BUILD SIMUL
https://doi.org/10.1007/s12273-020-0602-9

Study on the application of reinforcement learning in the operation optimization of HVAC system

Xiaolei Yuan[1], Yiqun Pan[1] (✉), Jianrong Yang[2], Weitong Wang[3], Zhizhong Huang[4]

1. School of Mechanical Engineering, Tongji University, 4800 Cao'an Road, Shanghai 201804, China
2. Shanghai Research Institute of Building Sciences, Shanghai, China
3. Kuaishou Co. Ltd., Beijing, China
4. Sino-German College of Applied Sciences, Tongji University, Shanghai 201804, China

Abstract

Supervisory control can be used to optimize the HVAC system operation and achieve building energy conservation, while reinforcement learning (RL) is considered as a promising model-free supervisory control method. In this paper, we apply RL algorithm to the operation optimization of air-conditioning (AC) system and propose an innovative RL-based model-free control strategy combining rule-based and RL-based control algorithm as well as complete application process. We use a variable air volume (VAV) air-conditioning system for a single-storey office building as a case study to validate the optimization performance of the RL-based controller. We select control strategies with the rule-based control controller (RBC) and proportional-integral-derivative (PID) controller respectively as the reference cases. The results show that, for the air supply of single zone, the RL controller performs the best in terms of both non-comfortable time and energy costs of AC system after one-year exploration learning. The total energy consumption of AC system reduced by 7.7% and 4.7%, respectively compared with RBC and PID strategies. For the air supply of multi-zone, the performance of RL controller begins to outperform the reference strategies after two-year exploration learning and two-year buffer stage. From the seventh year on, RL controller performs much better in terms of both non-comfortable time and operating costs of AC system, while the operating cost of AC system is reduced by 2.7% to 4.6% compared with the reference strategies. In addition, RL controller is more suitable for small-scale operation optimization problems.

Keywords

reinforcement learning,
HVAC system,
operation optimization,
control strategy,
VAV system,
energy saving

Article History

Received: 30 August 2019
Revised: 19 November 2019
Accepted: 17 December 2019

1 Introduction

Building sectors are responsible for around 40% of total primary energy consumption and approximate 30% of related total CO_2 emission in the worldwide (Costa et al. 2013). Building energy conservation methods should be taken into consideration to reduce the CO_2 related environmental impacts and achieve sustainable development (Zhao and Magoulès 2012). The energy consumption of heating ventilation and air-conditioning (HVAC) systems takes up near 40% of total building energy use (DOE 2011; Dong et al. 2014). Thus, energy-efficient HVAC system will contribute to significant energy saving in the building sectors and environmental-friendly development (Niu et al. 2018). The application of supervisory control (optimal control) in the operation optimization of HVAC system is considered as one of promising building energy conservation methods (Wang and Ma 2008). The basic goal of the supervisory and optimal control is to minimize the energy consumption or operating costs on the basis of satisfied indoor comfort level and healthy environment in occupant areas (Jung and Jazizadeh 2019). The essence of the supervisory and optimal control in HVAC system is the building operation optimization. For the adaptation of changeable outdoor weather and indoor loads, the dynamic adjustments of setting values and operating rules in the building system to

E-mail: yiqunpan@tongji.edu.cn

improve building energy efficiency has been the research focus of building operation optimization (Dong et al. 2018; Gunay et al. 2019).

The supervisory control methods applied in HVAC systems can be divided into 4 categories: model-based method, hybrid method, performance map-based method and model-free method (Wang and Ma 2008). Many researchers have studied the model-based supervisory control method in HVAC systems. House and Smith (1995) used physical model-based supervisory control methods to optimize the air-conditioning (AC) system. Li et al. (2010) used gray-box model-based method to develop and validate a dynamic zone model to achieve energy saving and indoor environment improvement. In addition, Curtiss et al. (1994) developed artificial neural networks (ANNs) black-box model-based supervisory control to minimize the energy consumption of HVAC system. Although lots of studies on the model-based supervisory control method have been done, some drawbacks still exist. It is time-consuming and labor-consuming to establish and validate the model based on the model-based method, hindering its application in practical engineering projects (Killian and Kozek 2016). In addition, the quality of the control strategy depends heavily on the quality of the model. Once the model deviates from the real situation of the building or HVAC system, the quality of the control strategy is unconvincing (Široký et al. 2011). Besides, the calculation of model predictive control (MPC) is complex and difficult, having much higher requirements for hardware equipment (Goyal et al. 2013).

Compared to model-based supervisory control method, model-free method can directly obtain the control strategy without establishing the mathematical model of building HVAC system (Baldi et al. 2015). Model-free supervisory control mainly includes expert system-based and reinforcement learning (RL) based control methods (Ling and Dexter 1994). According to Ling and Dexter (1994), expert system-based control method has the characteristics of simple structure and strong stability, but its parameters setting depends heavily on the prior engineering experience of engineers, rather than on the optimization algorithm. Thus, expert system-based control method has weak dynamic adjustment.

According to Mason and Grijalva (2019), RL has been widely used in practical engineering optimization and control areas. Compared to the characteristics of model-based control method, RL-based model-free strategy is a data-driven control method, using feedback information to update the control strategy after constantly trying and minimize the dependence on prior knowledge (Doll et al. 2016; Mbuwir et al. 2017; Russek et al. 2017; Halperin et al. 2019). Many researchers have applied RL control strategy in the operation optimization of building HVAC system. Liu and Henze (2006) used the RL to optimize the operation of active and passive building thermal storage inventory. They found that classic Q-Learning algorithm has the drawback of inefficiency in high-dimensional spatial learning. Barrett and Linder (2015) proposed an auto-control method for HAVC system based on RL. This auto-control method can realize the intelligent temperature control in the controlled areas by learning the characteristics of HVAC equipment and occupant habits. Costanzo et al. (2016) applied RL controller to building demand response, and concluded that the application of RL control strategy can achieve a 90% of the mathematical optimum solution. Urieli and Stone (2013) and Ruelens et al. (2015) applied the model-based and model-free RL algorithms to the HVAC system with heat pump, respectively. The results show that the application of both two algorithms could improve the operating efficiency of heat pump, achieving significant energy savings. Li and Xia (2015) proposed multi-scale RL to accelerate the process of solving optimal control strategies. They concluded that the multi-scale RL control strategy has advantages in energy saving and comfort. In addition, Wei et al. (2017) proposed the deep RL-based control method of HVAC system and validate the scalability of deep RL controller. They found that deep RL controller suffers from the problem of too long training time.

Above all, the related researches on RL mainly focus on the validation of the performance of RL in different operation optimization scenarios of HVAC systems. However, the systematic combing of the application of RL controller is not well addressed as well as the development process of RL supervisory controller. In addition, that the reliability reinforcement of RL controller and the decrease of time required for controller learning process are still the main research direction at present and should be further studied. In this paper, RL algorithm is applied to the operation optimization of AC system, while an innovative RL-based model-free control strategy combining rule-based and RL-based control methods is proposed as well as complete application process. The new RL-based controller is applied in a variable air volume (VAV) AC system for a single-storey office building as a case study to validate its performance.

