Applied Energy 281 (2021) 116117

Contents lists available at ScienceDirect

Applied Energy

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

Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning☆

Yan Du [a], Helia Zandi [b], Olivera Kotevska [b], Kuldeep Kurte [b], Jeffery Munk [b], Kadir Amasyali [b], Evan Mckee [a], Fangxing Li [a,*]

[a] Dept. of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, USA
[b] Oak Ridge National Laboratory, Oak Ridge, TN, USA

HIGHLIGHTS

- A deep reinforcement learning (RL) control strategy for residential HVAC is proposed.
- The control strategy is based on the deep deterministic policy gradient (DDPG) method.
- Simulation results prove the economy and time efficiency of the DDPG method.
- DDPG is compared with deep Q network (DQN) and baseline cases for verification.
- The generalization of the DDPG method is further verified in different scenarios.

ARTICLE INFO

Keywords:
Actor-critic learning
Demand response
Deep deterministic policy gradient (DDPG)
Deep reinforcement learning (deep RL)
Multi-zone residential HVAC

ABSTRACT

Residential heating, ventilation, and air conditioning (HVAC) has been considered as an important demand response resource. However, the optimization of residential HVAC control is no trivial task due to the complexity of the thermal dynamic models of buildings and uncertainty associated with both occupant-driven heat loads and weather forecasts. In this paper, we apply a novel model-free deep reinforcement learning (RL) method, known as the deep deterministic policy gradient (DDPG), to generate an optimal control strategy for a multi-zone residential HVAC system with the goal of minimizing energy consumption cost while maintaining the users' comfort. The applied deep RL-based method learns through continuous interaction with a simulated building environment and without referring to any prior model knowledge. Simulation results show that compared with the state-of-art deep Q network (DQN), the DDPG-based HVAC control strategy can reduce the energy consumption cost by 15% and reduce the comfort violation by 79%; and when compared with a rule-based HVAC control strategy, the comfort violation can be reduced by 98%. In addition, experiments with different building models and retail price models demonstrate that the well-trained DDPG-based HVAC control strategy has high generalization and adaptability to unseen environments, which indicates its practicability for real-world implementation.

1. Introduction

In the worldwide scope, buildings account for 40% of total primary energy consumption and 30% of all CO_2 emissions, among which a large portion can be attributed to thermal comfort overhead [1,2]. Therefore, it is important to study the effective energy management of building demand to achieve economic and environmental benefits.

The heating, ventilation, and air conditioning (HVAC) system is currently the most widely used device for maintaining building thermal comfort. It also serves as an important demand response resource for peak load reduction and stabilizing system-wide operation via proper

☆ DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan.)

* Corresponding author.
E-mail address: fli6@utk.edu (F. Li).

https://doi.org/10.1016/j.apenergy.2020.116117
Received 17 April 2020; Received in revised form 21 October 2020; Accepted 25 October 2020

Nomenclature

$T_{out}(t)$	Outdoor temperature at time step t
$T_{in,z}(t)$	Indoor temperature for room zone z at time step t
$T_{lower}(t)$	Lower bound of the user comfort level at time step t
$\lambda^{retail}(t)$	Retail price at time step t
$Setpt_z(t)$	Setpoint for room zone z at time step t
$E_{HVAC}(t)$	Power consumption of HVAC system at time step t
Δt	Control interval of the HVAC system
$c^{penalty}(t)$	Penalty for user comfort violation at time step t
θ^Q, θ^π	Neural network parameters of the critic network and the actor network in the DDPG algorithm
$Q(s,a;\theta^Q)$	Action-value of the state-action pair (s,a) under the current critic network θ^Q
$\pi(s;\theta^\pi)$	Control policy under the current actor network θ^π
$q^{target}(t)$	Target action-value for updating the critic network

demand-side energy management strategies [3]. In the literature, there are many studies focusing on optimizing HVAC control strategies for improving energy efficiency. In [4], the energy management of HVAC systems is modelled under load forecast errors, where a primal–dual algorithm is applied to seek the optimal operating states of HVAC for the consumer, and the pricing strategy for the energy provider. In another work, a regression approach is applied for temperature forecast in day-ahead scheduling of responsive residential HVAC demand [5]. The authors in [6,7] discuss the potential of using the HVAC system to provide primary frequency regulation to the bulk system via a hierarchical control strategy. A Lyapunov optimization technique is introduced in [8] for HVAC load control without needing to estimate the uncertain system factors such as price and temperature. A distributed transactive control market mechanism for commercial building HVAC systems is presented in [9] to demonstrate the effectiveness of HVAC in peak shaving and load shifting.

All the above methods can be categorized as model-based methods, where the detailed thermal dynamics of the HVAC with consideration of ambient environment effects need to be modelled, along with the requirement of analytical solution toolboxes for practical runtime control. The model-based methods may suffer from measurement errors (e.g., building model inaccuracy), as well as computational inefficiency, since the building and equipment models must be tailored to a specific building to achieve accurate results. This represents a serious challenge for widespread deployment of model-based methods.

Meanwhile, there has been significant development in machine learning technologies such as deep learning and reinforcement learning evidenced by the achievement of AlphaGo [10]. In power systems research community, general vision of new research directions related to machine learning is discussed in [11] with a number of research applications [12,13]. In industrial applications, AI-based implementation starts to be deployed in real control centers such as [14], which is the first reported control-room application of AI-driven distributed feature selection for a large, real power grid.

More specifically, in recent years, deep reinforcement learning (RL), which is a combination of a deep neural network (DNN) and RL, has attracted broad attention in solving high-dimensional control and optimization problems with tremendous complexity. A double Q learning method [15] and a continuous deep deterministic policy gradient (DDPG) method [16] have been applied for optimizing the energy management strategies of hybrid electric vehicles, respectively. In [17], the asynchronous advantage actor-critic is employed to find the economic operation schedules of multiple distributed energy resources within an energy Internet. In [18], a deep Q learning method is designed for supporting the maintenance decision-making of the bulk power system. Given the potential operation constraints encountered during the implementation of deep RL-based control actions, a safe deep RL method is explored in [19] to obtain the optimal control scheme of the active distribution network with the consideration of voltage level limits, which introduces a safe layer on top of the conventional actor network to avoid any possible violations of the voltage constraints.

With specific respect to the HVAC system control problem, there have also been some pioneering works in the literature focusing on utilizing the powerful deep RL approach to achieve higher energy efficiency and economic efficiency. In [20], a deep Q network (DQN) is constructed for coordinated control of joint datacenter and HVAC load, in which the neural network is utilized to estimate the Q value of a state-action pair. In [21], a convolutional neural network (CNN) is deployed as the approximator of the state-action value function to better capture the spatial and temporal correlations within the input state data with its convolutional operation. A deep policy gradient (DPG) method is investigated in [22] for controlling multiple responsive demands including ACs, electric vehicles and dishwashers. In [23], an actor-critic method is applied for optimizing the thermal comfort and energy consumption of HVAC. In [24], a practical HVAC control framework based on advantage actor critic is established for a whole building energy model. In [25], the DQN is applied to achieve optimal control balancing between different HVAC systems.

All the above research works have demonstrated the effectiveness of the applied deep RL methods in optimizing the HVAC thermal control strategy compared with the designed benchmarks. However, the majority of the existing researches treat the continuous control actions of the HVAC system, such as HVAC setpoint or air flow rate, in a discretized way to narrow down the search space. Discretization can achieve satisfying performance when the granularity is low or without the combination of action spaces. However, it encounters the issue of exponential explosion when the action space is high-dimensional, for example, multiple room zones in the case of HVAC control. As a result, more simulations are needed for training the deep RL methods and the algorithm performance decreases.

In [26], the authors adopt the DDPG method to realize the continuous thermal control of HVAC without discretization. However, this research work still focuses on single-zone HVAC control, which has been previously addressed by the above-mentioned discretion methods. In addition, the method applied is only compared with other RL methods, and no benchmark cases are designed to verify the optimality and the generalization of the obtained control strategy. In [27], a multi-agent deep RL method with an attention mechanism is applied to minimize the energy costs of an HVAC system in a multi-zone commercial building, where a set of actor and critic networks are designed for each zone, and they are updated in parallel during the training. While this research work provides some inspiring insights, one concern is that in the proposed algorithm, the number of neural networks needing training will grow with the number of zones, which could cause excessive computational burden. In [28], the long-short-term-memory (LSTM) recurrent neural network is combined with the DDPG to better simulate the real-world operation of multiple air handling units (AHUs), where a deep RL agent is designed for each AHU to control a separate section of the building. The same concern occurs regarding the number of RL agents and the growing computational cost.

Motivated by the above concerns, in this paper, we also apply the DDPG method for optimizing the continuous thermal control strategy of residential HVAC. The main contributions of this work, as compared with the existing research, are summarized as follows:

- We apply the DDPG RL method to optimize the continuous control of multi-zone residential HVAC. The multi-zone residential HVAC control involves more complex thermal dynamics and environment uncertainties, and a high-dimensional action space, which requires more delicate problem formulation including the definitions of state, action, and reward during the learning process;

- We conduct a comprehensive comparison between the applied DDPG method and the widely-used DQN method to demonstrate the effectiveness of the former in dealing with the continuous action space, which is a more common case in many real-world situations; we also design benchmark cases without RL to prove that the applied DDPG can achieve higher economic benefits while maintaining user comfort;
- We verify that the well-trained deep RL method has obtained high generalization and robustness, and can adapt to new environment with different price signals and physical conditions to provide the optimal HVAC control strategy.

