frontiers in Energy Research

ORIGINAL RESEARCH
published: 02 February 2021
doi: 10.3389/fenrg.2020.610518

Hybrid-Model-Based Deep Reinforcement Learning for Heating, Ventilation, and Air-Conditioning Control

Huan Zhao[1], *Junhua Zhao*[1,2]*, *Ting Shu*[3] *and Zibin Pan*[1]

[1]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China, [2]Shenzhen Research Institute of Big Data, Shenzhen, China, [3]Guangdong-Hongkong-Macao Greater Bay Area Weather Research Center for Monitoring Warning and Forecasting, Shenzhen, China

Buildings account for a large proportion of the total energy consumption in many countries and almost half of the energy consumption is caused by the Heating, Ventilation, and air-conditioning (HVAC) systems. The model predictive control of HVAC is a complex task due to the dynamic property of the system and environment, such as temperature and electricity price. Deep reinforcement learning (DRL) is a model-free method that utilizes the "trial and error" mechanism to learn the optimal policy. However, the learning efficiency and learning cost are the main obstacles of the DRL method to practice. To overcome this problem, the hybrid-model-based DRL method is proposed for the HVAC control problem. Firstly, a specific MDPs is defined by considering the energy cost, temperature violation, and action violation. Then the hybrid-model-based DRL method is proposed, which utilizes both the knowledge-driven model and the data-driven model during the whole learning process. Finally, the protection mechanism and adjusting reward methods are used to further reduce the learning cost. The proposed method is tested in a simulation environment using the Australian Energy Market Operator (AEMO) electricity price data and New South Wales temperature data. Simulation results show that 1) the DRL method can reduce the energy cost while maintaining the temperature satisfactory compared to the short term MPC method; 2) the proposed method improves the learning efficiency and reduces the learning cost during the learning process compared to the model-free method.

Keywords: deep reinforcement learning, model-based reinforcement learning, hybrid model, heating, ventilation, and air-conditioning control, deep deterministic policy gradient

OPEN ACCESS

Edited by:
Yan Xu,
Nanyang Technological University, Singapore

Reviewed by:
Guibin Wang,
Shenzhen University, China
Taskin Jamal,
Ahsanullah University of Science and Technology, Bangladesh
Ke Meng,
University of New South Wales, Australia
Mir Nahidul Ambia,
Other, Australia

***Correspondence:**
Junhua Zhao
zhaojunhua@cuhk.edu.cn

Specialty section:
This article was submitted to Smart Grids, a section of the journal Frontiers in Energy Research

Received: *26 September 2020*
Accepted: *28 December 2020*
Published: *02 February 2021*

Citation:
Zhao H, Zhao J, Shu T and Pan Z (2021) Hybrid-Model-Based Deep Reinforcement Learning for Heating, Ventilation, and Air-Conditioning Control. Front. Energy Res. 8:610518. doi: 10.3389/fenrg.2020.610518

INTRODUCTION

Improving the energy efficiency of commercial buildings is a critical task in many countries for energy-saving, cost-saving, and environmental protection (Paone and Bacher 2018). The target of the Heating, Ventilation, and Air-conditioning (HVAC) system is to minimize the energy/CO2 consumption while maintaining users' comfort and the HVAC system is the major energy consumer in the building (Belic et al., 2015). Improving the efficiency of the HVAC system contributes to greater energy savings within the building (Zhao et al., 2009). Therefore, balancing the indoor satisfaction of users and energy consumption is a critical issue.

The HVAC control is complex due to the cooperation of subsystems in the system and the thermal dynamic of buildings. Researchers in past years majorly focused on the model predictive control methods, which use the model of system and process to obtain control signal with certain objections and constraints (Afram and Janabi-Sharifi 2014; Xie et al., 2018; Afram et al., 2017; Gomez-Romero et al., 2019). The advantage of MPC is its flexibility and ability to consider different kinds of constraints. However, the performance of MPC methods majorly influenced by the accuracy and complexity of the model. The control efficiency of the MPC method performs unsatisfactorily under complex building thermal dynamics (Amasyali and El-Gohary, 2018). The temperature dynamic is hard to track for the system may change in various conditions. For example, the energy consumption prediction model is hard to predict, which is related by many factors such as weather conditions, occupancy schedule, thermal properties of materials, etc. This promotes the idea to use the model-free method in practice.

