding reduced pump head of different cases is marked as reduced heads I, II, and III. When the most unfavorable loop's valve openness approaches 100%, the DP setpoint will reach its optimum value and cannot be further decreased without sacrificing user comfort.

3.2 RL-based DP setpoint reset strategy

The proposed RL-based DP setpoint reset strategy is describe below. An RL agent model includes four key elements, i.e., state, action, policy and reward.

3.2.1 States

A state is a time slice of the current situation and should fully reflect the characteristics of the researched problem. In an HVAC system, the influencing factors include thermal and hydraulic environmental factors. Therefore, three state variables are proposed: the outdoor dry-bulb air temperature T_{out}, the valve opening degree ζ_{max}, ζ_{med}, ζ_{min} in statistical format (which refer to the maximum, medium and minimum values of the valve opening degree sequence, respectively) and the differential pressure setpoint DP_{set}.

The state variables include both observed variables, which are used to assist environment evaluation, and controlled

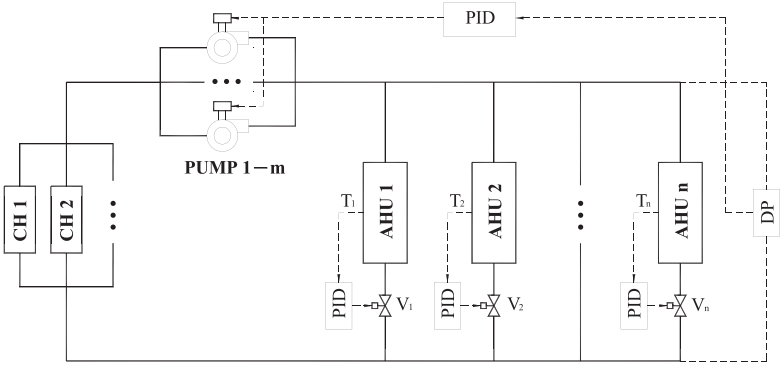


Fig. 2 A typical delta-P control water system

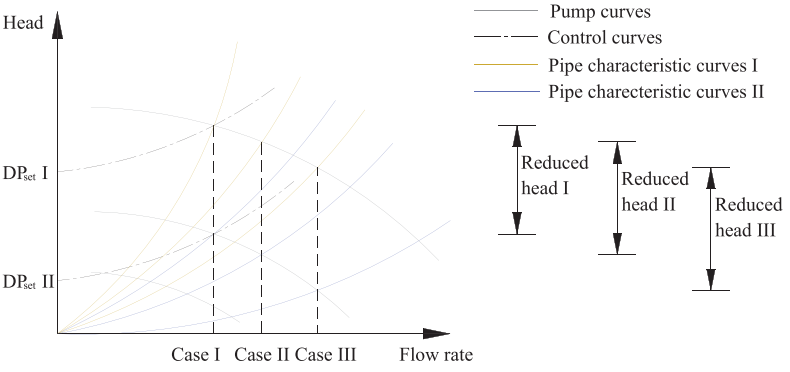


Fig. 3 Potential energy savings of the system at different DP set values

variables, which are regulated by actions and update at each iteration step. T_{out} is an observed variable introduced to describe outdoor environmental changes. Theoretically, other variables, including wet bulb temperature, wind speed, and others could also be added. However, the dry bulb temperature is still considered the most representative. The valve opening degree is also an observed variable that reflects changes in the indoor thermal environment and hydraulic environment. When the number of water valves is large, a statistical method is recommended to reduce the dimensionality of the variables while maintaining the accuracy of the results. DP_{set} is a controlled variable and plays a central role in the whole regulation process.

The state variables can be divided into continuous variables and discrete variables, depending on the problem. Regardless of which variable type is chosen, the core of the algorithm remains unchanged. The general principle is a trade-off between the algorithm's convergence speed and its calculation accuracy. For the normal DP setpoint reset problem, discrete variables are sufficiently effective when the space is relatively small. Specifically, T_{out} is accurate to the integer level. ζ is discretized into ten states, namely, state I (when the valve opening degree is between 0% and 10%), state II (when the valve opening degree is between 11% and 20%), state III (when the valve opening degree is between 21% and 30%),…, and state X (when the valve opening degree is between 91% and 100%). DP_{set} is updated by a fixed interval value ΔP.

3.2.2 *Actions*

An action is a description of an agent's change behavior. State changes can be divided into two modes: step or jump mode. In step mode, an action is usually a fixed set of values. For example, in a maze game, players' actions are limited to up, down, left and right. The next player position can be realized in only one of four adjacent positions to the previous position, as shown in Figure 4(a). Similarly, the action of the DP setpoint reset agent can be defined as $\{-\Delta P, 0, \Delta P\}$, and the DP_{set} of the next step is calculated by

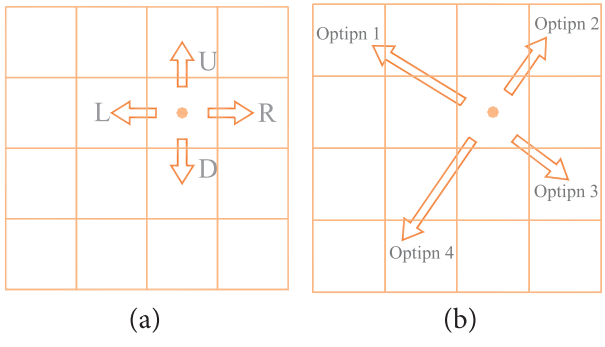


Fig. 4 Different state change modes

adding or subtracting a fixed value (ΔP) from the previous state or by leaving it unchanged.

In jump mode, the actions are considerably more flexible, and the difference between two exclusive states can be substantial. For example, in the maze game, the player's next position is uncertain, as shown in Figure 4(b), and in the DP setpoint reset problem, the DP_{set} value of the next state can be equal to any value as long as it is within the reasonable range.

In an actual chilled water system with multiple PID controllers, the value is likely to be out of the safe adjustment range when using jump mode. Thus, continuous mode is recommended and it is utilized in the following case study.

3.2.3 Policy

As mentioned above, the goal of the RL agent is to obtain the best performance strategy through continuous trial and exploring behaviors (where a trial means an agent trying actions that it has never been chosen before and exploring means the agent choosing actions that already exist and gaining explorational experience to update the rewards). In Monte Carlo theory, a sample's trajectory should be unbiased to ensure the accuracy of true-value estimation. Therefore, to avoid initial value interference, a certain randomness is introduced to ensure that almost all states or state-action pairs in space have an opportunity to be taken. Thus, state updating is actually a gamble between trial and exploration, where the weight of a trial is controlled by different parameters in different policies. The commonly used policies include the greedy, ε-greedy, Gaussian, and Boltzmann distribution strategies. In this study, the ε-greedy strategy is employed; it uses a parameter ε to describe the selection's greediness' degree.

First, the program generates a random number *Rand* and compares it with $1-\varepsilon$. If $1-\varepsilon$ is larger, then the agent moves greedily, which means that the agent will choose the action whose state-action pair maps to the maximum Q-value in the table. Otherwise, all actions will be selected at the same probability.

$$Action = \begin{cases} a_i \left(q_i = \max(q_1,\ldots,q_n)\right) & \text{when } 1-\varepsilon \geq Rand \\ a_j \left(a_j = \text{randof}(a_1,\ldots,a_n)\right) & \text{when } 1-\varepsilon < Rand \end{cases} \tag{2}$$

Under this strategy, the action variables will satisfy the following probability distribution:

$$P_\pi\left(a_i \middle| s\right) = \begin{cases} 1-\varepsilon+\dfrac{\varepsilon}{n} & \text{if } a_i = \max_a Q(s,a) \\ \dfrac{\varepsilon}{n} & \text{if } a_i \neq \max_a Q(s,a) \end{cases} \tag{3}$$

where $P_\pi\left(a_i \middle| s\right)$ is the probability of taking action a_i under state s, and n is the number of alternative actions corresponding to state s.