2 Reinforcement learning theory

2.1 Reinforcement learning introduction

Reinforcement learning (RL) is a special and adaptive machine learning method with environmental feedback as input, while its main principle is to interact with the environment and optimize the decision-making based on the feedback signal of evaluation (Gao et al. 2004; Jaafra et al. 2019). Figure 1 shows the sketch map of the RL. The

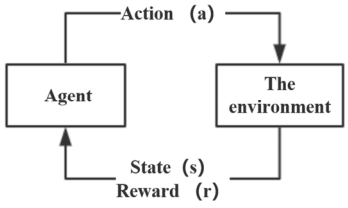

Fig. 1 Sketch map of RL

mathematical description of the RL problems is Markov Decision Process (MDP), including state space S, action space A, transition function P and reward function R (Costanzo et al. 2016). At each decision moment t, the agent selects the execution action $a_t \in A$. After one time step Δt, the environment E is converted from state s_t to s_{t+1}, while the agent calculates the reward r_t during this time step according to the reward function at time $t + 1$. The transition function describes the execution of action $a_t \in A$ when the environment state of the agent is s_t at the time t. In addition, the transition function also expresses the possibility that the environment state of the agent is converted to s_{t+1} when the random disturbance is w_t. MDP has the Markov property, that is, the reward r_t and environment state s_{t+1} at the next moment are only relevant to the current state s_t as well as the current action a_t (Gao et al. 2004; Han et al. 2019). The goal of the RL is to obtain a strategy function $a = \pi(s)$, mapping the relationship between state and control action, and maximize the cumulative reward by executing the strategy functions (Han et al. 2019). Under this circumstance, this strategy function is the optimal control strategy. RL method is goal-oriented and good at solving optimization problems of decision chains under unknown environment (You et al. 2019).

2.2 Model-free learning algorithms

In the RL tasks, the transition probability and the reward function of the environment are often unknown or hard to be obtained. A model-free learning algorithm need not build up the environmental model, while Q-Learning algorithm is one of the model-free learning algorithms (Halperin 2019).

Q-Learning algorithm is a RL method for solving Markov decision problems with incomplete information (Cheng et al. 2016). The object of this algorithm is the value function of state-action pairs, namely Q-value function, expressed by $Q(s, a)$. The Q-value function represents the cumulative reward awarded to the system by executing action a under state s. Tabular Q-Learning refers to an algorithm storing Q-value of finite state-action pairs in a table (Mason and Grijalva 2019; You et al. 2019). Q-value is random in the initial stage, while the samples in the form of tuples (s, a, r, s') are collected during the continuous interaction between agent and the environment. Equation (1) is used to update the Q-value in the table.

$$\Delta Q(s,a) = \alpha\left(R(s) + \gamma \max_{a' \in A(s')} Q(s',a') - Q(s,a)\right) \tag{1}$$
$$Q(s,a) \leftarrow Q(s,a) + \Delta Q(s,a)$$

where $A(s)$ is a set of actions that can be taken in state s; s' is the next state after executing action a under state s.

The reward discount factor $\gamma \in [0, 1]$ indicates the influence degree of current actions on future rewards (Mason and Grijalva 2019; Han et al 2019). There are many uncertainties during the operation optimization of HVAC system, such as the decrease of prediction accuracy of weather conditions with the increase of the forecast time from the current time. The uncertainty will accumulate in the time dimension, so the γ discount cumulative rewards are adopted as the long-term cumulative reward (Jaafra et al. 2019). When γ equals to 0, the agent only takes immediate rewards into consideration. However, when γ equals to 1, the agent considers both equal importance to long-term rewards as to immediate rewards. The learning rate $\alpha \in (0, 1)$ is the updating speed of Q-value. Larger learning rate can improve the convergence speed of the algorithm, while smaller learning rate can improve the stability of the algorithm (Watkins and Dayan 1992). Temporal-Difference term denotes the difference between the real value $r + \gamma \max_{a' \in A(s')} Q(s',a')$ and the estimated value $Q(s, a)$ (Sutton and Barto 1998; Brémaud 1999). However, the real value also includes the estimated value $Q(s',a')$ of the next state-action pair. Q-Learning algorithm makes the estimated value closer to the real value by continuously collecting samples and updating the Q-value (Watkins and Dayan 1992).

In the learning process, agents acquire experience constantly, and then they get strategies with certain performance after learning for a period of time. In addition, the parts excluding the existing strategies need to be explored to test the possibility of strategies improvement (Defazio and Graepel 2014). According to Costanzo et al. (2016), ϵ-greedy exploration is a widely used exploration method for balancing exploration and utilization. The agent selects an action with the random possibility of ϵ at each decision-making moment, and selects the action with the highest Q-value with the possibility of $1 - \epsilon$ (Defazio and Graepel 2014). The value of ϵ defines the proportion of agent exploration. After enough exploration, the agent stops exploring and implement the optimal control strategy according to the greedy control criterion (Costanzo et al. 2016).

2.3 Value function approximation

Tabular Q-value algorithm is a MDP for discrete state and

action space; however, tabular Q-Learning will come across curse of dimensionality because many optimization decision-making problems in practical engineering have large-scale or continuous state and action space (Nguyen et al. 2019). In order to solve this problem, generalization methods (e.g. parametric function approximation and non-parametric function approximation) are usually used for Q-value function approximation in RL. Artificial neural networks (ANNs) is one of typical approximation methods of parametric function, while regression Tree is one of typical approximation methods of non-parametric function (Baird 1995). Ernst et al. (2005) studied the performance of ensemble learning method in estimating value function, and then found the extreme random tree performed the best, which was suitable for estimating the Q-value function. In the use of extreme random tree, the number of trees and the minimum number of samples for spilt nodes are required to control the model scale.

2.4 Batch RL algorithm

Q-Learning algorithm is an online algorithm. Once a sample is collected, the online algorithm immediately uses this sample to update the Q-value function (Ernst et al. 2005). However, two problems exist in updating the Q-value function by using Q-Learning algorithm. On the one hand, the updating of sequential samples may cover the updating of Q-value of previous samples to some extent. On the other hand, each sample is used only once in the on-line algorithm causing low learning efficiency (van Hasselt 2010). However, these two problems of on-line algorithm can be solved by adopting batch RL algorithm (Lange et al. 2012). According to Huang (2017), the basic principle of batch RL is experience replay. The experience here is a sample in the form of quaternion (s, a, r, s'), while the method of repeated use of experience is experience replay. Compared with the online algorithm, batch RL algorithm has higher data utilization efficiency and stability.