The rest of the paper is organized as follows. The HVAC control problem formulation is introduced in Section 2; in Section 3, the two representative deep RL methods, the DQN and DPG methods are first briefly reviewed, followed by a detailed explanation of the DDPG method, which is an extension of the former two; the simulation results of the DDPG method are presented in Section 4, plus comparison with the DQN and benchmark cases; finally, Section 5 concludes the paper.

2. Multi-zone residential HVAC system control problem formulation

2.1. A brief introduction of the multi-zone HVAC system control problem

In this study, we consider a residential building with multiple zones. The indoor temperature of each zone can be controlled by adjusting the setpoint of the HVAC system. The HVAC system can work in various modes including "Cooling", "Heating" and "Auto". The "Auto" mode means that the HVAC system can automatically switch between cooling and heating according to the indoor temperature and the assigned setpoint. Whenever there is a difference between the indoor temperature and the setpoint, the HVAC system will be automatically turned on to push the indoor temperature near to the setpoint to maintain user comfort. Without losing generality, in this work, we will focus on the case when all zones need heating. The goal of controlling the HVAC system is to minimize the energy cost while keeping the indoor temperature within the user comfort band.

2.2. Mapping HVAC control problem to Markov decision process (MDP)

In this subsection, we will formulate the above multi-zone residential HVAC control problem as a Markov Decision Process (MDP), which will later be solved by a model-free deep RL-based algorithm in Section 3. According to the simplified thermal dynamics model of HVAC in [29], the indoor temperature at the current time interval is only related to the previous state parameters such as the indoor temperature at the previous time interval, and is not affected by indoor temperature at any other time intervals. Therefore, the HVAC control problem can be regarded as a finite Markov process and be solved using the RL method.

An MDP is composed of four essential elements: state (s), action (a), state transition probability (p), and reward (r). In the context of a multi-zone residential HVAC control problem, the four elements are defined as follows:

- State: 1) current outdoor temperature $T_{out}(t)$; 2) current indoor temperature $T_{in,z}(t)$ for the all the zones z; 3) the lower bound of the user comfort level $T_{lower}(t)$; 4) retail price $\lambda^{retail}(t)$, where t is the current time step.

Note that the state parameters include the lower bound of the user comfort level, which changes along with the time. This is because we assume that the HVAC users have a time-variant comfort preference. This is reasonable since during the daily work hours when no one is at home, the comfort range of the indoor temperature can be lowered to save the energy cost. The comfort range can be brought back during the off-work hours when the house is occupied.

The state parameters also include the current retail price to realize the pre-heating effect of HVAC. Pre-heating means setting the setpoint of the HVAC at a relatively high value when the retail price of energy is low to heat up the indoor temperature in advance, thus avoiding excessive energy consumption when cold outdoor temperatures occurs, when the retail price of energy is higher.

- Action: the setpoint $Setpt_z(t)$ for the zone z;

The HVAC setpoint in each zone is a continuous variable. Given the setpoint, the on/off status of the HVAC unit with a thermostat at each zone obeys the following control logic:

$$HVAC\ status = \begin{cases} 1,\ if\ T_{in}(t) < setpoint - deadband \\ 0,\ if\ T_{in}(t) > setpoint \\ remain\ at\ the\ current\ status,\ elsewise \end{cases} \quad (1)$$

The HVAC model considered in this paper is only utilized for heating. In Eq. (1), the deadband is a small temperature span, in which the thermostat will not change its on/off status to prevent short cycles. It can be observed in Eq. (1) that if the indoor temperature is above the setpoint, the HVAC will remain off; otherwise, the HVAC will be started automatically to heat the room to maintain the user comfort.

- Reward: the energy consumption cost plus the comfort violation cost for the control interval, which is defined as follows:

$$r(t) = -\omega_c \sum_{t'=t-\Delta t}^{t} \lambda^{retail}(t') E_{HVAC}(t') - \omega_p \sum_{t'=t-\Delta t}^{t} c^{penalty}(t') \quad (2)$$

In Eq. (2), the first term is the energy cost of the HVAC system, where $\lambda^{retail}(t')$ is the retail price, $E_{HAVC}(t')$ is the power consumption, and Δt is the control interval; the second term is the penalty for user comfort violation, which is calculated as follows:

$$c^{penalty}(t') = \begin{cases} 1,\ for\ T_{in}(t') < T_{lower}(t') - T_{th} \\ 0,\ elsewise \end{cases} \quad (3)$$

In Eq. (3), T_{th} is a threshold with a small value. The temperature violation is not counted if the magnitude of the violation is smaller than T_{th}. Given the existence of the deadband within the HVAC system, it is not possible to always keep the indoor temperature at the exact setpoint. The threshold allows for some deviations of the indoor temperature.

Because the reward encloses both the energy cost and the penalty, which leads to a multi-objective function, weight factors are added to the two objectives, which are represented by ω_c and ω_p in Eq. (2). The final objective of HVAC thermal control is to minimize the total energy consumption cost plus the penalty over the entire control cycle, which can be written as the cumulative sum of $r(t)$: $\sum_{t=1}^{N_T} r(t)$. Therefore, a far-sighted control strategy is needed to prevent against uncertain future circumstances, which leads to a multi-stage decision making problem.

Notice that the state transition probability p is not defined for the above MDP. The state transition probability refers to the probability of transferring to a certain next state after taking action $Setpt_z(t)$. With a known state transition probability, the MDP is fully observed and the cumulative reward can be analytically solved via model-based dynamic programming or other iterative methods. However, in the HVAC control problem, to obtain an accurate probability model of the state transitions is not a trivial task, because it is difficult to formulate the exact thermal-dynamic model of HVAC buildings. The heat transfer within the building is related to multiple resistances (R) and capacitors (C) from different building components, like the exterior walls, the interior walls and furnishings, and the attic, the values of which require estimation and validation through experimenting. All these factors can have a significant impact on the temperature response of the indoor air [30]. Furthermore, the indoor temperature is also affected by uncertain external factors such as outdoor temperature, solar irradiance, and

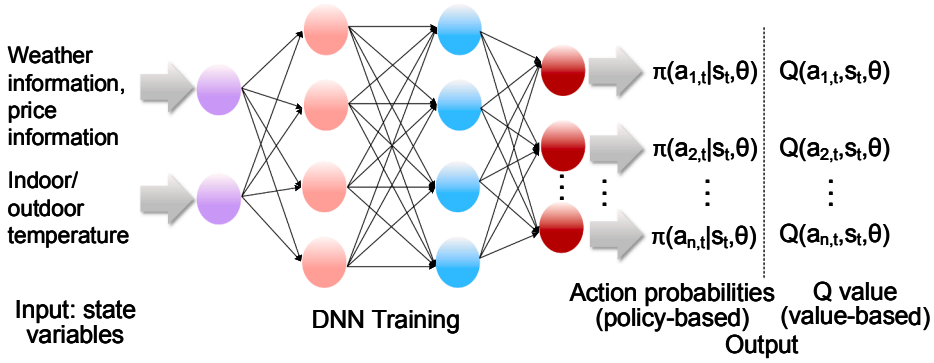

Fig. 1. DNN structure for function approximation in RL.

wind, which calls for additional modelling and computational efforts. As a consequence, a model-based method is not a robust or adaptive solution for HVAC system optimization.

Driven by the above considerations, in this paper the model-free deep RL method is leveraged to overcome the unobservability in the multi-zone residential HVAC control problem. The model-free RL method does not require any knowledge of the environment or the state transitions in advance. It gradually improves its decision-making strategy by continuously interacting with the environment and receiving feedback. In this way, the forecast errors of uncertain factors, as well as the measurement errors of building thermal mass, can be avoided. More details of the deep RL method will be revealed in the next section.

3. DDPG-based control strategy for multi-zone HVAC system

3.1. A brief review of deep reinforcement learning methods

The RL method is a type of machine learning method that optimizes the decision-making strategy in an MDP. In the RL algorithm, the reward defined in the MDP is served as the guideline for algorithm evolution. A large, positive reward will encourage the algorithm to search deep in the current direction, and vice versa. The RL method is especially suitable for handling decision-making problems with temporal constraints or with hidden state space.

There are two main types of RL method: the value-based RL method and the policy-based RL method. The difference between the two methods lie in their action evaluation strategies. The value-based method estimates the Q value of a state-action pair (s,a), which is the cumulative discounted reward starting from taking action a at state s, and selects the action with the highest Q value; the policy-based RL method generates the probabilities of all the feasible actions at the current state, and selects the action with the highest probability.

The combination of RL with a DNN is called the deep RL method. In deep RL, the DNN is utilized as a regression tool to estimate either the Q value, as in the value-based RL method; or the action probability, as in the policy-based RL method. A general DNN structure for regression in RL is shown in Fig. 1.

The main advantage of the deep RL method over the conventional RL method is that the application of the DNN makes it possible to achieve high level control for extremely complex problems, such as with continuous state space or action space, without the tabular constraints. In deep RL a more generalized regression model is established instead of maintaining a concrete Q table to store all the possible action values, as in the case of traditional Q learning. This generalized regression model offers more robust and flexible strategies against unseen states in the case of continuous control. In the following section, we will first introduce the DQN, as a representative of the valued-based deep RL methods; and the DPG method, as a representative of the policy-based deep RL methods. Then, a continuous control method, the DPG method, which is a combination of the above two methods, will be explained in detail for solving the optimal multi-zone residential HVAC control problem.