To summarize, a smaller ε causes greedier selection. When $\varepsilon = 0$, the strategy involves only exploration, and when $\varepsilon = 1$, the strategy involves only trials. Moreover, rather than remaining fixed, ε can be a variable that changes in each iteration and gradually increases the greedy degree, such as

$$\varepsilon\_nextstep = \varepsilon - a \times circletime \tag{4}$$

where $\varepsilon\_nextstep$ is the ε value of next step, a is a scaling factor used to control the changing rate of ε, and *circletime* is the current number of algorithm iterations.

3.2.4 Reward

A reward is a quantified feedback. Since the goal of the DP reset problem is to achieve a balance between energy savings and thermal comfort, the variables selected to calculate the feedback reward should include both these aspects. The reward function is defined as follows:

$$R = \alpha_1 \frac{E_{\max} - E_p}{E_0} + \alpha_2 \frac{T_{\text{limit}} - T_{\text{in}}}{T_0} \tag{5}$$

where R is the quantified reward; E_p is the real-time water pump energy consumption (kW); $E_{\max}$ is the upper limit of water pump power consumption inferred from the historical data (kW); E_0 is the auxiliary value used to nondimensionalize pump energy consumption (kW); T_{in} is the average of all zones' real-time indoor air temperature (K); T_{limit} is the upper limit of indoor air temperature defined by the user (K); T_0 is the auxiliary value used to nondimensionalize indoor air temperature (K); α_1 is the weight of energy savings in the comprehensive result; α_2 is the weight of thermal comfort in the comprehensive result.

When the strategy tends toward saving energy, α_1 is higher, and when the thermal comfort and people's feelings are more important, α_2 is higher. The weights are set by considering both the building characteristics and the owner's requirements.

3.3 Implementation of the RL controller

The overall update process of the RL-based DP setpoint reset strategy is shown in Figure 5. The algorithm consists of two parts: agent and environment. In addition to the abovementioned information, in an actual building automation (BA) system, the central control platform usually does not directly communicate with the underlying equipment; instead, it transmits commands through middle layers, which are mostly PID or DDC controllers. Thus, the RL algorithm is designed to communicate with the upper control layer through the API interface without changing

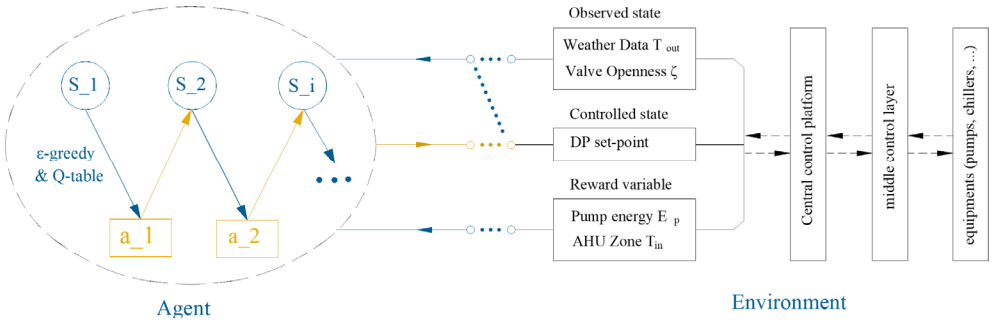


Fig. 5 Update process for the RL-based DP setpoint reset strategy

the original hierarchical control framework. The lower level uses traditional controllers to maintain sub-loop stability, while the upper level modifies the key controller parameters based on the agent's final action decision. Compared with a structure in which the algorithm directly interacts with bottom equipment, this design can effectively avoid totally changing existing hardware and software and can save a portion of the cost of upgrading an older system.

4 Case study

4.1 Model description

To validate the proposed strategy, a test model must be established that can accurately simulate a real HVAC system. According to the operation routine of a typical commercial building located in Shanghai, an HVAC system containing three different air conditioning zones was established, as shown in Figure 6.

The HVAC system is a typical primary pump variable flow system. Specifically, zones 1–3 represent three different kinds of air conditioning zones, and the basic information is set before beginning the simulation, as shown in Table 1. The modules are connected to weather buses; the weather data are downloaded from the EnergyPlus official website. Dampers 1–3 are used to adjust the supply air volume. Fans 1–3 deliver cooled air to bear the cooling load. AHUs 1–3 are equipped with centralized air treatment devices. Valves 1–3 adjust the supply water flow rate: the opening degree is controlled by local PID controllers that monitor the differential between the real indoor temperature and a predefined setpoint to calculate the control signal. Pipes are used to transmit corresponding liquids. A pressure sensor is responsible for monitoring the differential pressure in

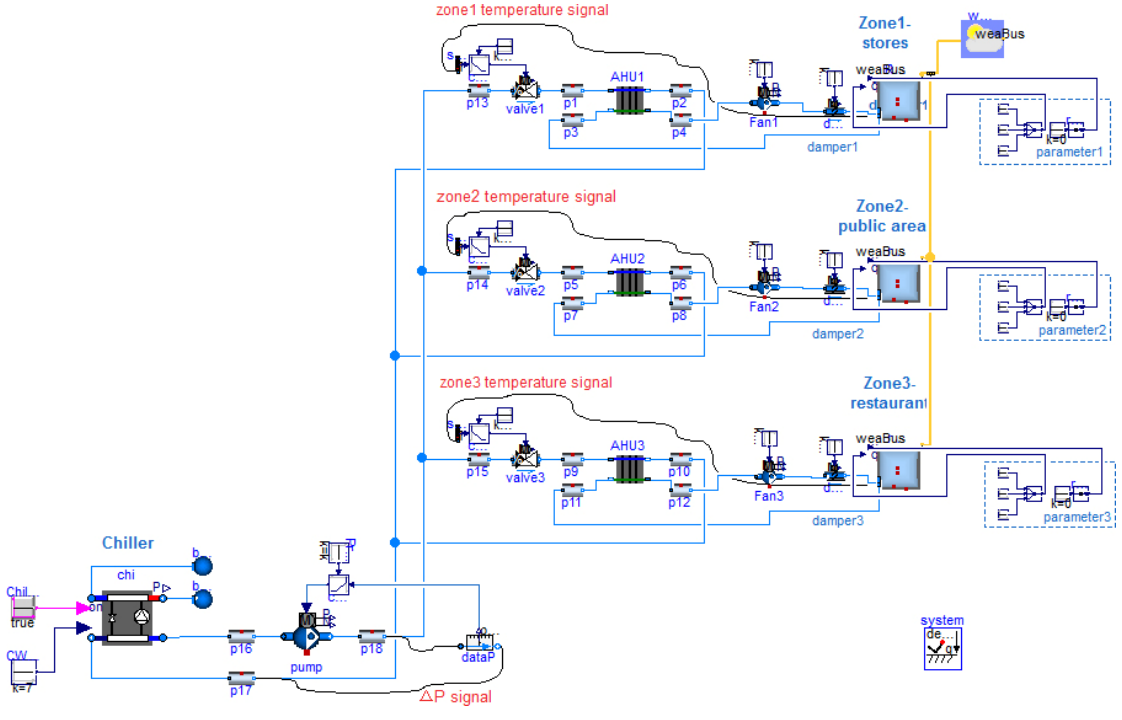


Fig. 6 Schematic diagram of a typical variable flow water system built in Dymola (related models have been packed and are available in the Electronic Supplementary Material (ESM) in the online version of this paper)

end-user loops and sending the signal to the pump frequency regulation controller. The chiller, which has perfect adjustment characteristics, can output self-real-time PLR.

To describe different user situations, the daily personnel occupancy rates of the three zones are shown in Figure 7.

Table 1 Basic information of HVAC system built in Dymola

Location	Shanghai
Orientation	North
Weather data	CHN_Shanghai.Shanghai.583620_SWERA.mos
External walls *U*-value	0.45 W/(m^2·K)
Window *U*-value	2.7 W/(m^2·K)
Zone 1 use	Public area
Total floor area	2200 m^2
Height	5 m
Zone 2 use	Stores
Total floor area	3500 m^2
Height	5 m
Zone 3 use	Restaurant
Total floor area	1500 m^2
Height	5 m

The public area's occupancy rate is relatively stable; some people are always present. The independent store occupancy rate has three peaks throughout the day: 10:00, 15:00 and 19:00. For restaurants, the peaks occur at typical mealtimes (e.g., 12:00 and 18:00). Equipment and lighting occupancy rates change with the same trends. The zone total cooling load during the entire cooling season (from June 1st to September 30th) is shown in Figure 8.