Reinforcement learning (RL) (Sutton and Barto 2018) is a model-free method and the agent utilizes "trial and error" to learn the optimal policy without the requirement of the system and process prior knowledge. Compared to MPC methods, RL shows the ability to obtain the building dynamic in real-time, less environmental parameters required, and more efficiently control results. Most of the existing papers prefer to use Q-learning for HVAC control. Fazenda et al. (2014) verified the Q-learning based method for the bang-bang heater and setpoint heater in the HVAC system considering both tenant and thermal zone changes. Barrett and Linder (2015) comprise the Bayesian approach to model room occupancy and the Q-learning method to learn a control policy for the thermostat unit. Vázquez-Canteli et al. (2017) proposed batch Q-learning for the heat pump control. Chen et al. (2018) proposed a Q-learning method to control both window and HVAC systems, which trying to fully utilize natural ventilation and coordinate its operation with the HVAC system.

Deep reinforcement learning (DRL) (Mnih et al., 2015) improves the RL with deep learning method and utilize deep neural networks to approximate the value function and policy function. The DRL largely extends the ability of RL to the larger state-action space in many different areas (Zhang et al., 2018; Zheng et al., 2018; Ye et al., 2019; Yan and Xu 2020; Yan and Xu 2019). Inspired by the advantage of the quick on-line decision-making process in the large-scale solution space, many methods are applied to HVAC control in recent years. In (Wei et al., 2017), Wei et al. proposed a Deep Q-Network based method to reduce energy cost while maintaining room temperature within the desired range. Gao et al. (2019) extended the problem to a continuous action space with the deep deterministic policy gradient (DDPG) method. In Yu et al. (2020a), proposed a multi-agent DRL method with the attention mechanism to minimize energy cost in a multi-zone building. In building energy. Zou et al (2020) applied DDPG in the data-based Long Short Term Memory (LSTM) environment model. Ding et al (2019) proposed a Branching Dueling Double Q-Network to solve the high dimensional action problem for four building subsystems, including the HVAC, lighting, blind, and window system. In Yu et al. (2020b), utilized DDPG to minimize the energy cost of a smart home with the HVAC and energy storage system. We may find a trend to solve the problem of the HVAC subsystem and cooperate with other systems.

In practice, these methods may take a long time to converge to a stable policy in the HVAC control problem and cause unpredictable learning costs during the learning process. A naive way is to train the agent in the simulator first and then apply it to the real environment. For example, in Zhang et al. (2019), a practical framework is proposed with the Asynchronous Advantage Actor-critic (A3C) algorithm that the agent learning in the simulator first and then deploy to the real environment. However, the agent may overfit the simulator and lead to unsatisfactory performance. In the sub-area of RL, the traditional way is to improve data efficiency by the model, which is called model-based RL. Model-based RL uses the log data to create and update the environment model (Sutton, 1990), which the agent can freely and unlimitedly interact with. The PILCO (Deisenroth and Rasmussen, 2011) employs non-parametric probabilistic Gaussian processes for the dynamic model. The PETS (Chua et al., 2018) combines uncertainty-aware deep network dynamics models with sampling-based uncertainty propagation. Based on these ideas, we proposed a hybrid model-based RL framework for the HVAC control problem.

The contributions of this paper are summarized as follows. Firstly, we formulate the HVAC control problem to a specific MDPs that the reward function contains energy cost, temperature violation, and action violation. The continuous constraint action space is considered in the paper. Secondly, a hybrid-model-based DRL (HMB-DRL) framework is proposed for HVAC control, which utilizes the knowledge-driven model in the pre-training process. Also, the knowledge-driven model and data-driven model are both utilized in the online learning process. The HMB-DDPG algorithm is proposed based on the framework, which can increase the training efficiency and reduce low learning cost periods comparing to DDPG. Lastly, the protection mechanism and adjusting reward methods are proposed. The protection mechanism utilizes the knowledge-driven model to avoid low reward action during the online learning process, and adjusting reward changes the parameter value of the action violation item to accelerate the learning process between the pre-training process and the online learning process.

MATERIALS AND METHODS

Problem Formulation

The target of the HVAC system is to minimize the energy cost and keep the zone temperature within the comfort range. We consider the HVAC system with Variable Air Volume (VAV) unit and constant air temperature supply. The zone temperature in the next time step is decided by current zone temperature, ambient temperature, and HVAC system (2)–(5). The power model P_f of VAV with fan efficiency k_f is defined by (2), where $f_{z,t}$ (kg/s) is the airflow rate at zone z time t, $\sum_{z=1}^{N} f_{z,t}$ is the total airflow rate into all zones at time t, and z is the zone index which is from 1 to N. The internal and external heat gains and losses cause the