In this paper, fitted Q-iteration algorithm is selected as the batch RL algorithm to solve the operation optimization problem of building HVAC system. The fitted Q-iteration algorithm is a common batch RL method proposed by Ernst et al. (2005), which can improve data utilization efficiency and enhance algorithm stability by value function approximator and experience repay. The experience set D is built up by Eq. (2). The algorithm continually iterates and establishes the training data for Q-value function estimation, while the input data of training data are all state-action pairs (s_n, a_n) in experience set D and the output data are calculated according to Eq. (3).

$$D = \left\{\left(s_n, a_n, r_n, s_n'\right)\right\}_{n=1}^{\#D} \tag{2}$$

$$Q_M\left(s_n, a_n\right) = \left(r_n + \gamma \max_{a_n' \in A(s_n')} \hat{Q}_{M-1}\left(s_n', a_n'\right)\right) \tag{3}$$

where $\hat{Q}_{M-1}$ represents the value function approximator obtained by the algorithm in the last iteration.

Equation (3) demonstrates that the Q-value of each state-action pair is obtained by the sum of the state-action corresponding rewards and the optimal Q-value of next state calculated by the value function approximator obtained from the last iteration. Table 1 shows the detailed fitted Q-iteration algorithm process.

Table 1 Fitted Q-iteration algorithm process

1. Initialize Q-valued function approximator (QVFA);
2. Set the update condition of QVFA;
3. Obtain the current world state s;
4. Repeat the following process:
   a) Calculate Q-values of different actions under state s;
   b) Select the action a based on the exploratory strategy, and keep action a on until the next decision-making moment;
   c) Get the current time state s' and the reward r within the control time step, and store the (s, a, r, s') to the experience set;
   d) Judge and determine if the Q-value function approximator needs to be updated;
      i. Construct data set for updating QVFA by using experience set $D = \left\{\left(s_n, a_n, r_n, s_n'\right)\right\}_{n=1}^{\#D}$;
      ii. Calculate the corresponding Q_n of each state-action pair according to Eq. (1);
      iii. Update the QVFA by the data set $\left\{\left(s_n, a_n, Q_n\right)\right\}_{n=1}^{\#D}$.
   e) Assign s' to s.

3 Application of RL algorithm in the operation optimization of HVAC system

Many researches have focused on the appliction of RL in the HVAC syetm operation optimization. The operation methods of HVAC system entirely based on RL algorithm suffer from the problems of long learning time and poor reliability (Wei et al. 2017). Figure 2 shows the schematic map of the rule-assisted RL application in the operation optimization of HVAC system in this study. The simulation model of the building HVAC system is not needed, while

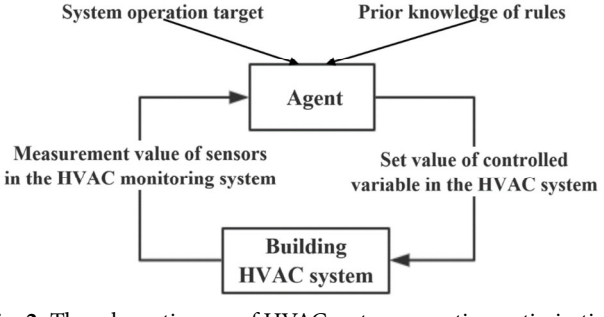

Fig. 2 The schematic map of HVAC system operation optimization based on rule-assisted RL

rule-based control method and RL algorithm is combined. The rule-based control method transforms engineering experience into the rule of conditional judgement statement form, and then determines the reasonable interval of system control variable under various working conditions. Meanwhile, RL algorithm collects and analyzes the operational data feedback from the HVAC monitoring system and continuously improves the control strategy, and then determines the optimal set value of the system control variables under various working conditions.

Rule-based control method can be effectively combined with RL-based control method. Rules can be used to optimize the operation of HVAC systems based on RL in two stages. In the initial stage of HVAC system operation, the data can be collected according to the rule-based control method for the initial RL controller so that the controller can quickly get a more reasonable control strategy and be improved. In the exploration stage after initializing the RL controller, the exploration space can be reduced by establishing rules, so as to ensure the agent avoids damaging the HVAC equipment and meet the comfort requirements of the controlled areas of the building as far as possible at the same time.

The combination of rule-based control method and RL-based control method can utilize their respective advantages. For one side, the engineering experiences can be introduced by the rules, reducing the exploration number required by RL controller and ensuring the reliability of the learning methods. For another side, the dynamic adjustment of rule-based control method can be enhanced by RL algorithm. This application process of combined control method can be divided into four stages in chronological order: preparation stage, initial stage, exploration stage and operation stage. Table 2 shows the complete flow chart of HVAC system operation method based on the combination of rule-based control method and RL algorithm.

Table 2 Operation process of HVAC system based on rule-based control method and RL algorithm

Preparation stage:

1) Model the operation optimization of HVAC system as a MDP;
2) Set the rule-based control method used in the initial stage, defining rules to reduce exploration space;
3) Determine the exploration strategies and the duration of the exploration stage, select the value function approximator, and set the discount factor and the update condition of the value function approximator.

Initial stage:

4) Run the Rule-based control HVAC system and collect initial samples.

Exploration stage:

5) Initialize value function approximator using data collected in initial stage;
6) Get the current state s;
7) Repeat the following processes:
   a) Calculate Q-value of different actions under state s;
   b) Select action a based on the exploration strategy and execute action a until the next decision-making moment;
   c) Get the current state s' and the reward r, and store (s, a, r, s') into the experience set;
   d) Judge and determine whether the Q-value function approximator needs to be updated; If so, update the value function approximator using the data in the experience set; or if not, execute the next step;
   e) Assign s' to s.

Operation stage:

8) Get the current state s;
9) Repeat the following processes:
   a) Calculate Q-value of different actions under state s;
   b) Select action a based on the exploration strategy and execute action $a$until the next decision-making moment;
   c) Get the current state s' and the reward r, and store (s, a, r, s') into the experience set;
   d) Judge and determine whether the Q-value function approximator needs to be updated; If so, update the value function approximator using the data in the experience set; or if not, execute the next step;
   e) Assign s' to s.

4 Application of RL controller in VAV system: a case study

4.1 Background

RL controller is applied in a single-storey office building with total AC area of 475 m^2 and with a VAV system as a case study. RL controller is responsible for adjusting the air supply volume and optimizing the set value of the air supply volume in the controlled area. The goal of the application of RL controller is to reduce the operating cost of the AC system as much as possible while meeting the indoor comfort requirement.

Regulating the air supply volume in the controlled area belongs to the typical building environment control. Killian and Kozek (2016) compared three existing control methods: proportional-integral-derivative (PID) control, PID control with external temperature compensation (PIDc) and model predictive control (MPC), and then concluded that both application of PID and PIDc cannot meet requirements of thermal conditions over a period of time. MPC can use the forecasted future outdoor weather changes and be combined with building AC system model to determine the current optimal control strategy by rolling optimization. Although MPC seems better than PID control in terms of control strategy, an accurate and efficient simulation model in the MPC is a prerequisite but hard to obtain in the building AC system. Thus, the focus of this VAV case is on achieving effective building environment control by using measurable

data, predicting the future changes of outdoor weather and strengthening learning algorithm without the model.