Horizontally, the different point colors reflect the daily load fluctuations, while vertically, the middle parts in Figure 8 are lighter, while the edges are darker. This situation reflects the fact that the cooling load reaches its peak value in early August and declines slowly toward the beginning or end of the cooling season. The zones are not further subdivided into smaller rooms because the model size would be larger and the simulation time would also be longer.

Dymola includes various packages, including fluids, HVAC equipment, PID controllers and Python interfaces (Mehlhase 2012) that can simulate the details of the system well. The input parameters for the model are shown in Table 2.

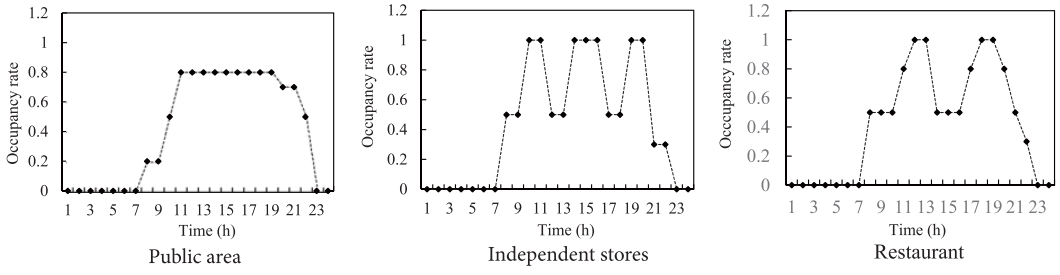


Fig. 7 Personnel occupancy rate of different zones

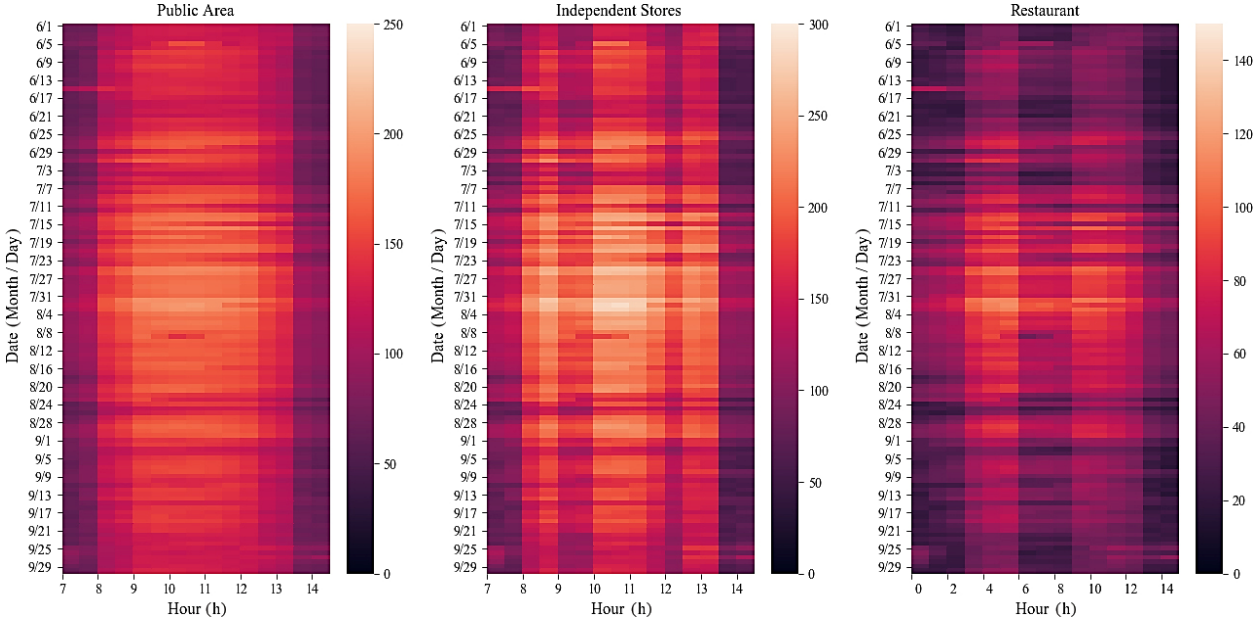


Fig. 8 Cooling loads of different zones throughout the entire cooling season (unit: kW)

Table 2 Parameters of the model built in Dymola

Model	Component	Parameters					
Chiller	Buildings.Fluid.Chillers. Electric EIR	Rated capacity	830 kW	Rated COP	6.97		
		Nominal water flow rate 1	30 kg/s	Nominal resistance 1	50,000 Pa		
		Nominal water flow rate 2	40 kg/s	Nominal resistance 2	50,000 Pa		
Pump	Buildings.Fluid.Movers. flowmachine_Nrpm	Nominal water flow rate	30 kg/s	Nominal supply head	282,500 Pa		
		V_flow={0.0025,0.0148,0.03,0.035} kg/s, dp={320000,282500,217500,7000} Pa					
Valves 1–3	IBPSA.Fluid.Actuators. baseclasses.partialtwowayvalve	Nominal water flow rate 1	15 kg/s	Nominal resistance 1	60,000 Pa		
		Nominal water flow rate 2	12 kg/s	Nominal resistance 2	45,000 Pa		
		Nominal water flow rate 3	5 kg/s	Nominal resistance 3	30,000 Pa		
AHUs 1–3	Buildings.Fluid.heatexchangers. drycoilcounterflow	Nominal water flow rate 1	15 kg/s	Nominal airflow rate 1	50 kg/s		
		Nominal UA 1	40 kW/K	Nominal resistance 1	35,000 Pa		
		Nominal water flow rate 2	12 kg/s	Nominal airflow rate 2	40 kg/s		
		Nominal UA 2	26 kW/K	Nominal resistance 2	30,000 Pa		
		Nominal water flow rate 3	5 kg/s	Nominal airflow rate 3	20 kg/s		
		Nominal UA 3	10 kW/K	Nominal resistance 3	30,000 Pa		
Fans 1–3	Buildings.Fluid.Movers. flowmachine_Nrpm	Nominal n	1500 r/min	V_flow={0,42,83} kg/s, dp={4000,1500,0}Pa			
		Nominal n	1500 r/min	V_flow={0,30,60} kg/s, dp={1600,800,0}Pa			
		Nominal n	1500 r/min	V_flow={0,20,40} kg/s, dp={2000,1200,0}Pa			
Pipes 3, 4	IBPSA.Fluid.BaseClasses. PartialResistance	Nominal airflow rate	50 kg/s	Nominal resistance	1,500 Pa		
Pipes 7, 8		Nominal air flow rate	20 kg/s	Nominal resistance	600 Pa		
Pipes 11, 12		Nominal airflow rate	12 kg/s	Nominal resistance	400 Pa		
Pipes 1, 2		Nominal airflow rate	15 kg/s	Nominal resistance	10,000 Pa		
Pipes 5, 6		Nominal airflow rate	12 kg/s	Nominal resistance	10,000 Pa		
Pipes 9, 10		Nominal airflow rate	5 kg/s	Nominal resistance	10,000 Pa		
Pipes 13–18		Nominal airflow rate	30 kg/s	Nominal resistance	30,000 Pa		
PIDs 1–4	Buildings.Controls.Continuous. LimPID	Input signal u_1	Zone 1 temperature (K)	Output signal y_1	Valve 1 opening degree		
		Input signal u_2	Zone 2 temperature (K)	Output signal y_2	Valve 2 opening degree		
		Input signal u_3	Zone 3 temperature (K)	Output signal y_3	Valve 3 opening degree		
		Input signal u_4	DP (Pa)	Output signal y_4	Pump rotating speed		
		$U_{set,1}$	300 K	K_1	0.15	$T_{i,1}$	5,000
		$U_{set,2}$	300 K	K_2	0.15	$T_{i,2}$	5,000
		$U_{set,3}$	300 K	K_3	0.15	$T_{i,3}$	5,000
		$U_{set,4}$	180,000 Pa	K_4	0.05	$T_{i,4}$	Changeable

Note: for chiller, 1: evaporator, 2: condenser; for others, 1: public area, 2: independent stores, 3: restaurants.