3.2. Understanding the basic principles behind typical deep RL methods

1) Deep Q Network (DQN)

The DQN is a combination of Q-learning and a DNN. In the DQN, the input is the current state, and the output is the Q value for each potential action at the current state. The advantage of the DQN over the tabular Q-learning method is that when the state and action are slightly changed, the DQN can still estimate the associated Q value without re-training, which is highly time-efficient.

Unlike the supervised learning algorithm, in deep RL there are no labeled samples for the DNN to learn. To handle this issue, two DNNs are designed for the DQN algorithm: one is called the target network, and the other is called the behavior network. The function of the target network is to serve as a reference, similar to the ground truth in the supervised learning, to guide the evolution of the algorithm.

Both networks are initialized with the same parameters and the same structure. As the training proceeds, the behavior network is updated at a faster speed than the target network. The loss function in the DQN is defined as the mean square error (MSE) between the target Q value and the behavior Q value. Once the loss function is calculated, the parameters of the behavior network will be updated based on its gradient to the loss function. The algorithm will continue updating until the output from the target network and the behavior network are close to each other, which indicates the convergence of the learning. More details of the DQN method can be found in [31].

2) Deep Policy Gradient (DPG)

The DPG method utilizes a strategy different from the DQN for control optimization. The output from the DNN is the probabilities of each potential action at the current state, or the policy. The policy refers to the probability of selecting action $a(t)$ at state $s(t)$, and can be written as $\pi(a|s,\theta) = \Pr\{a(t) = a|s(t) = s,\ \theta(t) = \theta\}$. θ stands for the parameters of the probability function. The loss function of the DPG method is also different from that of the DQN, which intends to maximize the expected total reward under the policy $\pi(a|s,\theta)$, and can be expressed as follows:

$$max\ \ J(\theta) = E_{\pi(a|s,\theta)}(\sum_{t=1}^{N_T} r(t)) = \sum_{\tau} \pi_\theta(\tau) R(\tau) \tag{4}$$

In Eq. (4), τ is called an episode generated under the policy $\pi(a|s,\theta)$: $\tau=\{s(1), a(1), s(2), a(2),\ldots, s(N_T), a(N_T)\}$. $R(\tau) = \sum_{t=1}^{N_T} r(t)$, which is the total reward of the episode. The goal of the DPG method is to get the parameters of the policy π that leads to the maximum value of the expected total reward. More details of DPG algorithm can be found in [22].

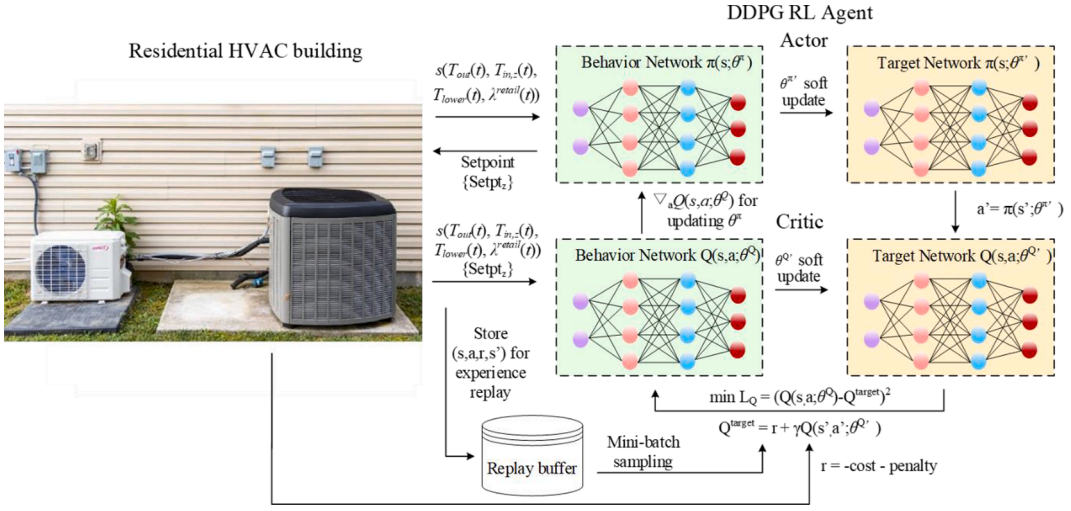

Fig. 2. Multi-zone HVAC control framework with DDPG.

3.3. Realizing continuous control of HVAC system with the DDPG

1) An introduction to the DDPG

The DDPG method is specially designed for solving problems with continuous variables. Unlike the DQN or DPG, where the Q values or action probabilities of all feasible actions are generated by the DNN for the agent to select, the term "deterministic" in the DDPG refers to the fact that there is only one output from the DNN, which is determined. In this way, the action space can be continuous since there is only one output unit.

Another advantage of the DDPG over the DQN and DPG is that it is a combination of the two methods. In the DDPG, there are two types of neural networks applied: the actor network, which assembles the DPG, and the critic network, which assembles the DQN. Their functions are explained as follows.

The input to the actor network is the current state, and the output is a deterministic action; the input to the critic network is the current state plus the action generated by the actor network, and the output is the Q value of the state-action pair. This Q value will be further used to update the parameters of the actor network. The loss function of the actor network is defined to maximize the Q value with the current policy, which follows the logic of the DPG method; and the loss function of the critic network is the MSE of the Q value, which follows the logic of the DQN method. In summary, the function of the actor network is to select actions, and the function of the critic network is to evaluate the selected action.

In addition, similar to the DQN algorithm, for both actor network and critic network in the DDPG, two neural networks are designed, a behavior network and a target network. Hence there are four neural networks in total. The reason for applying the target network is to stabilize the algorithm convergence. More details of the DDPG algorithm are presented in the next subsection.

2) DDPG algorithm for developing optimal HVAC control strategy

The details of the proposed DDPG algorithm are shown in **Algorithm 1**, which is customized from a general-purpose DDPG algorithm in [32]. The DDPG algorithm follows a process similar to that of the DQN, except that an actor network is built to select a deterministic action. The applied DDPG algorithm is further explained as follows:

To begin with, two neural networks, i.e., the actor network and the critic network are randomly initialized, and their associated target networks are initialized with the same set of parameters, as shown in lines 1–2. Starting from line 3, for each iteration, the system state is first initialized, then an HVAC control action, i.e. the setpoint, is chosen based on the current actor network $\pi(s;\theta^{\pi})$, as shown by line 7. A noise is added to the selected action to boost the exploration of the algorithm.

Next, in lines 8–9, the selected action is executed in the environment for the entire control interval Δt, and the received reward and the next state are observed. The transition $(s(t), \text{Setpt}_z(t), r(t), s(t+\Delta t))$ is stored in a replay buffer to be further used for algorithm training. When a sufficient number of transitions is collected, a mini-batch of transitions is randomly selected to update the parameters of the actor network and the behavior network, as shown by line 11. The random selection can cut off the temporal correlations among the transitions, which will maintain the independent, identically distributed assumption in the learning model. Also, the transitions can be sampled multiple times, which increases their utilization efficiency.

The neural network parameters θ^{Q} and θ^{π} are updated according to the loss functions. The loss function of the critic network is defined as the MSE between the target Q value and the current Q value from the behavior critic network, as shown by line 12. The temporal-difference error is used to update the Q value, where the target Q value is the sum of the current reward plus a discounted Q value from the target critic network $\theta^{Q'}$ for the next control interval $t+\Delta t$. γ is called the discount factor. Once the loss function is calculated, the parameters of the behavior critic network θ^{Q} are updated based on the gradient, as shown by line 13. η_Q is called the learning rate.

The loss function of the actor network is defined to maximize the Q value:

$$max \quad \frac{1}{M}\sum_{i=1}^{M} Q(s^{(i)}(t), a^{(i)}(t); \theta^{Q}) | a^{(i)}(t) = \pi(s^{(i)}(t); \theta^{\pi}) \tag{5}$$

In Eq. (5), $a^{(i)}(t)$ is generated from the actor network $\pi(s;\theta^{\pi})$. Hence, the chain rule is applied in line 14 to calculate the gradient of the Q value to the θ^{π}. In line 16, the parameters of the target critic network and the target actor network, $\theta^{Q'}$ and $\theta^{\pi'}$, are updated at a slower rate than the behavior network, where τ is a number between 0 and 1 and close to 1. The function of this slower update is to increase the stability of the learning. The complete deep RL-based control framework of a multi-zone HVAC system is shown in Fig. 2.