4.2 Preparation stage

According to the operation process of AC system based on rule-based control method and RL algorithm in Table 2, the first step in the preparation stage is to establish the MDP model for the operation optimization. The influencing factors on the operating costs of the AC system in the VAV case is firstly analyzed in Fig. 3. On this basis, the action is defined, namely the controlled variable, as the set value of the air supply volume in the controlled area in this case. The minimum air supply volume is determined as the fresh air volume needed by indoor personnel, while the maximum air supply volume is determined as six times of indoor ventilation. In addition, the action interval is divided into four levels $\{a_1, a_2, a_3, a_4\}$ with control time step of 0.2 h.

Then, the time, outdoor weather conditions and temperature of the controlled area are selected as the state of the VAV case. Time is used to reflect time-related information such as indoor thermal disturbance and electricity tariff in different time. Outdoor weather conditions are determined by the current outdoor temperature, current outdoor solar radiation and predicted temperature changes in the next hour. In addition, the temperature of controlled area is divided into two parts, including indoor air temperature and wall temperature. The wall temperature is used to reflect the building thermal storage, but it cannot be measured in practical project. Thus, the difference between the average indoor air temperature of the past four time steps and the current indoor air temperature is used in this case to approximate the building heat storage situation. In addition, the controller can also be provided with more complete information when only one dimension of the state value is added.

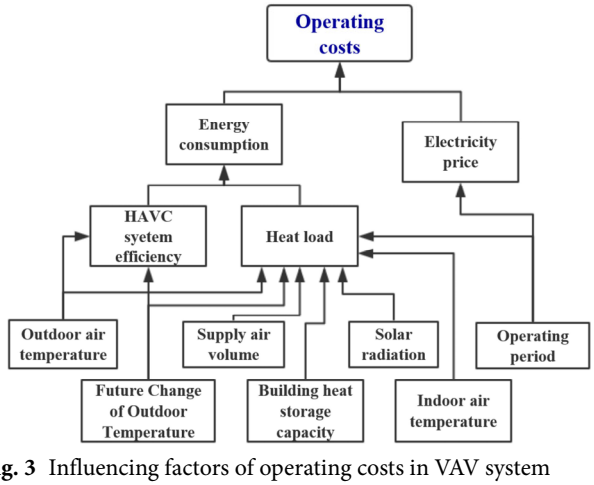

Fig. 3 Influencing factors of operating costs in VAV system

$$r_{t-1} = -\text{cost}(s_{t-1}, a_{t-1}) + \text{penalty}(T_t) \tag{4}$$

$$\text{penalty}(T_t) = \begin{cases} -\beta(T_t - T_{\text{lower bound}})^2 + b & \text{if } T_t < T_{\text{lower bound}} \\ -\beta(T_t - T_{\text{upper bound}})^2 + b & \text{if } T_t > T_{\text{upper bound}} \\ 0 & \text{if } T_{\text{lower bound}} \le T_t \le T_{\text{upper bound}} \end{cases} \tag{5}$$

At last, the reward function is defined in Eqs. (4) and (5) and divided into two parts. The first part is to control the energy consumption in the time step, which is the sum of the energy consumption of chillers, chiller water pumps, cooling water pumps, cooling tower and fans. Table 3 shows the electricity tariff in different time. The other part is the penalty when the indoor temperature excessed the comfort range. According to the Chinese design code of heating, ventilation and air conditioning for civil building (MOHURD 2012), the indoor temperature in Winter should be kept between 24 °C and 28 °C; Thus, 24 °C and 28 °C are selected as the lower and upper limit temperature, respectively.

The second step in the preparation stage is to set the rule-based control method used in the initial stage and define rules to reduce the exploration space. In the VAV case, the operation mode of start-stop control is adopted in the initial stage. Two rules are defined to ensure that system operation meets the comfort constraints. When the temperature in the controlled exceeds 27.5 °C, the supply air volume is set to the maximum and maintained to the next control step to avoid further temperature rising. In addition, when the indoor air temperature is less than 24.5 °C, the air supply volume is set to the minimum and maintained to the next control step to reduce energy consumption. These two rules are used in both the exploration and operation stages, aiming to accelerate the learning process of RL, avoiding meaningless explorations and improving the reliability of the system operation.

The third step in the preparation stage includes determining the exploration strategies and the duration of the exploration stage, selecting the value function approximator, setting the discount factor and updating the value function approximator conditions. The ϵ-greedy selection action is

Table 3 Electricity price information during on-peak and off-peak time

Period	Electricity price (RMB / kWh)	Time
Off-peak	0.28	22.00 – 6.00
On-peak	1.17	8.00 – 11.00; 13.00 – 15.00; 18.00 – 21.00
Other period	0.72	6.00 – 8.00; 11.00 – 13.00; 15.00 – 18.00; 21.00 – 22.00

selected as exploration strategy, expressed as Eq. (6). It is possible to perform more explorations at the beginning of the exploration stage and collect samples of different state-action pairs. Using the ϵ-greedy selection action in the operation of the first year is set as the exploration stage, while the AC system will be transferred to the operation stage from the second year onwards.

$$\epsilon = \max\left(\epsilon_0 - d \times \Delta\epsilon, 0.1\right) \tag{6}$$

where ϵ_0 is 0.5; $\Delta\epsilon$ is 0.02; d is operating days in the exploration stage.

Neural network is used as Q-value function approximator. There are 3 hidden layers in total, and the number of neurons in each hidden layer is 32, 64 and 32, respectively. The first two hidden layers use the rectified linear unit (ReLU) as the activation function, while the last hidden layer uses a linear unit (Linear) as the activation function.

The number of micro-batch data is set to 500. The discount factor γ is set to 0.5, while the value function approximator is updated once per day. When the indoor temperature violates the comfort constraint condition, it means that the control effect of the enhanced learning controller is still ideal. Thus, once the room temperature violates the comfort constraint, the value function approximator would be updated immediately.

4.3 Simulation model

The performance of RL controller is tested by Transient System Simulation (TRNSYS) Program software in this case. Modularization idea is adopted in the TRNSYS to establish simulation model, while each module represents a device or process. In the process of building TRNSYS system model, users only need to comb the input and output of each component, and then connect the components correctly to form a complete system and set the parameters of each components reasonably. The building model in the TRNSYS can reflect the basic mechanism of building thermal response, thus TRNSYS software is used in this case to analyze the system operation characteristics. In addition, Python language is also used to write RL controller. TRNSYS is responsible for building and AC system simulation, while co-simulations between TRNSYS and Matrix Laboratory (MATLAB) are conducted by using Internet socket through user datagram protocol (UDP). Figure 4 shows the co-simulation diagram.

Table 4 shows the parameters set of the components in the TRNSYS, while Fig. 5 shows the schedule setting for the occupant, lighting and equipment. The cooling season is from 1 June to 20 September. In addition, the simulation model of building AC system in TRNSYS is shown in Fig. 6. In this case, the meteorological parameters are simulated by the typical meteorological year (TMY) data of Shanghai, China. The predicted temperature value is calculated by adding the non-normality random error $N(0, 0.2)$ to the real temperature.