To verify the accuracy of the model, the hourly temperature fluctuations in different zones are drawn. As shown in Figure 9 (i), the indoor air temperature of the public area is kept at 27 °C (300 K), and should fluctuate by no more than ±1 °C. Two representative fragments (marked in dotted rectangular boxes) show temperature variation details; the corresponding time ranges are 1,400–1,448 h (Figure 9 (ii)) and 1,850–1,898 h (Figure 9 (iii)). Although the two fragments are not quite the same, the temperature is successfully maintained at the setpoint, which indicates that the PID controllers are tuned successfully. The hourly temperature variations in the other zones are shown in Figures 10–11.

To perform a quantitative comparison, the distribution of temperature change in three different zones is presented in Figure 12, and the corresponding statistical characteristics are shown in Table 3.

To explore the influence of DP_{set} on the system hydraulic characteristics, the valve opening degree, the water loop differential pressure, and the pump energy consumption under different DP_{set} are plotted in Figs. 13–15. The details are also shown to help with the analysis. As shown, when DP_{set} decreases, the valve openness will increase and the differential pressure of water loops will decrease, which leads to a reduction in pump energy consumption.

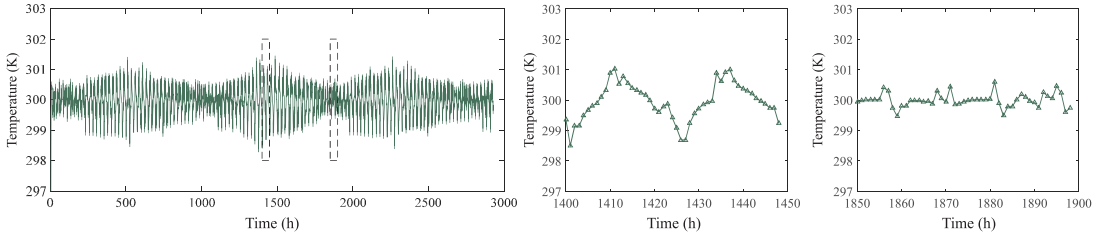


Fig. 9 Hourly temperature variations in the public area

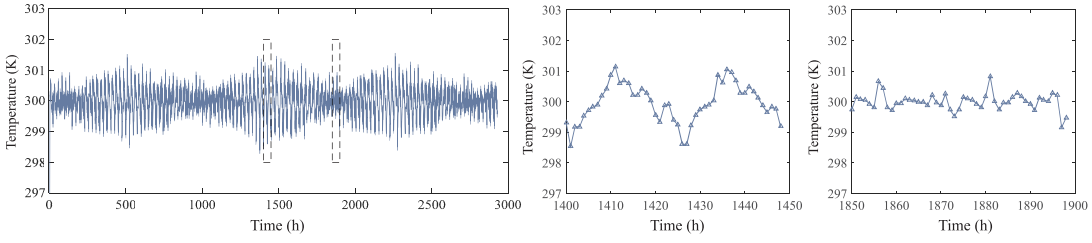


Fig. 10 Hourly temperature variation of independent stores

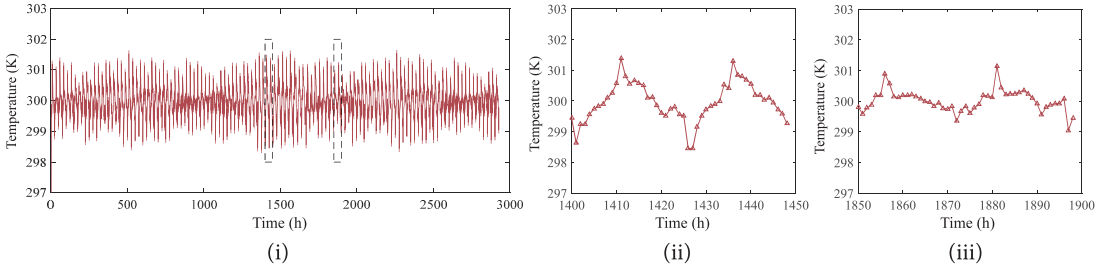


Fig. 11 Hourly temperature variations in restaurants

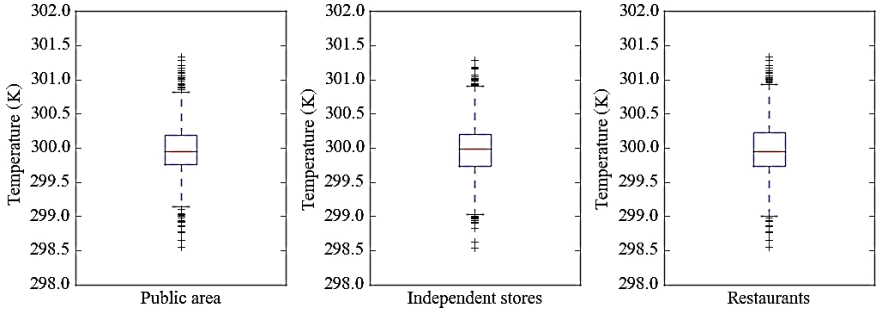


Fig. 12 Distribution of temperature changes for the three zones

Table 3 Statistical characteristics of the temperature sequences (unit: K)

Zone	Mean	Std	Min	25%	50%	75%	Max
Public area	299.98	0.3848	298.51	299.76	299.96	300.20	301.33
Stores	299.96	0.4287	298. 51	299.74	299.98	300.21	301.28
Restaurants	299.96	0.4414	298. 51	299.74	299.96	300.22	301.33

Note: std—standard deviation.

Similarly, the variable distributions are also depicted, as shown in Figure 16. The corresponding statistical characteristics are presented in Tables 4–6. It can be seen from the charts that the system is sensitive to the control variable, which provides a good environment for strategy testing. The changes of the variables are scaling similar to

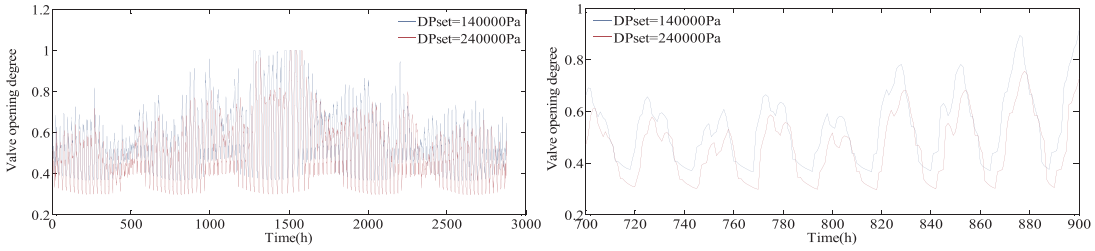


Fig. 13 Valve opening degree under different DP_{set} values (taking valve 3 as an example)