Algorithm 1. (*DDPG method for multi-zone HVAC control*)

Initialize the parameters of the critic network $Q(s,a;\theta^Q)$ and the actor network $\pi(s;\theta^\pi)$
Initialize the target networks $Q(s,a;\theta^{Q'})$ and $\pi(s;\theta^{\pi'})$ with θ^Q and θ^π
  for episode = 1 to arbitrary number **do**
    Initialize system state $s(T_{out}(0), T_{in,z}(0), T_{lower}(0), \lambda^{retail}(0))$
      for $t = 1$ to N_T **do**
        if $t == k\Delta t$, where k is an integer, **do**
          Select the multi-zone HVAC control action $\text{Setpt}_z(t)$ with $\pi(s;\theta^\pi)$ plus noise
          Execute $\text{Setpt}_z(t)$, receives the immediate reward $r(t)$ and the next state $s(t+\Delta t)$
          Store the transition ($s(t)$, $\text{Setpt}_z(t)$, $r(t)$, $s(t+\Delta t)$) in the replay buffer
        end if
        Collect a mini-batch of transitions ($s^{(i)}(t)$, $\text{Setpt}_z^{(i)}(t)$, $r^{(i)}(t)$, $s^{(i)}(t+\Delta t)$) with the size M from the replay buffer
        Calculate the MSE of the Q value:
          $q^{target(i)}(t) = r^{(i)}(t) + \gamma Q(s^{(i)}(t+\Delta t), \pi(s^{(i)}(t+\Delta t);\theta^{\pi'});\theta^{Q'})$
          $L(\theta^Q) = 1/M\sum_{i=1}^{M}(q^{target(i)}(t) - Q(s^{(i)}(t),\pi(s^{(i)}(t);\theta^\pi);\theta^Q)$
        Update the parameters of the critic network:
          $\theta^Q = \theta^Q - \eta_Q \nabla_\theta^Q L(\theta^Q)$
        Calculate the gradient of the Q value to the actor network parameter θ^π:
          $\nabla_\theta^\pi J \approx 1/M\sum_{i=1}^{M} \nabla_\pi Q(s^{(i)}(t),\pi(s^{(i)}(t);\theta^\pi);\theta^Q)\nabla_\theta^\pi \pi(s^{(i)}(t);\theta^\pi)$
        Update the parameters of the actor network:
          $\theta^\pi = \theta^\pi - \eta_\pi \nabla J$
        Update the parameters of the target network with a smaller step:
          $\theta^{Q'} = (1-\tau)\theta^Q + \tau\,\theta^{Q'}$
          $\theta^{\pi'} = (1-\tau)\,\theta^\pi + \tau\,\theta^{\pi'}$
        end for
    end for

4. Case study

In this section, the effectiveness of the applied DDPG-based continuous control method for multi-zone residential HVAC is demonstrated through simulations with real-world data, as well as by comparison with the DQN-based discrete control method and the benchmark cases, to fully verify the advantages of the DDPG method. Further, the generalization of the deep RL method is demonstrated by experimenting with unseen physical environments.

4.1. Simulation environment

A two-zone residential HVAC model [33] is implemented for training and testing the applied deep RL method, with real-world weather data from 2019 to 2020 obtained from [34]. For price signals, a simulated retail price sequence is generated, which includes a high price value and a low price value. The price is regularly switched between the two values every three hours. The reason for applying such a frequently changing price sequence is to find if the deep RL agent can identify the effect of price signals on the reward function and properly adjust its control strategies. It is further assumed that the lower bound of the user comfort level changes four times during the daily cycle, as shown in Table 1:

The control interval of the RL agent is set to 60 min, i.e., $\Delta t = 60$. Since we only focus on the heating effect of the HVAC system, the November weather data is used as the training data. During the training, one episode is defined as 24 h. In this way, 24 ($s^{(i)}(t)$, $\text{Setpt}_z^{(i)}(t)$, $r^{(i)}(t)$, $s^{(i)}(t+\Delta t)$) transitions will be generated from each episode. In total 300 episodes are simulated for the RL agent to learn. After the training, the RL agent will be applied to new test days with different weather conditions to examine its generalization and adaptability.

Table 1
Daily user comfort level.

Time period	0:00 – 6:00	6:00 – 12:00	12:00 – 18:00	18:00 – 24:00
T_{lower} (°C)	18	17	18	19

Table 2
DNN structure applied in DDPG and DQN algorithms.

Algorithm	DDPG		DQN
	critic network	actor network	
Size of input	[1,7]	[1,5]	[1,5]
No. of hidden layers	2	2	2
Size of each hidden layer	[7,20], [20,10]	[5,20], [20,10]	[5,20], [20,10]
Size of output	[1]	[2]	[25]
Activation function for the hidden layer	ReLU	ReLU	ReLU
Optimizer	Adam	Adam	Adam
Learning rate (η)	0.001	0.01	0.01
Discount factor (γ)	0.99	–	0.99
Batch size	48		
Weights of the reward	ω_c: 10, ω_p: 1		

4.2. Design of the DNN structure in deep RL

The detailed design of the actor and critic network in the DDPG is shown in Table 2. The design of the DQN is also listed for comparison. The designs of both the DDPG and the DQN are obtained via a trial-and-error process, and the current configurations provide the best possible results among all the trials.

For the DDPG method, the input to the critic network is a vector containing both state variables and action variables, and the output is the estimated Q value, which is a scalar; the input to the actor network is a vector containing only state variables, and the output is a vector containing the setpoint for each zone. Although the setpoint is a continuous variable, in reality there is always a range of the setpoint for maintaining user comfort. Therefore, the output layer from the actor network utilizes tanh as the activation function, which confines the output with a range of [−1, 1]. The actual setpoint is calculated as $\text{Setpt}_z = T_{lower} + \Delta T \cdot (y_{out} + 1)$, where y_{out} is the output from the actor network, and ΔT is the upper range of the setpoint. In the simulation, ΔT is set to 2 °C. Therefore, the setpoint selected by the DDPG lies within the range of $[T_{lower}, T_{lower} + 2]$.

For the DQN method, the inputs are also the state variables. Since the DQN requires a discrete action space, we discretize the range of setpoints with a step size of 0.5 °C. As a result, there are 5 actions for each zone and 25 combinations of actions for the 2-zone HVAC. The output from DQN is a vector containing 25 Q values, with each corresponding to one combination of actions.

4.3. Performance of the continuous HVAC control method

1) Convergence of the DDPG

In Fig. 3, the average returns gained after each episode during the training process in the DDPG and the DQN are presented. Notice that

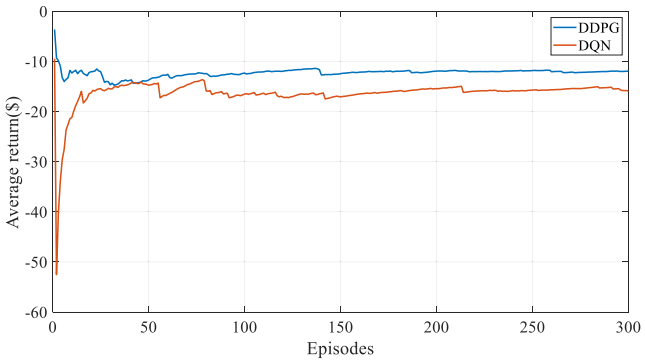

Fig. 3. Convergence of different deep RL methods.

Table 3
Test results of different HVAC control methods.

Control method	DDPG	DQN	Rule-based	Fixed setpoint
Total cost ($)	55.21	65.03	39.08	71.48
Temperature violation (min)	48	230	2617	0
Average temperature violation (°C)	0.13	0.93	1.85	0

the average returns in the first few episodes appear to be higher than that of the last few episodes. This is because for each episode, one training day is randomly chosen. Some training days may have moderate outdoor temperatures, which can lead to low energy cost and low penalty, and vice versa. However, as the training proceeds, the number of episodes grows, and the average return is neutralized. It can be observed that both curves gradually become steady as the training evolves. However, the average return gained by the DDPG method is higher than that of the DQN method. This is because the size of the output from the DQN is larger than that of the DDPG, and the combination of actions have not been fully explored after 300 episodes, leading to a lower average return.

2) Computational efficiency

After the training process, the DDPG RL agent is applied to 10 test days in January 2020 from the real-world data in [34] to generate the optimal HVAC control strategy. The time cost is around 19 s for testing, which is highly time-efficient. The code is written in Python 3.6 with the open-source deep learning platform TensorFlow [35]. The hardware environment is a laptop with Intel®CoreTM i7-7600U 2.8 GHz CPU, and 16.00 GM RAM.

3) Comparison of the DDPG with the DQN and the benchmark cases

In this study, the well-trained deep RL agents from both the DDPG and the DQN are run on new test days to verify their learning performance. We also design two benchmark cases without the RL agent as comparisons. The benchmark cases are described as follows: a) Rule-based case: the setpoint is set at the lowest value at the peak price hours, and the highest value at the off-peak price hours, to realize the pre-heating effect to save energy cost; b) Fixed setpoint case: the setpoint is always at the highest value of the setpoint range to avoid any temperature violation.

The final optimized results of the RL methods and the benchmark cases are shown in Table 3:

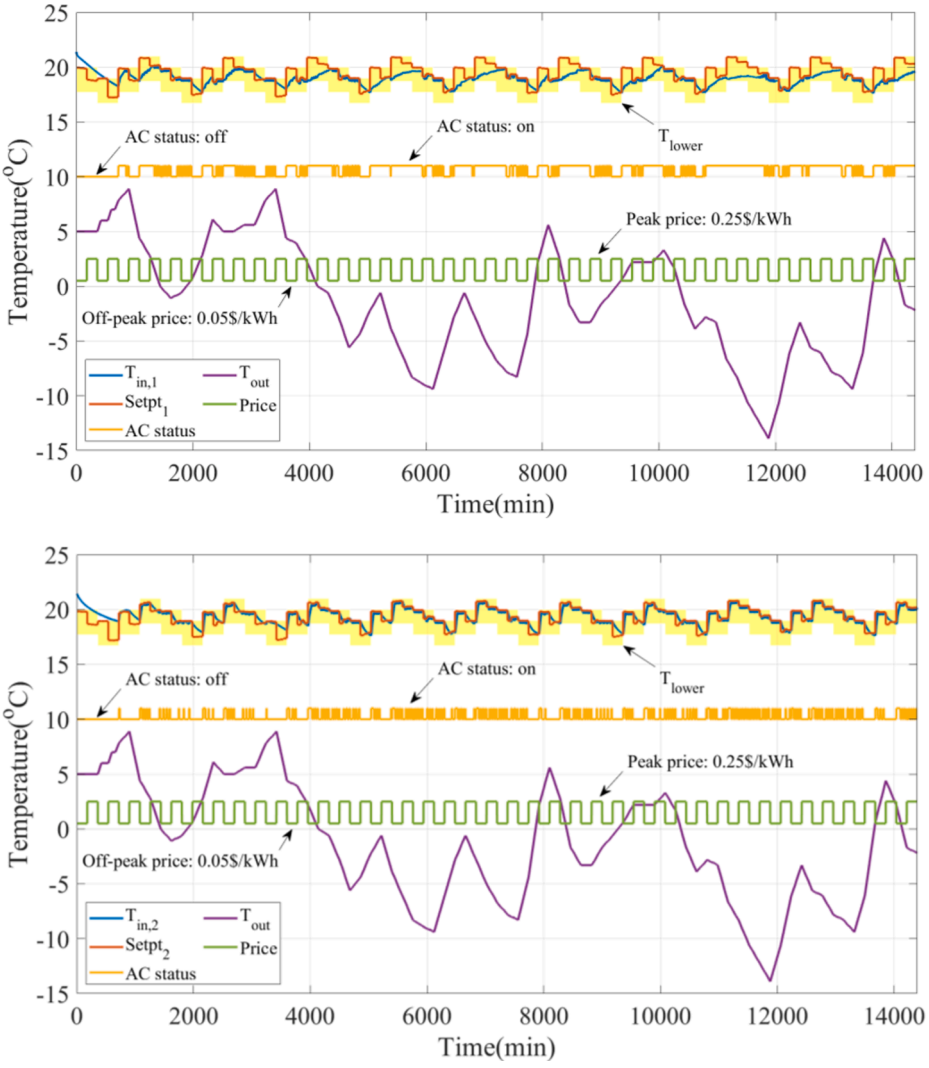

Fig. 4. Setpoint control strategy based on DDPG for 10 test days (top: zone 1; bottom: zone2).

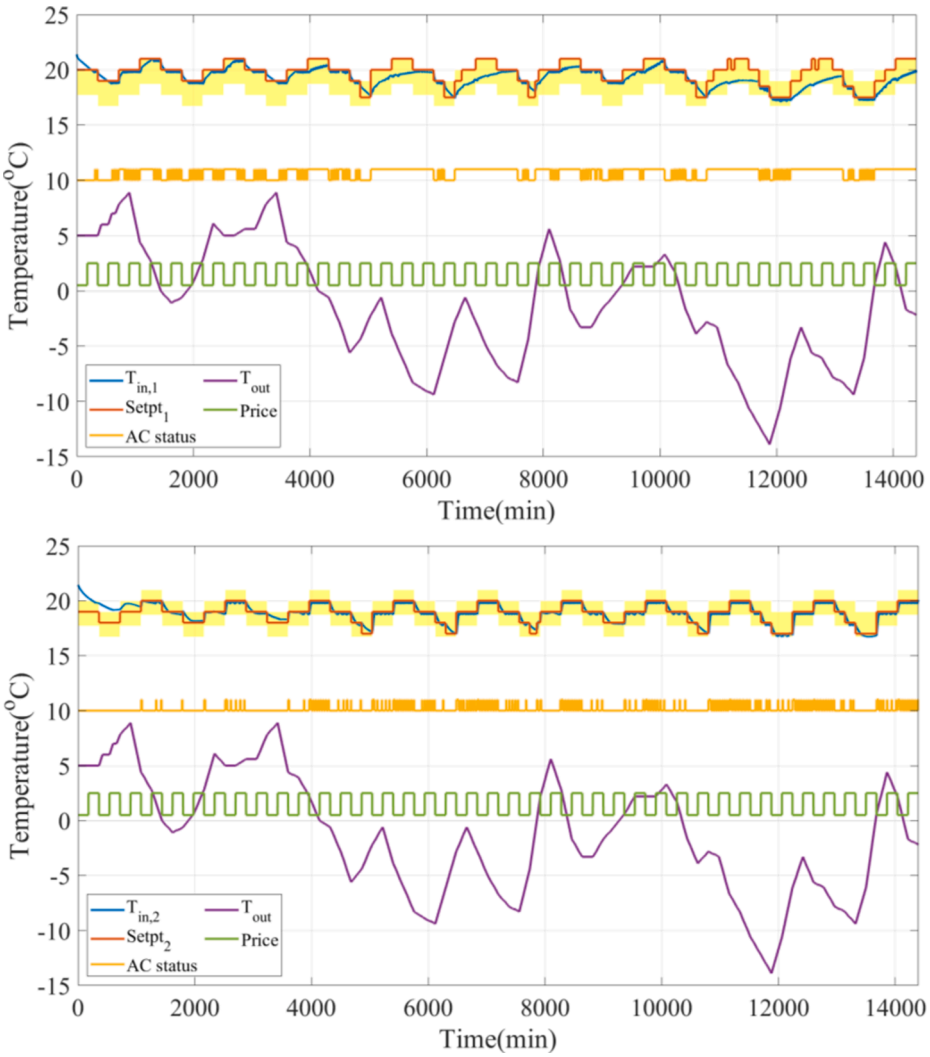

Fig. 5. Setpoint control strategy based on DQN for 10 test days (top: zone 1; bottom: zone2).

In Table 3, the well-trained deep RL agents are applied to generate the HVAC control strategies for the first 10 days in January 2020. The weather conditions of the test days are different from those of the training days, since the outdoor temperature is much lower in January than in November. The total cost in the table refers to the total energy cost over the 10 days, and the temperature violation in the table refers to the total number of minutes that the indoor temperature falls below T_{lower} - T_{th}, as shown by Eq. (3). T_{th} is set to 0.3 °C. The average temperature violation indicates on average by how many degrees the indoor temperature is lower than the setpoint. As shown in the table, the control strategy derived from the DDPG method has both lower energy cost and fewer temperature violations than that of the DQN. With regard to the benchmark cases, in the rule-based case, because the pre-heating logic is applied based on the price structure, it obtained the lowest cost among all four cases. However, by always setting the setpoint to the lowest value at peak price hours, this control strategy results in severe temperature violation. In the fixed setpoint case, since the setpoint is always set at the highest value, there is no temperature violation. However, the energy cost is also the highest among the four cases. The control strategy and the associated indoor temperature in the four cases are further illustrated in Figs. 4–6.

In all the figures, the yellow rectangular area represents the feasible region of the setpoint [T_{lower}, $T_{lower} + 2$ °C]. As can be observed, the setpoint range changes at a daily cycle. In addition, the indoor temperature in zone 1 is lower than that of zone 2, this is because in the building model, zone 1 is at the 1st floor and zone 2 is at the 2nd floor, and the warmer air goes to upper floors.

In Fig. 4, the DDPG RL agent develops a setpoint control strategy such that when the outdoor temperature is relatively high, i.e. in the first 4000 min, the setpoint will be set at the lowest value at the peak price hour, and at the highest value at the off-peak hour, to realize the pre-heating effect and to reduce energy cost, which is similar to the control logic of the rule-based case. When the outdoor temperature is low, i.e., in the last 2000 min, the setpoint is always set at the highest value to avoid the indoor temperature violation. On the contrary, in the rule-based case, the control strategy still follows the price structure even when the outdoor temperature is extremely low, which results in severe indoor temperature violation, as shown in Fig. 6. Such comparisons indicate that after the training, the DDPG RL agent has acquired the knowledge that the price signal and the outdoor temperature have a significant impact on the reward, and it learns to intelligently set the setpoint based on this state information to reach a higher reward value.

The control strategy of the DQN RL agent is shown in Fig. 5. It can be observed that when the outdoor temperature is relatively high, i.e. in the first 4000 min, the setpoint is set at a relatively high value, and it does not follow the change of retail price. When the outdoor temperature is extremely low, i.e., around 12,000 min, the setpoint is set at the lower bound, which results in temperature violation. The DQN RL agent does