4.4 The comparison of control effect

Figure 7 shows the temperature variation in controlled area

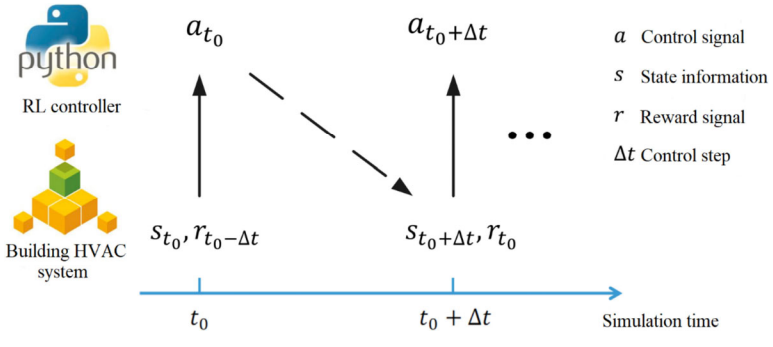

Fig. 4 The co-simulation diagrams

Table 4 The parameter of the components in the TRNSYS

Parameter		Value	Parameter	Value
U-value	External wall	0.888 W/(m²·K)	Building operating time	8.00–18.00
	Roof	0.638 W/(m²·K)	Set indoor air supply temperature	15 °C
	External windows	2.73 W/(m²·K)	Per capita fresh air volume	30 m³/(h·person)
Density	Personnel	0.25 person/m²	Centrifugal refrigeration unit COP	5
	Equipment	20 W/m²	Chilled water pump	Variable frequency
	Lighting	11W/m²	Cooling water pump	Constant frequency

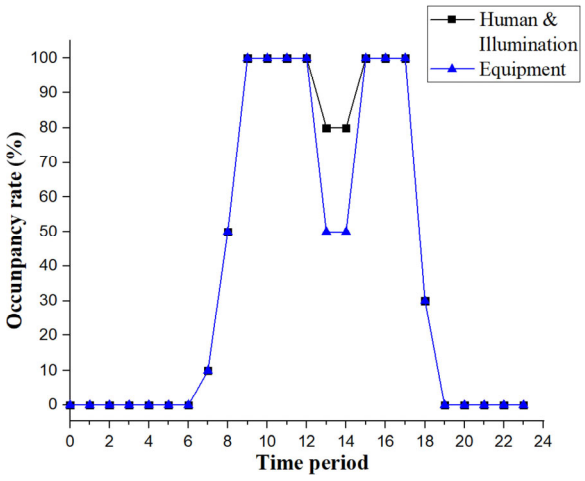

Fig. 5 The schedule setting for the personnel, lighting and equipment

with RL controller during a part of continuous operation stages. With the continuously changing of outdoor temperature, the RL controller can ensure that the comfort requirement can be met in the vast majority of the operation period. In order to quantify the control effect of RL controller, the rule-based control strategy (RBC) similar to start-stop control and PID controller with fixed set temperature are selected as the reference control strategies and comparison objects. The rules of set temperature limit in the RBC is the same to that in RL controller. The set temperature of the PID controller is 26 °C, and the indoor temperature is stabilized by adjusting the air supply volume.

Figure 8 shows the temperature variation of the controlled area in application of reference strategies. For both RBC and PID controllers, the temperatures in the controlled area meet the comfort requirements in most operation period. In

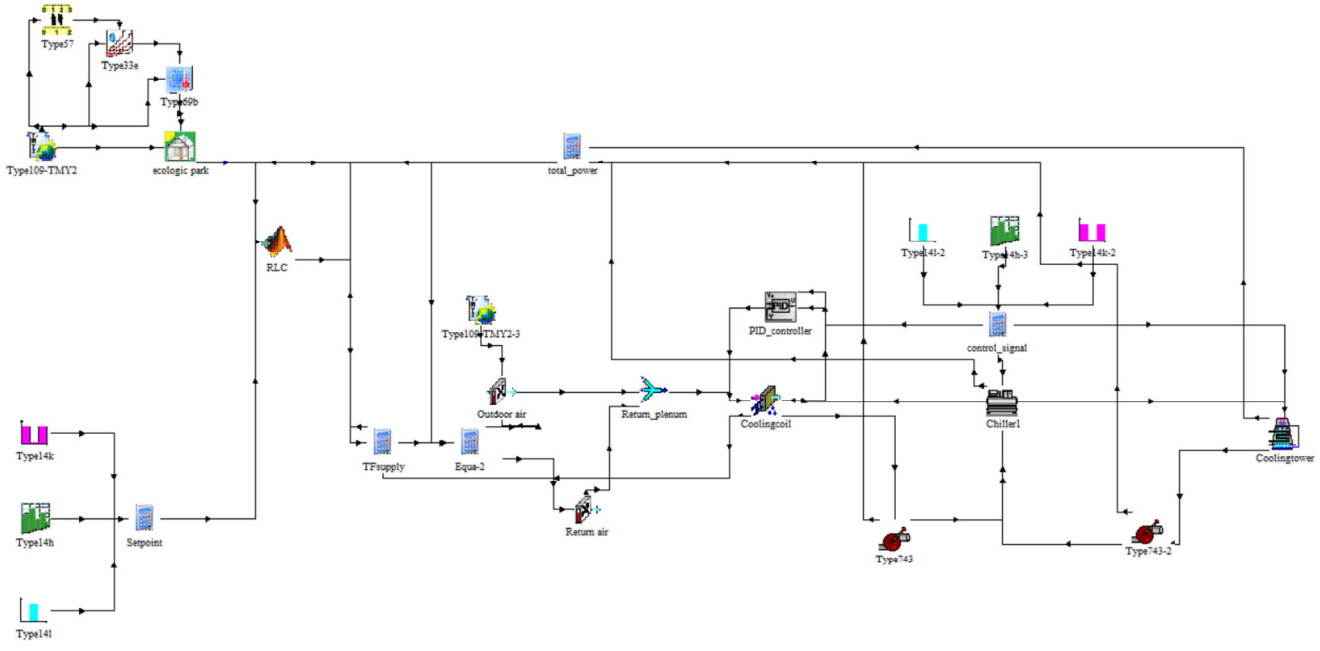

Fig. 6 The simulation model of building and AC system in TRNSYS

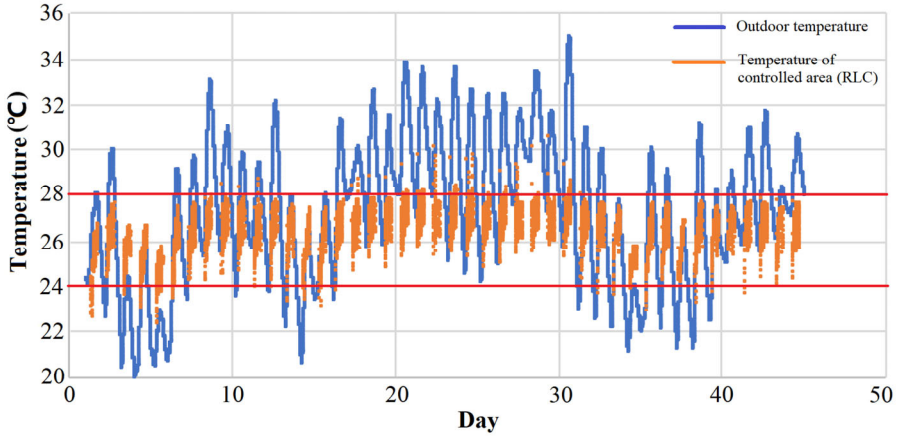

Fig. 7 Temperature variation in controlled area with RL controller during operation