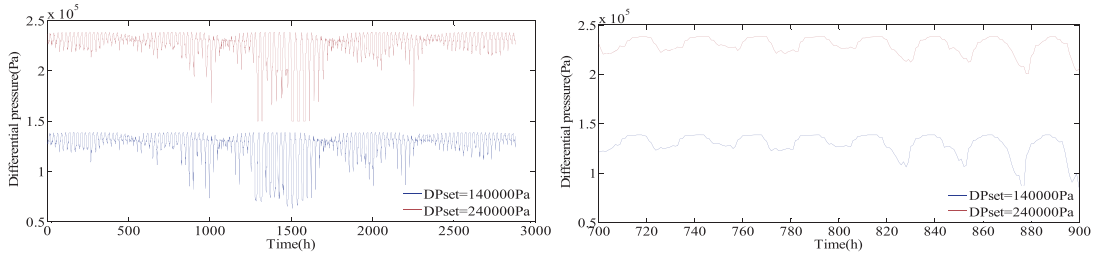


Fig. 14 Differential pressure of water loops under different DP_{set} values

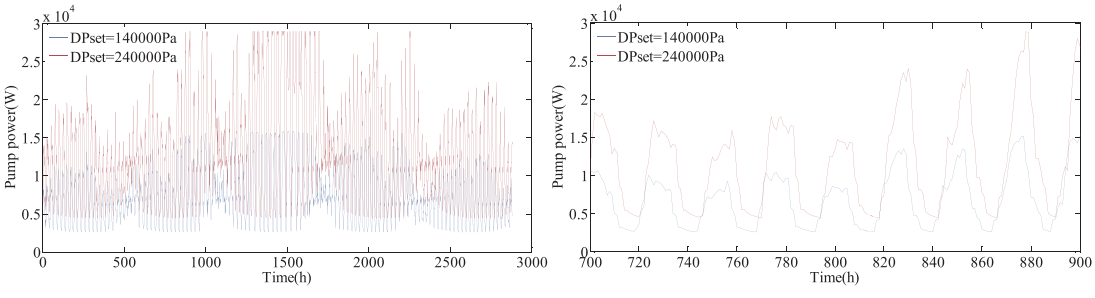
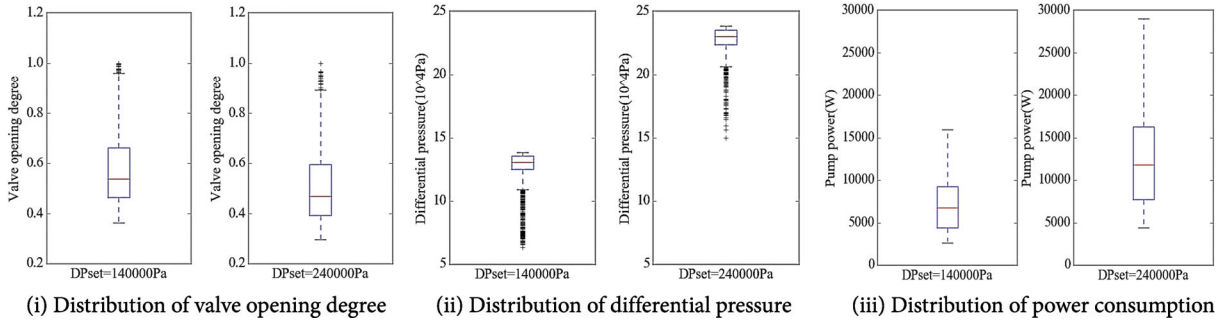


Fig. 15 Pump energy consumption under different DP_{set} values



(i) Distribution of valve opening degree (ii) Distribution of differential pressure (iii) Distribution of power consumption

Fig. 16 Distribution of observed variables under different DP_{set} values

Table 4 Statistical characteristics of the valve opening degree sequence

DP_{set}	Mean	Std	Min	25%	50%	75%	Max
140000 Pa	0.57	0.1484	0.36	0.47	0.54	0.66	1.00
240000 Pa	0.50	0.1429	0.30	0.40	0.47	0.5	1.00

Table 5 Statistical characteristics of the differential pressure sequence (unit: Pa)

DP_{set}	Mean	Std	Min	25%	50%	75%	Max
140000 Pa	1.27×10^5	14187	0.63×10^5	1.25×10^5	1.31×10^5	1.36×10^5	1.40×10^5
240000 Pa	2.26×10^5	14585	1.50×10^5	2.23×10^5	2.30×10^5	2.35×10^5	2.40×10^5

Table 6 Statistical characteristics of the power consumption sequence (unit: W)

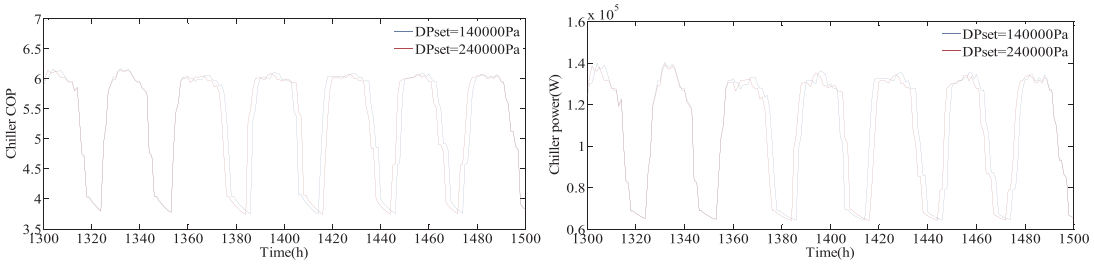
DP_{set}	Mean	Std	Min	25%	50%	75%	Max
140000 Pa	7.25×10^3	3482	2.57×10^3	4.43×10^3	6.78×10^3	9.27×10^3	1.59×10^4
240000 Pa	1.28×10^4	6476	4.42×10^3	7.77×10^3	1.18×10^4	1.63×10^4	2.90×10^4

the change of the building cooling load to some extent.

Since pumps are essential equipment of water loops, they will have an impact on other important components of the HVAC system during the operation. To further explore the influence degree, chiller characteristics under different DP_{set} values are shown in Figure 17. Corresponding variable distribution is shown in Figure 18. Statistical results are shown in Tables 7–8. It can be concluded that when DP_{set} changes from 1.4×10^5 Pa to 2.4×10^5 Pa, the average chiller COP changes by 0.6% and the average chiller power consumption changes by 1.16%. The impact is relatively small compared with the optimization goal, thus, the abovementioned agent structure can be utilized to improve sub-loop performance. More discussion can be seen in Section 5.

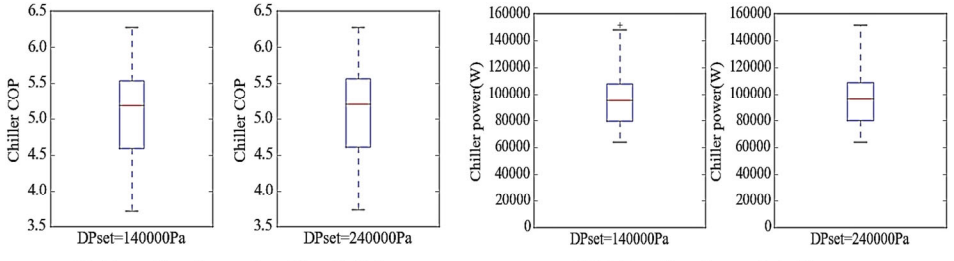
4.2 Traditional rule based control method

To evaluate the energy saving potential of the RLC strategy, a traditional rule-based control method named the variable differential pressure control (VPC) is introduced for comparison. In the VPC, sensors collect and send system operational data at fixed time intervals. When all valve opening degrees are less than 90% and this state lasts over 20 minutes, the DP_{set} will be reduced by the ΔP value. In contrast, if the temperature overshoots and that condition



(i) Chiller COP under different DP_{set} values (ii) Chiller power under different DP_{set} values

Fig. 17 Chiller parameters under different DP_{set} values



(i) Distribution of chiller COP (ii) Distribution of chiller power

Fig. 18 Distribution of chiller parameters under different DP_{set} values

Table 7 Statistical characteristics of the chiller COP sequence

DP_{set}	Mean	Std	Min	25%	50%	75%	Max
140000 Pa	5.04	0.6574	3.72	4.60	5.19	5.54	6.27
240000 Pa	5.07	0.6677	3.73	4.62	5.21	5.57	6.28

Table 8 Statistical characteristics of the chiller power sequence (unit: W)

DP_{set}	Mean	Std	Min	25%	50%	75%	Max
140000 Pa	9.48×10^4	19200	6.39×10^4	7.96×10^4	9.55×10^4	1.08×10^5	1.52×10^5
240000 Pa	9.59×10^4	19850	6.41×10^4	8.01×10^4	9.61×10^4	1.09×10^5	1.52×10^5

lasts for over 20 minutes, DP_{set} will be increased by the ΔP value. The flow chart is shown in Figure 19.