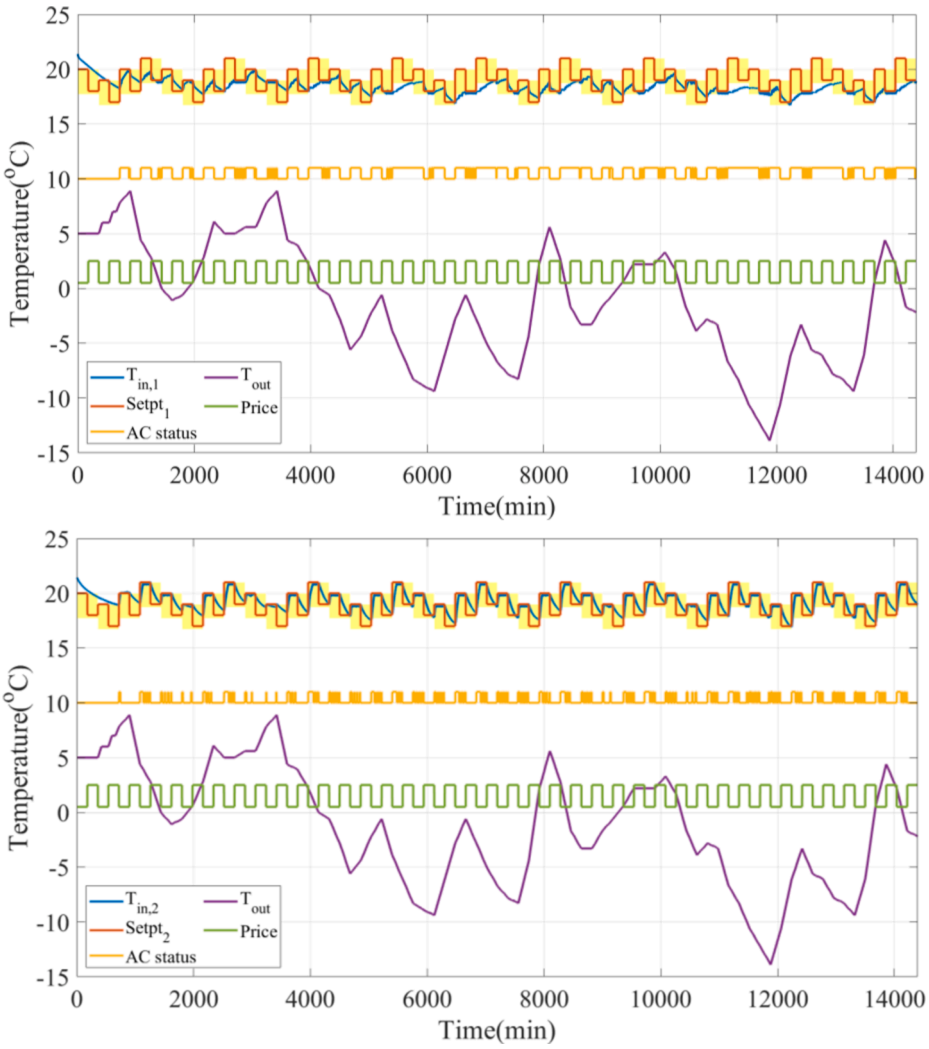

Fig. 6. Setpoint control strategy from the rule-based case for 10 test days (top: zone 1; bottom: zone2).

not successfully capture the impacts of the state variables on the reward function. This can be attributed to the large number of action combinations encountered by the Q network. In such cases, the DQN RL agent has not fully explored all the possible action combinations to maximize the reward, and thus obtains a control strategy with a higher energy cost and more temperature violations.

Finally, in Fig. 7, the fixed setpoint case, since the setpoint is always set at the highest value, the indoor temperature for both zones also remains at the highest level among the four test cases. However, this fixed setpoint case results in the highest energy cost.

4) Generalization of the DDPG Algorithm

a) Extending the DDPG RL agent to different residential buildings

The well-trained DDPG RL agent is further tested in new residential building models with HVAC systems to fully validate its generalization and robustness. Ten building models are generated with different thermal mass parameters, the variation of which follows a normal distribution. The same 10 test days in January 2020 are applied in this case. The energy cost and the temperature violation for the 10 building models under the DDPG control strategy and under the two benchmark cases are compared in Table 4 and Fig. 8. As can be read from the table, similar to the results in Table 3, the rule-based control strategy provides the lowest energy cost, while the fixed setpoint control strategy provides the lowest violation. The well-trained DDPG RL agent can obtain an HVAC control strategy that properly weighs the two objectives, resulting in a relatively lower energy cost and fewer temperature violations for different test building models. Therefore, it can be safely concluded that the DDPG RL agent can flexibly adapt to unseen physical environments and provides an economic HVAC control strategy after its offline training with the fixed environment.

b) DDPG performance under different retail price signals

In the above simulations, a simulated retail price sequence is generated for training and testing the deep RL agent, which is composed of only two price signals. To demonstrate that the well-trained DDPG RL agent has developed high generalization to an unseen environment without additional training, the DDPG RL agent is further tested with a retail price sequence that is generated from the PJM wholesale hourly locational marginal price (LMP) data [36]. The retail price is set as triple of the wholesale market price. The PJM price is very irregular, changing hourly and fluctuating within a large range. The final optimized results of the two deep RL methods and the benchmark case are shown in Table 5:

In Table 5, the fixed setpoint case applies a control strategy where the setpoint is always set at the middle of the setpoint range. This is because the PJM price sequence contains more than just two values, and it cannot be simply divided into two groups as high price and low price. As a result, the setpoint is set at the middle point to avoid possible

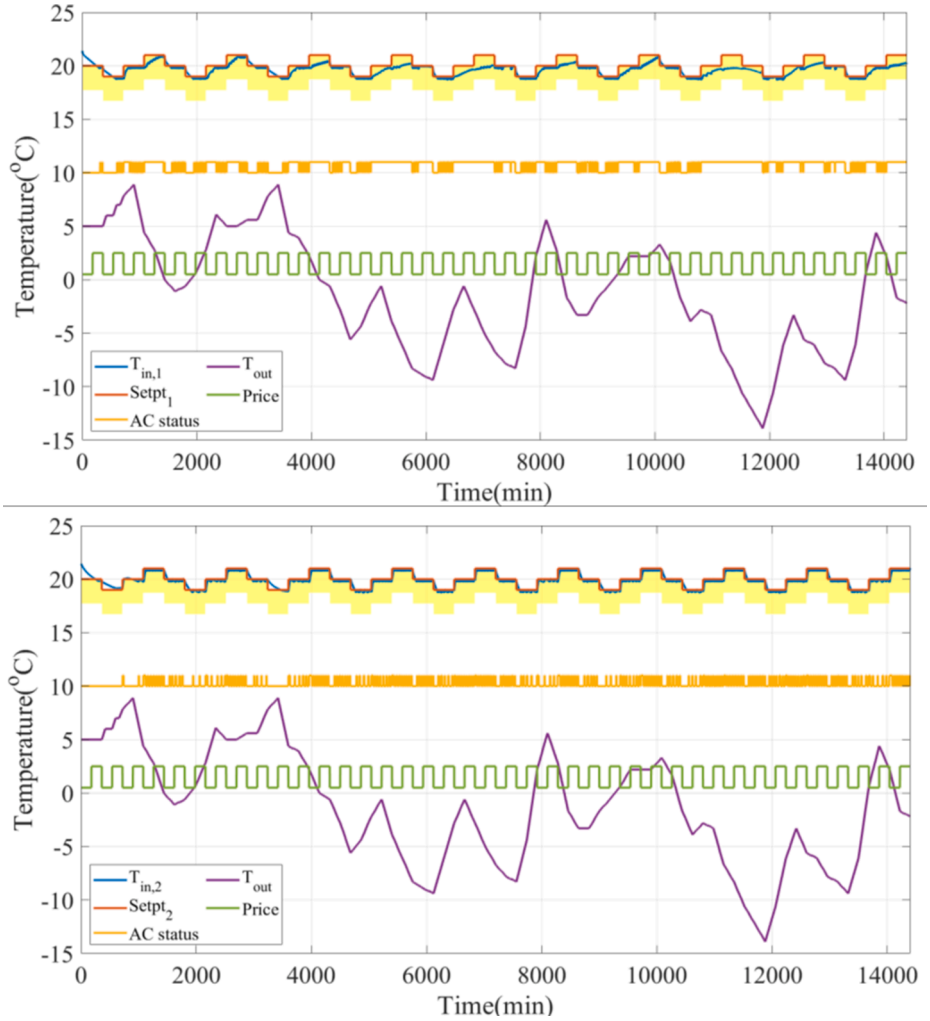

Fig. 7. Setpoint control strategy from the fixed setpoint case for 10 test days (top: zone 1; bottom: zone2).

Table 4
Comparison of optimization results for different building models.

Building index	DDPG		Rule-based		Fixed setpoint	
	Cost ($)	Temperature violation (min)	Cost ($)	Temperature violation (min)	Cost ($)	Temperature violation (min)
1	42.22	31	27.78	1296	57.98	0
2	44.13	41	29.22	1586	60.13	0
3	52.14	45	36.51	2347	68.52	0
4	59.66	101	43.94	3364	75.61	0
5	45.84	41	31.30	1879	62.91	0
6	42.49	39	27.68	1398	59.06	0
7	37.47	24	23.51	1012	53.42	0
8	61.21	81	45.42	3520	76.44	0
9	35.34	25	21.98	818	49.90	0
10	43.19	59	28.41	1323	58.46	0

temperature violations while minimizing the energy cost.

The control strategy and the associated indoor temperature in the three cases are further illustrated by Figs. 9–11. As can be observed in the figure, the PJM price demonstrates a very different pattern from the simulated price sequence. For most of the time the price remains at a relatively low level, with some occasional spikes and fluctuations. However, the well-trained DDPG RL agent still attempts to follow the price tendency, and intelligently sets the setpoint to realize the pre-heating effect. For example, a price spike appears around 12,500 min. The DDPG RL agent catches this sudden change, and lowers the setpoint. At around 13,500 min the retail price sequence demonstrates some fluctuations, and the DDPG RL agent also adjusts the setpoint accordingly. Note that under the price signals that are more time-variants like the PJM market price, it is difficult to develop a simple rule-based control strategy, because the price range is uncertain. However, the well-trained DDPG RL agent can still work intelligently under such an uncertain environment, and it obtains satisfying economic benefits. Therefore, it can be safely concluded that the DDPG algorithm has gained adaptability after training and has potential for real-world online applications.

In Fig. 10, the HVAC control strategy developed by the DQN RL agent also intends to follow the retail price tendency. However, at the price

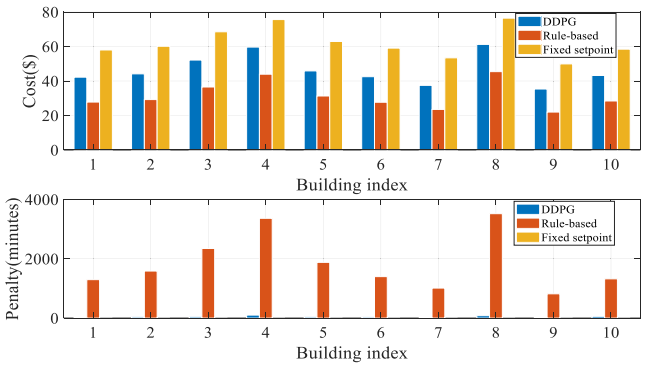

Fig. 8. Comparisons of cost and penalty from three control methods.