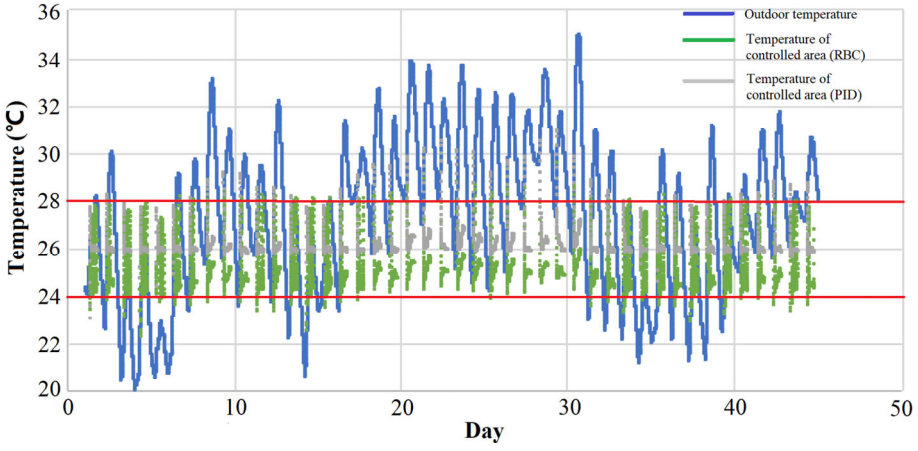

Fig. 8 Temperature variation of the controlled area using reference strategies

the use of PID controller, the temperature of the controlled area can be kept around the set value 26 °C at a small fluctuation. In the use of RBC strategy, the temperature in the controlled area fluctuates between 24 °C and 28 °C.

The annual operating costs of the AC system and the comfort of the controlled area are statistically compared in the three control strategies (RBC, PID controller and RL controller). Figure 9 shows the annual operating costs of RBC, PID controller and RL controller, while Fig. 10 shows the indoor comfort level under three control strategies. In the *x*-axis of Fig. 9 and Fig. 10, "Y" is the abbreviation of year. "Y1" and "Y2" represent the first and second year, respectively, and so on.

As shown in Fig. 10, during the exploration stage in the first year, the percentage of non-comfortable time increases as the RL controller continually attempts to update the strategies. However, in the operation stage, the RL controller can utilize the learned experience and has better performance than the reference strategy. As shown in Fig. 9, compared to RBC and PID controller, RL controller can reduce the

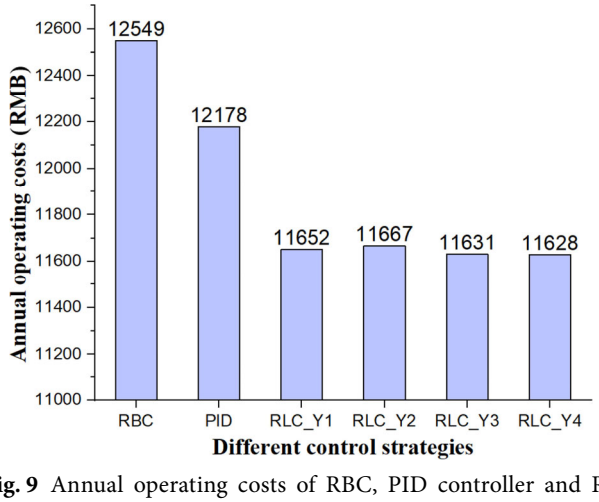

Fig. 9 Annual operating costs of RBC, PID controller and RL controller

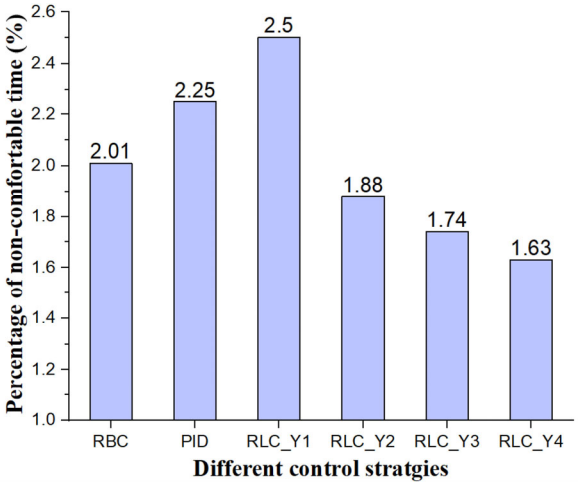

Fig. 10 Percentage of non-comfortable time in the control strategies of RBC, PID and RL

operating costs by more than 7% and 4.5%, respectively under the circumstance of slightly improving the indoor comfort level. In addition, as the runtime progresses, the RL controller can further improve the control strategy. Both the operating cost and the percentage of non-comfortable time reach the minimum in the fourth year applying RL controller. Thus, Fig. 11 only shows the comparison of energy consumption in the AC system with RBC, PID controller and RL controller in the fourth year. Compared to RBC and PID controller, the application of RL controller in the fourth year can reduce the cooling energy consumption and transmission and distribution energy consumption of the AC units. Compared to the application of RBC and PID controller in the AC system, the total energy consumption reduced by 7.7% and 4.7%, respectively.

4.5 Multi-zone air supply volume control

Performance of RL controller in multi-area air supply volume

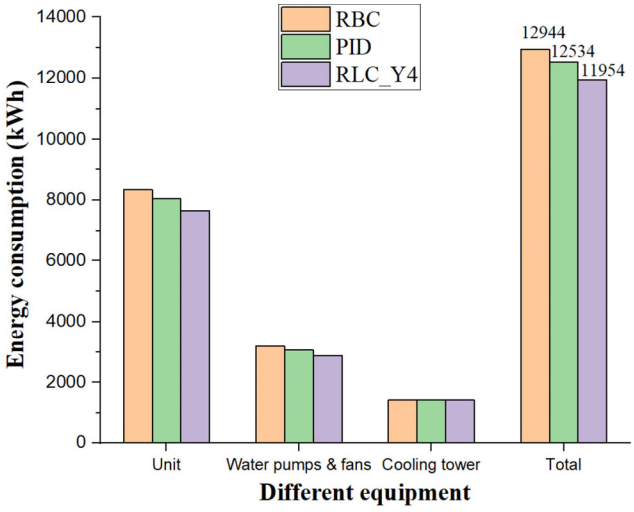

Fig. 11 Comparison of energy consumption by item in the control strategies of RBC, PID and RL

control should be further analyzed. The partition of the model is mainly determined by the room function, the air distribution form and the room orientation. Because the building area in this VAV case is small, it is not partitioned inside and outside. As shown in Fig. 12, the controlled area is the single area simulation is converted to the set partition form in this VAV case.