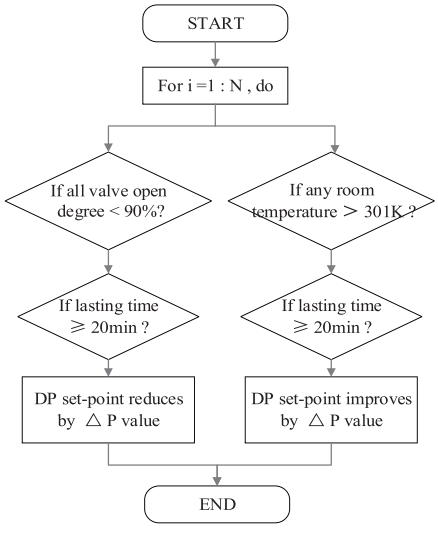


Fig. 19 VPC control logic

4.3 Simulation result

Before the simulation, the parameters were set as shown in Table 9.

Before beginning the cooling season simulation, three typical days at the end of July were selected to test the convergence. The heat flowing into the zone node is shown in Figure 20, and the corresponding simulation results are shown in Figure 21. In the VPC strategy, the DP_{set} for the next step is determined based on the current feedback information, and when the corresponding control signal is read into the system at the next step, a certain delay occurs due to the thermal inertia and load change, as shown by the dotted blue lines. This delay leads to signal repetition on a time scale, which is difficult to change because of the principle underlying the method. In contrast, in the RLC strategy, the relationship between the control signal and the subsequent influences is learned from end to end. The characteristics of intermediate links are considered as implicit experience in the algorithm; thus, the time delay problem can be avoided. The pump energy consumption under different strategies is shown in Figure 22. The DP_{set} of the RLC strategy is higher at night; however, the cooling load during this period is low; thus, the energy consumption difference of the two methods is actually small. A higher night DP_{set} allows the agent to reach an advantageous position faster during the daytime (DP_{set} can be changed only at

Table 9 Parameter settings

Dymola	Start time	12,960,000 s
	End time	23,328,000 s
	Simulation interval	3,600 s
	Monitoring interval	1,200 s
Python	Algorithm	Q-learning
	α (learning rate)	0.9
	ε	0.05
	γ (discount rate)	0.3
	Actions	{−40,000 Pa, 0, 40,000 Pa}
	Steps in an episode	2,880 (4×30×24)
	E_{max}	6,000 W
	E_0	2,000 W
	T_{limit}	300 K
	T_0	0.5 K
	α_1	1,000
	α_2	$100\times2^{Tin\text{-}Tlimt}$
	$DP_{set,min}$	80,000 Pa
	$DP_{set,max}$	280,000 Pa

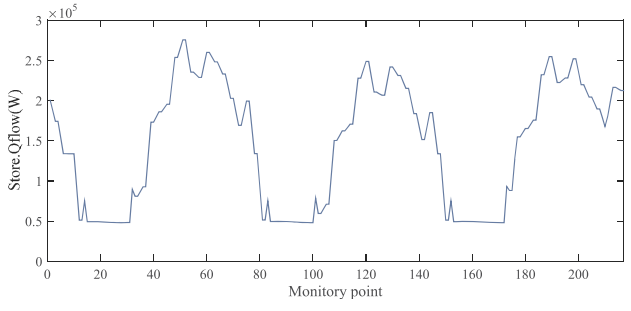


Fig. 20 Q-flow of independent stores during typical days

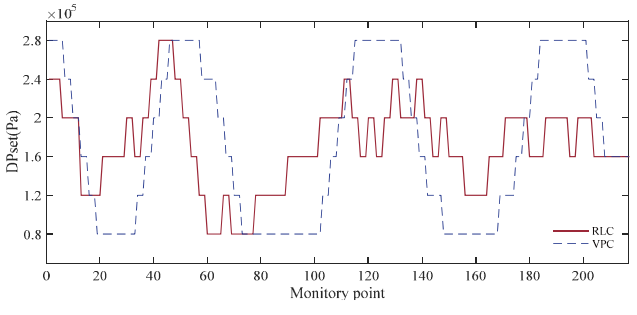


Fig. 21 DP_{set} of different strategies during typical days

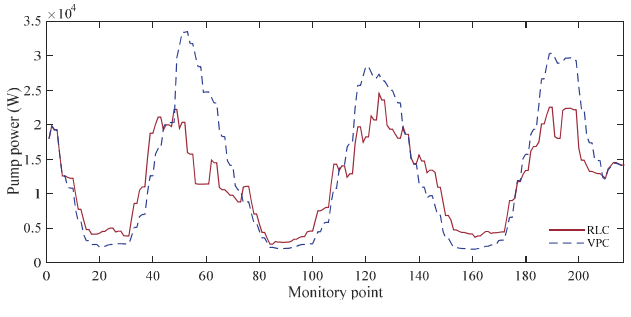


Fig. 22 Pump power of different strategies during typical days

fixed intervals), and this will have a greater impact on energy consumption when the cooling load is high. Eventually, the total energy consumption is effectively reduced. The overall effect, as demonstrated in Figure 23, is that both strategies work to effectively control the zone temperature within its required range.

Extending the timescale to the entire cooling season, according to the cooling load input of the system model, a training episode spans 2880 h in total. As noted in Table 9, the simulation interval is set to 1 h; thus, there are 2880 steps in each episode. In addition, the monitoring interval is set to 1200 s, which means that the software outputs a group of simulation results every 20 minutes. Figure 24 shows the DP_{set} distribution in different episodes. As the number of training sessions increases, the total number of DP_{set} increases in the lower value region {80000 Pa, 120000 Pa, 160000 Pa} and decreases in the higher value region {200000 Pa, 240000 Pa, 280000 Pa}. The figure also shows the DP_{set} distribution under the VPC strategy in a different color. In contrast to the DP_{set} distribution under the RLC strategy, higher values are obtained at both ends, while the values are relatively average in the middle, which has a relationship with the feedback mechanism of the VPC strategy.

Figure 25 shows the accumulated pump power consumption at monitoring points in different episodes, and the corresponding quantitative results are shown in Table 10. It can be estimated that the algorithm converges after 4 episodes, and the maximum energy savings reach 17.87%. Note that after the first episode, the RLC strategy

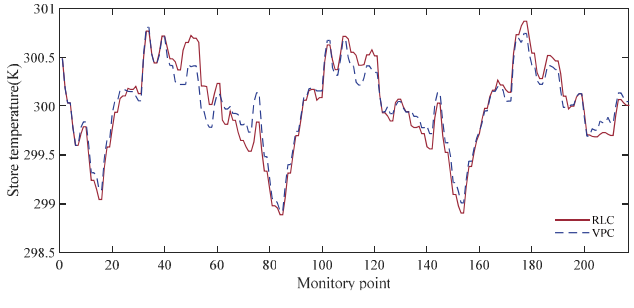


Fig. 23 Store air temperature during typical days

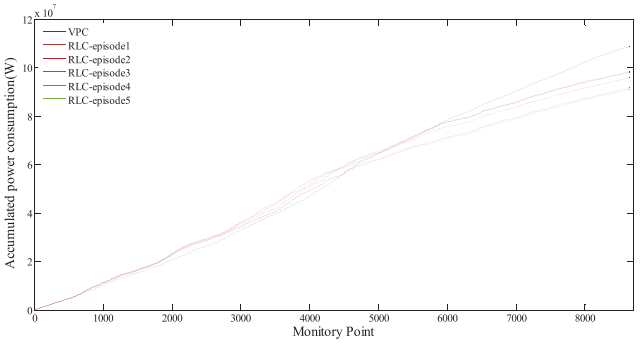


Fig. 25 Accumulated pump power consumption in different episodes

already shows some energy saving effect; thus, considering the existence of a "simulator-assisted" problem, here the attention focuses on a discussion of the first episode.