Table 5
Test results of different control methods (under PJM price).

Control method	DDPG	DQN	Fixed setpoint
Total cost ($)	32.90	31.80	32.71
Temperature violation (min)	0	222	31
Average temperature violation (°C)	0	1.00	0.27

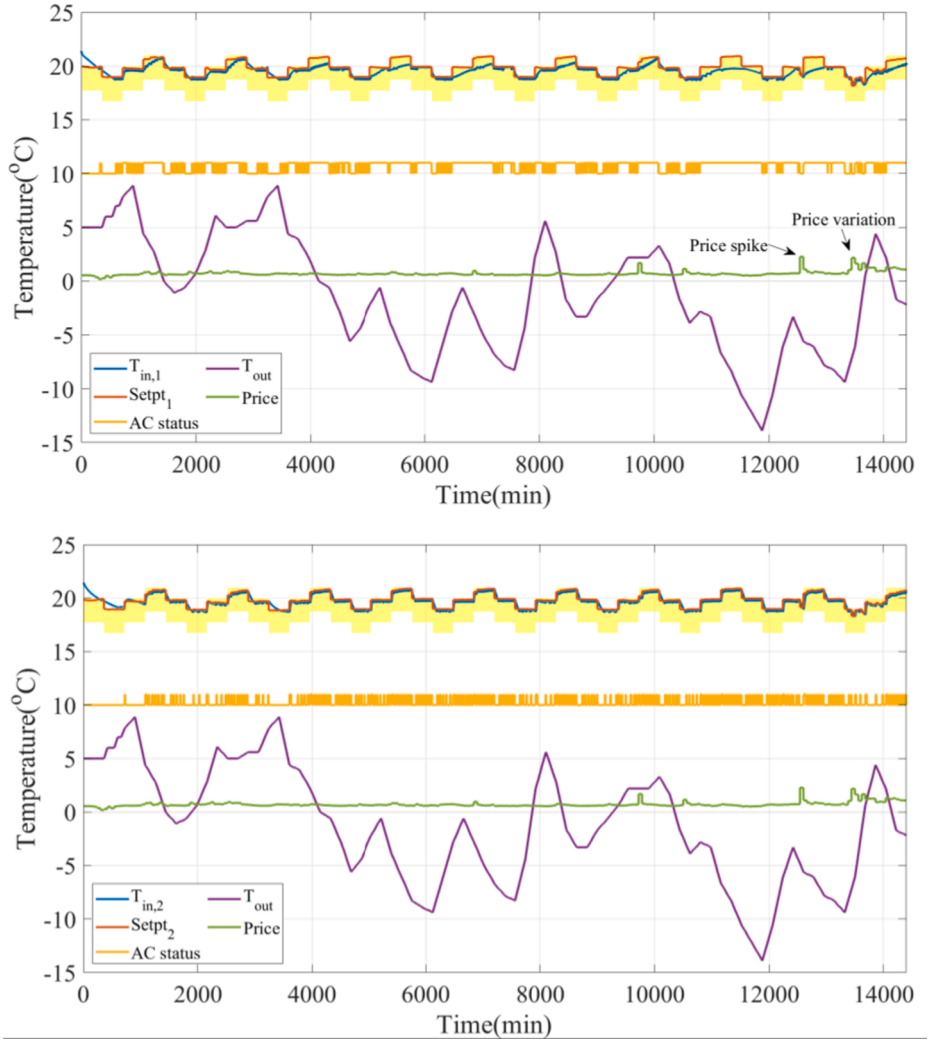

Fig. 9. Setpoint control strategy based on DDPG under PJM price for 10 test days (top: zone 1; bottom: zone2).

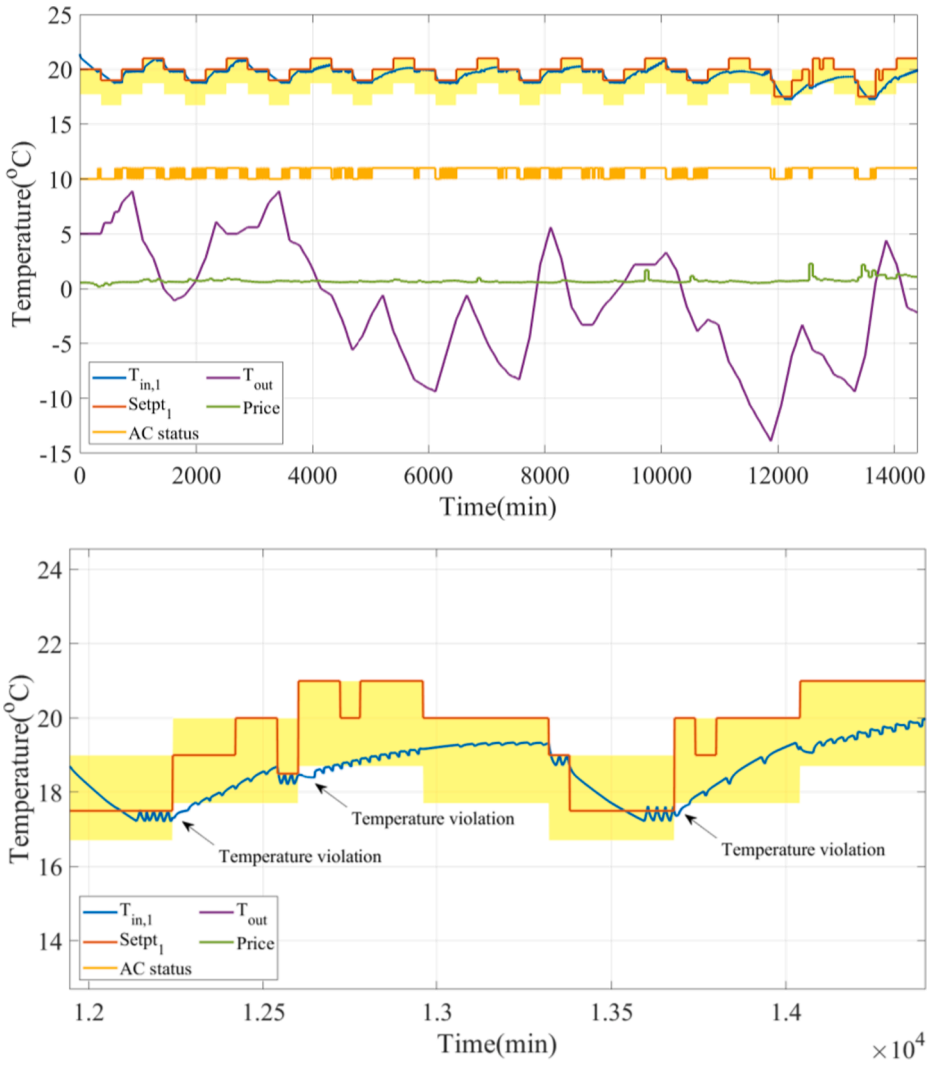

Fig. 10. Setpoint control strategy based on DQN under PJM price for 10 test days (top: zone 1; bottom: zoomed part of zone 1).

spike period (12,500 min) and the price variation period (13,500 min), the DQN RL agent chooses the lowest setpoint values, which results in a temperature violation in zone 1, as shown in the bottom figure.

Finally, in Fig. 11, the fixed setpoint case also leads to some temperature violations in zone 1 when the outdoor temperature is extremely low (after 10,000 min).

5. Conclusion

In this paper, the DDPG RL method is applied for controlling a multi-zone residential HVAC system to minimize the energy consumption cost while maintaining the user comfort. The DDPG can realize continuous control of the HVAC setpoint due to its application of DNNs. Simulation results demonstrate that the well-trained DDPG RL agent can act intelligently to balance the different optimization objectives, and that it also gains generalization and adaptability to unseen environment, which signifies its potential for future online applications in solving MDP problems with hidden information or with continuous search space.

For future works, we will mainly look into two directions for further improving the robustness of the RL-based control strategy: 1) considering different seasoning scenarios, the deep RL agent should learn to automatically switch between different operation modes, i.e. cooling and heating, in order to be applied to a longer control period, i.e. one year, to provide economic control strategies for HVAC users; 2) considering different user preferences, the deep RL agent should be able to learn a more variant setpoint schedule customized by users, and provide more flexible HVAC control strategies. By investigating these two directions, the deep RL agent will become more generalized and robust against uncertainties in real-world operation scenarios.

CRediT authorship contribution statement

Yan Du: Investigation, Methodology, Writing-original draft. **Helia Znadi:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing - review & editing. **Olivera Kotevska:** Investigation, Methodology, Writing - review & editing. **Kuldeep Kurte:** Investigation, Methodology, Writing - review & editing. **Jeffery Munk:** Methodology, Software, Writing - reviewing & editing. **Kadir Amasyali:** Investigation, Methodology, Writing - review & editing. **Evan Mckee:** Investigation, Methodology, Writing - review & editing. **Fangxing Li:** Funding acquisition, Methodology, Project administration, Writing-review & editing.