When applying RL controller to the multi-zone air supply volume control, the preparatory work in the preparation stage is roughly the same as that for single zone air supply volume control. The differences are the modification of the state, the action and the reward. The temperature of controlled area in state is changed to the indoor air temperature of each controlled area. The action is the air supply at the terminal of each VAV system. The reward is changed to the sum of the system energy consumption and the comfort penalty of each controlled area. In addition, the exploration stage is extended to two years due to the enlargement of the state-action space. In the first year, the exploration strategy remains unchanged, while in the second year, the ϵ-greedy exploration strategy with ϵ value of 0.05 is adopted.

The performance of the RL controller is tested by simulation and compared with the reference strategies. Figure 13 shows the performance comparison of different control strategies (RBC, PID and RL controller) for multi-zone air supply volume control. In the x-axis of Fig. 13, the "Y" is the abbreviation of year. "Y1" and "Y2" represent the first and second year, respectively, and so on. In the two reference strategies, the PID controller can reduce the operating cost by 3% compared with RBC, while the comfort performance is similar. Thus, PID controller is used as the comparison object in the VAV case. Compared with the PID controller, the application of RL controller reduces the operating cost by about 4% in the first three years, but

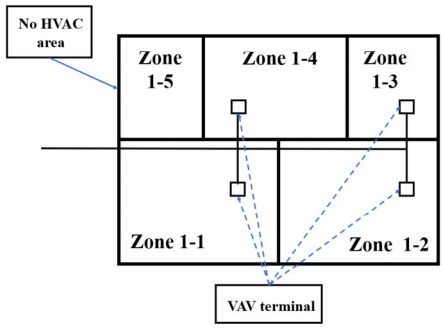

Fig. 12 Building partition

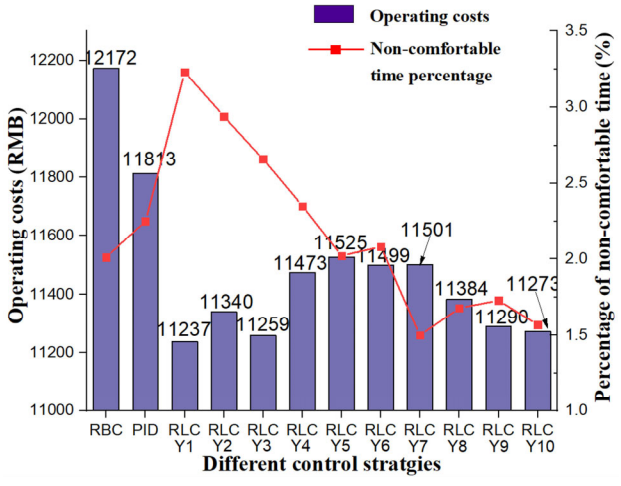

Fig. 13 Performance comparison of different control strategies for multi-area air supply volume

increases the non-comfortable time. During the following two years of RL controller operation, the operating cost reduces by 2.6% with similar non-comfortable time. However, when the operation time of RL controller is more than 6 years, it performs better in terms of non-comfortable time, and reduces the operating cost by 2.7% to 4.6%.

4.6 Discussion and analysis

When solving the problem of optimal control of building environment using VAV AC system, it is necessary to take full consideration of building dynamic thermal response characteristics, AC system performance characteristics, indoor thermal disturbance and outdoor weather conditions, so as to determine the set value of air supply volume in the controlled area within each control step. In the studied VAV case, the operation strategy of the AC system based on RL is applied. All the above information and feedback signal of control effect are provided to the RL controller. The exploration learning ability is utilized to summarize the system operation experience and continuously improve the

control strategy. In the VAV system case, the RL controller keeps the temperature of the controlled are to a higher level in the comfort zone and reduces the energy consumption of cooling, transmission and distribution of the system, thereby the operating costs.

For the air supply of single zone, after one-year exploration learning, the performance of RL controller is better than that of reference strategies (RBC and PID controller) in terms of both energy cost and non-comfortable time of AC system. For the air supply of multi-zone, compared to the reference cases, although the operating costs of the AC system with RL controller reduce significantly in the first four years, the non-comfortable times increase to varying degrees due to the obvious increase of state-action space in the first two-year exploration stage. During the two-year exploration stage, the non-comfortable time percentage is much high, while another two-year buffer stage is given to reduce the percentage. Actually, the non-comfortable time percentage continues to decrease during the first 4 years. In the following two years (years 5 to 6), compared to the reference cases, the application of RL controller in AC system can maintain the approximate non-comfortable time level, but reduce the energy costs of AC system. From the seventh year, the performance of RL controller is much better than that of reference cases in aspect of both non-comfortable time and operating costs of AC system. Thus, in the long term, RL controller can perform better than the reference control methods in both comfort degree and operating costs after exploration learning phase (including another 2-year buffer stage). In addition, RL controller can continuously improve control strategy as time goes on; however, the exploration cost will keep increasing with the increase of operation problem scale. Thus, RL controller has more advantages and practical application value in small-scale operation optimization problems instead of big-scale ones.

Currently, the performance research of RL controller on the operation optimization of AC system is only in theoretical and simulation research phase, which has not been validated on the actual operating AC systems. In addition, the results are may not be generalized for different AC systems. Thus, future works should be focused on applying the RL controller into different AC systems to analyze respective optimization performance, and validating its optimization performance on the actual AC systems.

5 Conclusions

RL is considered as a promising model-free supervisory control method to optimize HVAC system operation and achieve energy saving. Thus, RL algorithm is applied to the operation optimization of AC system in this paper, while an innovative RL-based model-free control strategy combining rule-based and RL-based control methods is proposed as well as complete application process. The new RL-based controller is applied in a VAV AC system for a single-storey office building as a case study to validate the performance of the RL-based controller. The RBC and PID controller are selected as the reference control strategies. The conclusions are shown as follows:

1) For the air supply of single zone, the RL controller performs the best in terms of energy cost and non-comfortable time after one year of exploration learning. Compared with RBC and PID strategies, the use of RL controller can reduce the total energy consumption by 7.7% and 4.7%, respectively.
2) For the air supply of multi-zone, the performance of RL controller begin to outperform the reference strategies after four-year studying, including two-year exploration stage and two-year buffer stage. From the seventh year on, the RL controller performs much better in terms of both non-comfortable time and operating costs of AC system than that of reference strategies. The operating cost is reduced by 2.7% to 4.6% compared with the reference strategies.
3) In the long term, RL controller can perform better than the reference control methods in both comfort degree and operating costs after exploration learning phase in the multi-zone application.

Also, some limitations should be mentioned that RL controller is more suitable for small-scale operation optimization problems instead of big-scale ones due to the exploration cost increase with the increase of operation problem scale. In addition, the results in this paper may not be generalized for different AC systems. Thus, future works should be done to apply RL controller to big-scale operation optimization problems and different AC systems to validate their respective optimization performances.