The DP_{set} of the RLC strategy in the first episode is shown in Figure 26. The variable is maintained at a low level most of the time, and its change trend is basically consistent with the change in the cooling load. In the VPC strategy, the control signal fluctuates periodically, and the amplitude and period of fluctuation vary under different cooling loads. The energy consumption under different strategies is shown in Figure 27.

To further explore the reason for energy savings under the RLC strategy, 489 different state pairs are obtained and analyzed after the first training episode. This number is less than 2,880, which indicates that some states appear more than once in the first episode and this can provide a chance to help agent learn experience better. The details are shown in Figure 28. The horizontal axis shows the number of times a certain state repeats in one episode, while the vertical axis reflects the total number of states sharing the same number of repetitions.

Finally, zone 3 is selected as a representative zone to validate the ability of the strategy to maintain indoor air temperature. Figure 29 shows that both strategies can maintain the indoor air temperature within its required range. The temperature distribution is also shown in Figure 30, and the corresponding statistical characteristics are presented in Table 11. The standard deviation of the temperature sequence

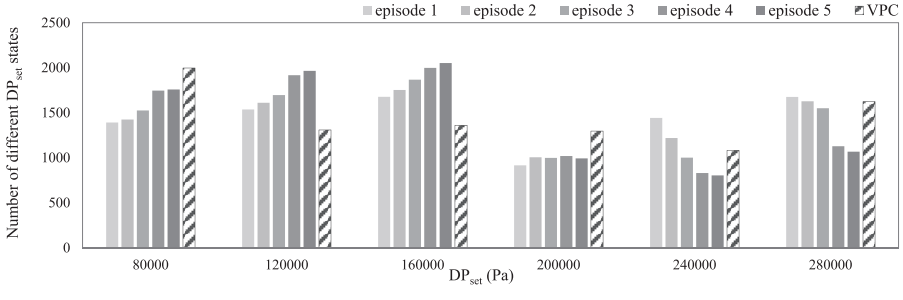
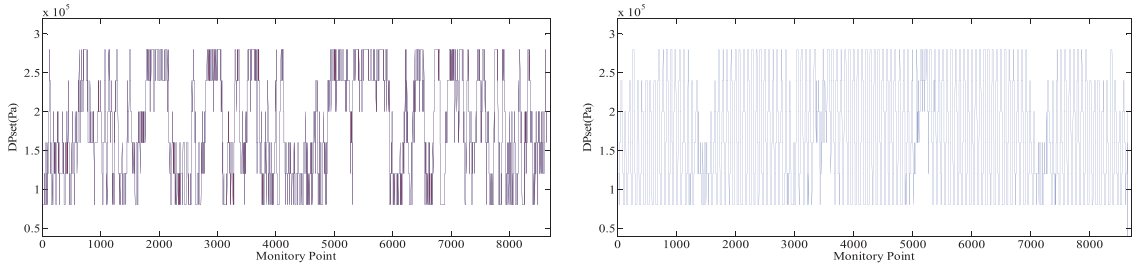


Fig. 24 DP_{set} distribution in different episodes

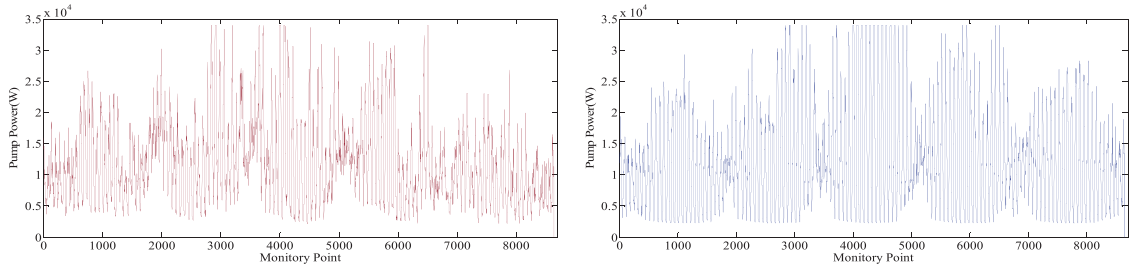
Table 10 Total energy consumption of the water pump in different episodes

Strategy	VPC	RLC				
		Episode 1	Episode 2	Episode 3	Episode 4	Episode 5
Total energy consumption	1.089×10^8 J	9.74×10^7 J	9.69×10^7 J	9.48×10^7 J	9.02×10^7 J	8.94×10^7 J
Potential energy saving rate	—	10.51%	10.98%	12.96%	17.17%	17.87%



(a) RLC-1st episode (b) VPC

Fig. 26 DP_{set} under different strategies



(a) RLC-1st episode (b) VPC

Fig. 27 Pump power consumption under different strategies

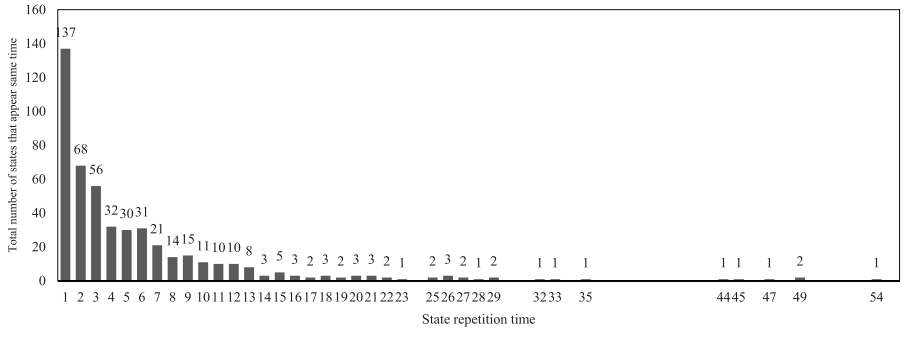
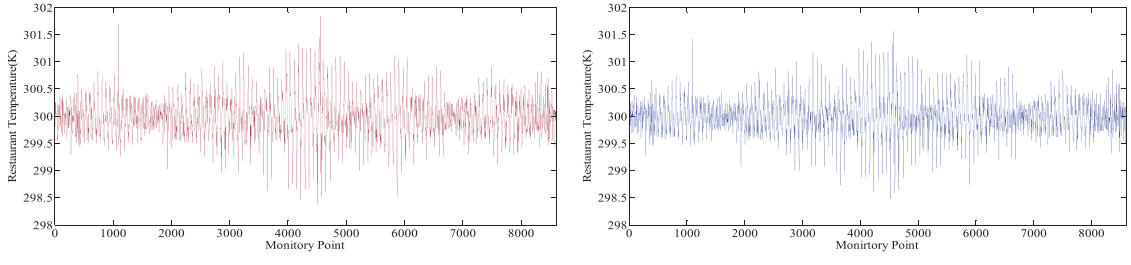


Fig. 28 Details of state repetition during the first episode



(a) RLC-1st episode (b) VPC

Fig. 29 Zone temperatures under different strategies

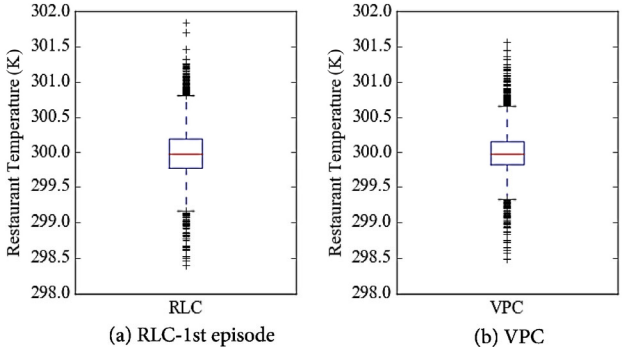


Fig. 30 Zone temperature distribution under different strategies

Table 11 Statistical characteristics of the temperature sequence (unit: K)

Strategy	Mean	Std	Min	25%	50%	75%	Max
RLC	300.00	0.3456	298. 39	299.78	299.99	300.20	301.84
VPC	300.00	0.2970	298.48	299.98	299.98	300.16	301.56

when employing the RLC strategy is slightly larger than that of the VPC strategy since the former just uses temperature as an implicit constraint.