Declaration of Competing Interest

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

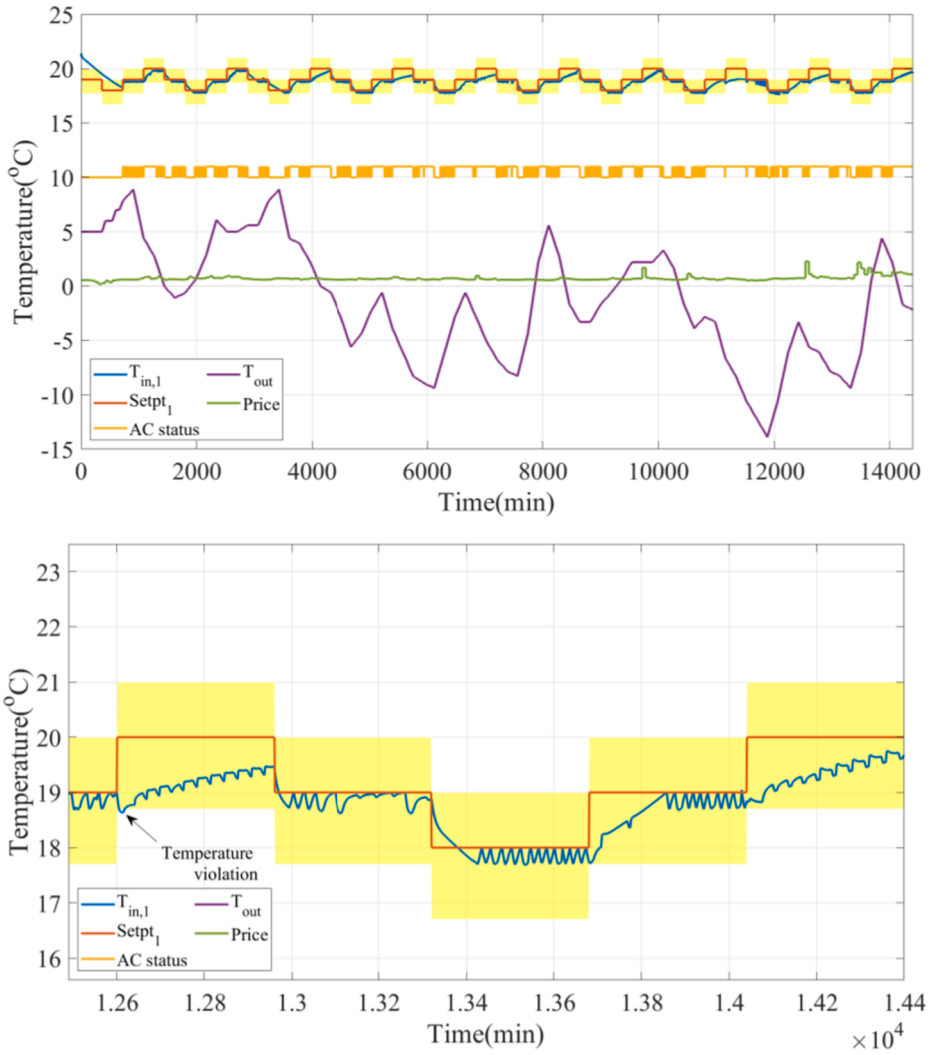

Fig. 11. Setpoint control strategy in the fixed setpoint case under PJM price for 10 test days (top: zone 1; bottom: zoomed part of zone 1).

Acknowledgement

The authors would like to acknowledge the support in part by the U. S. Department of Energy (DOE), including Office of Energy Efficiency and Renewable Energy under the Buildings Technologies Program, in part by CURENT which is an Engineering Research Center (ERC) funded by the U.S. National Science Foundation (NSF) and DOE under the NSF award EEC-1041877, and in part by the U.S. NSF award ECCS-1809458.

References

[1] Pérez-Lombard L, Ortiz J, Pout C. A review on buildings energy consumption information. Energ Build 2008;40:394–8.
[2] Costa A, Keane MM, Torrens JI, Corry E. Building operation and energy performance: monitoring, analysis and optimisation toolkit. Appl Energy 2013; 101:310–6.
[3] Kou X, Li F, Dong J, et al. A scalable and distributed algorithm for managing residential demand response programs using alternating direction method of multipliers (ADMM). IEEE Trans Smart Grid 2020. in-press.
[4] Ma K, Hu G, Spanos CJ. Energy management considering load operations and forecast errors with application to HVAC systems. IEEE Trans Smart Grid 2016;9: 605–14.
[5] Erdinc O, Taşcıkaraoğlu A, Paterakis NG, Eren Y, Catalão JP. End-user comfort oriented day-ahead planning for responsive residential HVAC demand aggregation considering weather forecasts. IEEE Trans Smart Grid 2016;8:362–72.
[6] Wu X, He J, Xu Y, Lu J, Lu N, Wang X. Hierarchical control of residential HVAC units for primary frequency regulation. IEEE Trans Smart Grid 2017;9:3844–56.
[7] Lin Y, Barooah P, Mathieu JL. Ancillary services through demand scheduling and control of commercial buildings. IEEE Trans Power Syst 2016;32:186–97.
[8] Yu L, Jiang T, Zou Y. Online energy management for a sustainable smart home with an HVAC load and random occupancy. IEEE Trans Smart Grid 2017;10:1646–59.
[9] Hao H, Corbin CD, Kalsi K, Pratt RG. Transactive control of commercial buildings for demand response. IEEE Trans Power Syst 2016;32:774–83.
[10] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, et al. Mastering the game of go with deep neural networks and tree search. Nature 2016; 529:484–9.
[11] Li F, Du Y. From AlphaGo to power system AI: what engineers can learn from solving the most complex board game. IEEE Power Energ Mag 2018;16:76–84.
[12] Du Y, Li F, Li J, Zheng T. Achieving 100x acceleration for N-1 contingency screening with uncertain scenarios using deep convolutional neural network. IEEE Trans Power Syst 2019;34:3303–5.
[13] Du Y, Li F. Multi-microgrid energy management based on deep neural network and model-free reinforcement learning. IEEE Trans Smart Grid 2020;11:1066–76.
[14] Huang T, Guo Q, Sun H. A distributed computing platform supporting power system security knowledge discovery based on online simulation. IEEE Trans Smart Grid 2017;8:1513–24.
[15] Wu Y, Tan H, Peng J, Zhang H, He H. Deep reinforcement learning of energy management with continuous control strategy and traffic information for a series-parallel plug-in hybrid electric bus. Appl Energy 2019;247:454–66.
[16] Han X, He H, Wu J, Peng J, Li Y. Energy management based on reinforcement learning with double deep Q-learning for a hybrid electric tracked vehicle. Appl Energy 2019;254:113708.
[17] Hua H, Qin Y, Hao C, Cao J. Optimal energy management strategies for energy Internet via deep reinforcement learning approach. Appl Energy 2019;239: 598–609.
[18] Rocchetta R, Bellani L, Compare M, Zio E, Patelli E. A reinforcement learning framework for optimal operation and maintenance of power grids. Appl Energy 2019;241:291–301.
[19] Kou P, Liang D, Wang C, Wu Z, Gao L. Safe deep reinforcement learning-based constrained optimal control scheme for active distribution networks. Appl Energy 2020;264:114772.

[20] Wei T, Ren S, Zhu Q. Deep reinforcement learning for joint datacenter and HVAC load control in distributed mixed-use buildings. IEEE Trans Sustainable Comput 2019. early access.
[21] Claessens BJ, Vrancx P, Ruelens F. Convolutional neural networks for automatic state-time feature extraction in reinforcement learning applied to residential load control. IEEE Trans Smart Grid 2016;9:3259–69.
[22] Mocanu E, Mocanu DC, Nguyen PH, Liotta A, Webber ME, Gibescu M, et al. On-line building energy optimization using deep reinforcement learning. IEEE Trans Smart Grid 2018;10:3698–708.
[23] Wang Y, Velswamy K, Huang B. A long-short term memory recurrent neural network based reinforcement learning controller for office heating ventilation and air conditioning systems. Processes 2017;5:46–63.
[24] Zhang Z, Chong A, Pan Y, Zhang C, Lam KP. Whole building energy model for HVAC optimal control: a practical framework based on deep reinforcement learning. Energy Build 2019;199:472–90.
[25] Ahn KU, Park CS. Application of deep Q-networks for model-free optimal control balancing between different HVAC systems. Sci Technol Built Environ 2019;26:61–74.
[26] Gao G, Li J, Wen Y. DeepComfort: energy-efficient thermal comfort control in buildings via reinforcement learning. IEEE Internet of Things J 2020 (early access).
[27] Yu L, Sun Y, Xu Z, Shen C, Yue D, Jiang T, et al. Multi-agent deep reinforcement learning for HVAC control in commercial buildings. IEEE Trans Smart Grid 2020 (early access).
[28] Zou Z, Yu X, Ergan S. Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network. Build Environ 2020;168:1–15.
[29] Lu N. An evaluation of the HVAC load potential for providing load balancing service. IEEE Trans Smart Grid 2012;3:1263–70.
[30] Cui B, Joe J, Munk J, Sun J, Kuruganti T. Load flexibility analysis of residential HVAC and water heating and commercial refrigeration. Oak Ridge, TN (United States): Oak Ridge National Lab; 2019.
[31] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature 2015;518:529–33.
[32] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971.
[33] Cui B, Munk J, Jackson R, Fugate D, Starke M. Building thermal model development of typical house in US for virtual storage control of aggregated building loads based on limited available information. In: 30th International Conference on Efficiency, Cost, Optimisation, Simulation and Environmental Impact of Energy Systems. San Diego, California, US; 2017.
[34] Clean Power Research. [online]: https://www.cleanpower.com/.
[35] TensorFlow. [online]: https://www.tensorflow.org/.
[36] PJM market. [online]: https://www.pjm.com/.