Acknowledgements

This study is supported by the Thirteenth Five-Year National Key Research and Development Program “Study on the Technical Standard System for Post-evaluation of Green Building Performance”, Ministry of Science and Technology of China (No. 2016YFC0700105).

References

Baird L (1995). Residual algorithms: Reinforcement learning with function approximation. In: Proceedings of the 12th International Conference on Machine Learning, Miami, FL, USA.

Baldi S, Michailidis I, Ravanis C, Kosmatopoulos EB (2015). Model-based and model-free “plug-and-play” building energy efficient control. *Applied Energy*, 154: 829–841.

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: Springer.

Brémaud P (1999). Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues. New York: Springer.

Cheng Z, Zhao Q, Wang F, Jiang Y, Xia L, Ding J (2016). Satisfaction based Q-learning for integrated lighting and blind control. *Energy and Buildings*, 127: 43–55.

Costa A, Keane MM, Torrens JI, Corry E (2013). Building operation and energy performance: Monitoring, analysis and optimisation toolkit. *Applied Energy*, 101: 310–316.

Costanzo GT, Iacovella S, Ruelens F, Leurs T, Claessens BJ (2016). Experimental analysis of data-driven control for a building heating system. *Sustainable Energy, Grids and Networks*, 6: 81–90.

Curtiss PS, Brandemuehl MJ, Kreider JF (1994). Energy management in central HVAC plants using neural networks. *ASHRAE Transactions*, 100(1): 476–493.

Defazio A, Graepel T (2014). A comparison of learning algorithms on the arcade learning environment. arXiv:1410.8620

DOE (2011). Building Energy Data Book. US Department of Energy. Available at http://buildingsdatabook.eren.doe.gov/.

Doll BB, Bath KG, Daw ND, Frank MJ (2016). Variability in dopamine genes dissociates model-based and model-free reinforcement learning. *Journal of Neuroscience*, 36: 1211–1222.

Dong B, O'Neill Z, Luo D, Bailey T (2014). Development and calibration of an online energy model for campus buildings. *Energy and Buildings*, 76: 316–327.

Dong B, Yan D, Li Z, Jin Y, Feng X, Fontenot H (2018). Modeling occupancy and behavior for better building design and operation—A critical review. *Building Simulation*, 11: 899–921.

Ernst D, Geurts P, Wehenkel PL (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6: 503–556.

Gao Y, Chen S, Lu X (2004). A review of reinforcement learning. *Journal of Automation*, 30(1): 86-100. (in Chinese)

Goyal S, Ingley HA, Barooah P (2013). Occupancy-based zone-climate control for energy-efficient buildings: Complexity vs. performance. *Applied Energy*, 106: 209–221.

Gunay HB, Ouf M, Newsham G, O'Brien W (2019). Sensitivity analysis and optimization of building operations. *Energy and Buildings*, 199: 164–175.

Halperin I (2019). The QLBS Q-learner goes NuQLear: Fitted Q iteration, inverse RL, and option portfolios. *Quantitative Finance*, 19: 1543–1553

Han M, May R, Zhang X, Wang X, Pan S, Yan D, Jin Y, Xu L (2019). A review of reinforcement learning methodologies for controlling occupant comfort in buildings. *Sustainable Cities and Society*, 51: 101748.

House JM, Smith TF (1995). System approach to optimal control for HVAC and building systems. *ASHRAE Transactions*, 101(2): 647–660.

Huang X (2017). Optimal control based on experience replay and Q-Learning. *Computer Engineering and Design*, 38(5): 1352–1355. (in Chinese)

Jaafra Y, Laurent JL, Deruyver A, Naceur MS (2019). Reinforcement learning for neural architecture search: A review. *Image and Vision Computing*, 89: 57–66.

Jung W, Jazizadeh F (2019). Human-in-the-loop HVAC operations: A quantitative review on occupancy, comfort, and energy-efficiency dimensions. *Applied Energy*, 239: 1471–1508.

Killian M, Kozek M (2016). Ten questions concerning model predictive control for energy efficient buildings. *Building and Environment*, 105: 403–412.

Lange S. Gabel ST, Riedmiller M (2012). Batch reinforcement learning. In: Wiering M, van Otterlo M (eds), Reinforcement Learning. Berlin: Springer. pp. 45–73.

Li J, Poulton G, Platt G, Wall J, James G (2010). Dynamic zone modelling for HVAC system control. *International Journal of Modelling, Identification and Control*, 9: 5–14.

Li B, Xia L (2015). A multi-grid reinforcement learning method for energy conservation and comfort of HVAC in buildings. In: Proceedings of IEEE International Conference on Automation Science and Engineering, Gothenburg, Sweden.

Ling KV, Dexter AL (1994). Expert control of air-conditioning plant. *Automatica*, 30: 761–773.

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

Mason K, Grijalva S (2019). A review of reinforcement learning for autonomous building energy management. *Computers & Electrical Engineering*, 78: 300–312.

Mbuwir BV, Ruelens F, Spiessens F, Deconinck G (2017). Battery energy management in a microgrid using batch reinforcement learning. *Energies*, 10: 1846.

MOHURD (2012). Design code for heating Ventilation and air conditioning of civil buildings (GB50736-2012). Ministry of Housing and Urban-rural Development of China. (in Chinese)

Nguyen ND, Nguyen T, Nahavandi S (2019). Multi-agent behavioral control system using deep reinforcement learning. *Neurocomputing*, 359: 58–68.

Niu F, O'Neill Z, O'Neill C (2018). Data-driven based estimation of HVAC energy consumption using an improved Fourier series decomposition in buildings. *Building Simulation*, 11: 633–645.

Ruelens F, Iacovella S, Claessens BJ, Belmans R (2015). Learning agent for a heat-pump thermostat with a set-back strategy using model-free reinforcement learning. *Energies*, 8: 8300–8318.

Russek EM, Momennejad I, Botvinick MM, Gershman SJ, Daw ND (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Computational Biology*, 3(9): e1005768

Široký J, Oldewurtel F, Cigler J, Prívara S (2011). Experimental analysis of model predictive control for an energy efficient building heating system. *Applied Energy*, 88: 3079–3087.

Sutton RS, Barto AG (1998). Reinforcement Learning: An Introduction. Cambridge, MA, USA: MIT Press, .

TRNSYS (2017). Transient System Simulation (TRNSYS) Program Documentation.

Urieli D, Stone P (2013). A learning agent for heat-pump thermostat control. In: Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), Saint Paul, MN, USA.

van Hasselt H (2010). Double Q-Learning. In: Proceedings of the 23rd International Conference on Neural Information Processing Systems.

Wang S, Ma Z (2008). Supervisory and optimal control of building HVAC systems: A review. *HVAC&R Research*, 14: 3–32.

Watkins CJCH, Dayan P (1992). Q-learning. *Machine Learning*, 8: 279–292.

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.

You C, Lu J, Filev D, Tsiotras P (2019). Advanced planning for autonomous vehicles using reinforcement learning and deep inverse reinforcement learning. *Robotics and Autonomous Systems*, 114: 1–18.

Zhao H, Magoulès F (2012). A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, 16: 3586–3592.