From a practical point of view, to obtain a good performance, three conditions should be met. That is, the agent's state pair should have good optimization potential, the agent's state space should be reasonably divided, and the algorithm parameters should be properly adjusted. For air conditioning systems, most of the variables selected as states have certain periodicity characteristics (e.g., day, month or year), which makes it easier to accumulate experience that can be repeatedly utilized. More states is not always better, because no matter for which kind of the optimizer, irrelevant features will increase problem complexity and convergence difficulty. State selection should consider optimized objects' characteristics and should be determined after comparison. Second, when dividing the state space, the grid density should be reasonable; this avoids a poor optimization effect when the grid is too sparse or lengthy convergence times when the grid is overly dense. Third, especially in practical applications, improper exploration processes will lead the system to deviate from a steady state. It is recommended that algorithm parameters controlling exploration speed should be more conservative while factors related to the feedback reward should be more sensitive to encourage the agents to adapt faster to environmental changes.

5 Conclusion and discussion

The main conclusions obtained from this study are as follows:

(1) In a typical delta-P control air conditioning water system, the pump curves, control curves and pipe characteristic curves influence each other and work together to maintain system stability. When the DP setpoint is decreased, the pump frequency will decrease while the valve openness will increase simultaneously, and when the system is stable again, pump energy consumption will be effectively reduced.

(2) By defining two interacting objects, agents and the environment, and four key elements, state, action, policy and reward, a specific RL-based DP setpoint reset strategy is proposed. The agents collect information from the environment and select the best action to maximize the cumulative reward. In an actual BA system, the RLC is designed to connect with the central control platform through an open interface, allowing the original hierarchical control framework of the system to be preserved.

(3) In a simulated case study, an HVAC system with three AHUs was built in Dymola to validate the RLC and compare it with the traditional VPC strategy. The results indicate that RLC effectively avoids the time delay problem by regarding intermediate links as implicit experiences, and it sacrifices smaller profits to move quickly to the best position at important times. Finally, the RLC reduces the total pump power consumption by 10.5% after the first episode while maintaining the indoor air temperature within its defined range.

(4) The purpose of this paper is to explore a common methodology of constructing RL device agents of HVAC system. That is, pump agents are built according to pumps' characteristics and similarly, chiller agents are built according to chillers' characteristics. In this way, the whole HVAC system can be divided into limited similar separated modules. Thus, users can customize the optimization modules according to their needs, which can increase actual optimization flexibility. When optimizing the whole system, main parameters of bottom modules can be centrally unified optimized or decentralized partial optimized. This paper provides a method of building bottom agent modules and serves a foundation for further research of higher level inter-module optimization algorithm.

Some other points that need further consideration are as follows:

1) Generally, in the RL algorithm's gambling process, problems may occur when the agent randomly moves to strange states during its exploratory stage. These actions could cause system imbalances and influence user comfort levels. At this point, some expert rules should be added to narrow down the scope of exploration, or some estimation methods can be integrated to help agent judge the system situation and be familiar with the environment in a faster way (Zhou et al. 2020).

2) For many small and medium-sized systems in reality, there are always exist problems such as sensor damage or missing, and the valid data is limited. In this case, algorithm with simple logic will work better. For example, training process of DQN usually requires more data because of the intermediate NN or CNN layers; in contrast, the table of Q-learning algorithm shows more practical benefits, which is convenient for operators to explicitly judge the operating state of the system. However, for large new systems with complete sensors, the data dimension usually exceeds a certain range and Q-table algorithm will become inadequate. More complicated methods could be utilized to optimize system (Wei et al. 2017). However, the core of the RLC strategy remains unchanged, that is, interacting with the environment and obtaining feedback to determine the next move. As long as the MDP process and evaluation standard are clearly designed, the agent will continually learn by itself.
3) To further explore RLC's control performance on whole HVAC system, more control objects should be added, such as chillers and cooling towers. Based on conclusions obtained from this paper, the control framework can be designed in two ways: First, similar independent control modules are designed according to different equipment, meanwhile, the centralized controller will organize the modules in a unified way. Second, a more complicated agent with all necessary equipment's characteristic parameters is established as the same logic proposed in this paper. These two different control logics can be compared to explore a better way to utilized RLC on actual complicated systems. Besides, more disturbance signals can be added to the model in different stages to better validate RLC's robustness and accuracy.

Electronic Supplementary Material (ESM): supplementary material is available in the online version of this article at https://doi.org/10.1007/s12273-021-0808-5.

References

Barrett E, Linder S (2015). Autonomous HVAC control, a reinforcement learning approach. In: Bifet A, et al. (eds), Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2015. Lecture Notes in Computer Science, vol 9286. Cham, Switzerland.

Chen Y, Norford LK, Samuelson HW, et al. (2018). Optimal control of HVAC and window systems for natural ventilation through reinforcement learning. *Energy and Buildings*, 169: 195–205.

Dayan P, Niv Y (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, 18: 185–196.

Gao D, Wang S, Shan K (2016). In-situ implementation and evaluation of an online robust pump speed control strategy for avoiding low delta-T syndrome in complex chilled water systems of high-rise buildings. *Applied Energy*, 171: 541–554.

Han M, May R, Zhang X, et al. (2019). A review of reinforcement learning methodologies for controlling occupant comfort in buildings. *Sustainable Cities and Society*, 51: 101748.

Hou J, Xu P, Lu X, et al. (2018). Implementation of expansion planning in existing district energy system: A case study in China. *Applied Energy*, 211: 269–281.

Ji Y, Peng S, Geng L, et al. (2009). Pressure loop control of pump and valve combined EHA based on FFIM. In: Proceedings of the *9th International Conference on Electronic Measurement and Instruments*, Beijing, China.

Jin X, Du Z, Xiao X (2007). Energy evaluation of optimal control strategies for central VWV chiller systems. *Applied Thermal Engineering*, 27: 934–941.

Lewis FL, Vrabie D, Vamvoudakis KG (2012). Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers. *IEEE Control Systems Magazine*, 32(6): 76–105.

Li W, Xu P, Lu X, et al. (2016). Electricity demand response in China: Status, feasible market schemes and pilots. *Energy*, 114: 981–994.

Liu S, Henze GP (2006a). Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory: Part 1. Theoretical foundation. *Energy and Buildings*, 38: 142–147.

Liu S, Henze GP (2006b). Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory: Part 2: Results and analysis. *Energy and Buildings*, 38: 148–161.

Liu X, Liu J, Lu J, et al. (2012). Research on operating characteristics of direct-return chilled water system controlled by variable temperature difference. *Energy*, 40: 236–249.

Ma Z, Wang S (2009). Energy efficient control of variable speed pumps in complex building central air-conditioning systems. *Energy and Buildings*, 41: 197–205.

Mehlhase A (2012). A python package for simulating variable-structure models with dymola. *IFAC Proceedings Volumes*, 45: 1081–1086.

Pérez-Lombard L, Ortiz J, Pout C (2008). A review on buildings energy consumption information. *Energy and Buildings*, 40: 394–398.

Recht B (2019). A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2: 253–279.

Sutton R , Barto A (2018). Reinforcement Learning:An Introduction. Cambridge, MA, USA: MIT Press.

Wang Z, Hong T (2020). Reinforcement learning for building controls: The opportunities and challenges. *Applied Energy*, 269: 115036.

Wei T, Wang Y, Zhu Q (2017). Deep Reinforcement Learning for Building HVAC Control. In: Proceedings of the 54th Annual Design Automation Conference, Austin, TX, USA.

Yang L, Nagy Z, Goffin P, Schlueter A (2015). Reinforcement learning for optimal control of low exergy buildings. *Applied Energy*, 156: 577–586.

Zhao T, Ma L, Zhang J (2016). An optimal differential pressure reset strategy based on the most unfavorable thermodynamic loop on-line identification for a variable water flow air conditioning system. *Energy and Buildings*, 110: 257–268.

Zhou Y, Chen J, Yu ZJ, et al. (2020). A novel model based on multi-grained cascade forests with wavelet denoising for indoor occupancy estimation. *Building and Environment*, 167: 106461.