temperature increase or decrease in the zone. Suppose the effect of all walls on the zone temperature are the same, the modified zone model in a discontinuous time step is given by the rate of change of heat in the walls $H_{w,z,t}$ (kW) and in the HVAC system $H_{h,z,t}$ (kW) in zone z at time t (3)–(5), where U_w (kW/m^2C^o) is the heat transfer coefficient of walls, A_w (m^2) is the total area of walls, $T_{a,\,t}$ (C^o) is the ambient temperature at time t, and $T_{z,\,t}$ (C^o) is the zone temperature at time t. The C_a (kJ/$kg\ C^o$) is the specific heat of air, $f_{z,t}$ is the control variable in **Eq. 2**, and T_s is the supply air temperature, which is a constant. The Δt is the time interval, V_z (m^3) is the volume of the zone, and ρ_a (kg/m^3) is the density of air. Therefore, the target of the HVAC system to minimize the long-term energy cost is as follows,

$$\min_{a_{z,t}} E\left(\sum_{t=1}^{T}\gamma^t P_{f,t}\lambda_t\right) \tag{1}$$

$$s.t.\ \ P_{f,t} = k_f\left(\sum_{z=1}^{N} f_{z,t}\right)^2 \tag{2}$$

$$H_{w,z,t} = U_w A_w\left(T_{a,t} - T_{z,t}\right) \tag{3}$$

$$H_{h,z,t} = f_{z,t}C_a\left(T_s - T_{z,t}\right) \tag{4}$$

$$T_{z,t+1} - T_{z,t} = \Delta t\left(H_{w,z,t} + H_{h,z,t}\right)/\left(V_z\rho_a C_a\right) \tag{5}$$

$$T_{zone}^{min} < T_{z,t} < T_{zone}^{max} \tag{6}$$

$$T_{a,t+1} = F_a\left(T_{a,t}\right) \tag{7}$$

$$\lambda_{t+1} = F_\lambda\left(\lambda_t\right) \tag{8}$$

$$f_{z,t} = a_{z,t} * f_z^{max};\ a_{z,t} \in [0,1] \tag{9}$$

where γ is the discount factor, λ_t is the electricity price at time t, T_{zone}^{max} and T_{zone}^{min} are the maximum and minimum acceptable zone temperature respectively, f_z^{max} is the maximum airflow rate and $a_{z,t}$ is the continuous control variable at time t, which is between [0, 1]. F_a and F_λ are the dynamic model of ambient temperature and electricity price. With known F_a and F_λ, this problem can be solved by dynamic programming. However, the ambient temperature and electricity price dynamics are hard to accurately model and the HVAC system parameters may be inaccurate. Also, the constraint and continuous action space make the problem hard to calculate. All these factors cause an unsatisfactory result.

As a model-free method, the RL perfectly solve the above challenges. The agent interacts with an environment and iteratively improves the policy without requiring the knowledge of the environment. The first step is to reformulate the problem as a Markov Decision Processes (MDPs). The key components of MDPs are reformulated as follows.

State: The HVAC system controls the air volume of multiple zones in the building. Considering the target of the system, the following observations are chosen as the state, including all zones temperature $T_{z,t}|\forall z \in N$, ambient temperature $T_{a,t}$, electricity price λ_t, and time index in the day t', i.e., $s_t = \left(T_{z,t}|\forall z \in N, T_{a,t}, \lambda_t, t'\right)$

Action: The action is defined as the power percentage of the VAV units in all zones $a_{z,t}$, which is a continuous action space and should between [0, 1]. When the action is defined as a continuous action space, many new issues meet. The most serious one is how to constraint the action in the resalable range during the learning process. For example, the action $a_{z,\,t}$ can only between [0, 1] in this framework and any real number out of this range cannot be executed physically. The general approach is mapping the out of range action value to the feasible range during the learning process and give punishment according to the level of the violation.

Reward: The target of the HVAC system is to minimize the energy cost and keep the zone temperature within the comfort range. A normal way to deal with constraint is to use the violation as a punishment in the reward function. Combining the action constraint, the reward function contains three components. The first item is the energy cost of the HVAC system C_1, i.e., $C_{1,t} = P_{f,t}\lambda_t$. The second item is the violation of zone temperature C_2, i.e., $C_{2,t} = \sum_{z=1}^{N}\left([T_{z,t} - T_{zone}^{max}]^+ + [T_{zone}^{min} - T_{z,t}]^+\right)$. The $[x]^+$ means to choose the larger value of x and 0, i.e. $max(x, 0)$. The third item is the violation of action C_3, i.e. $C_{3,t} = \sum_{z=1}^{N}\left([a_{z,t} - 1]^+ + [0 - a_{z,t}]^+\right)$. Using different coefficient α to represent the importance, the reward R is equal to the negative weighted sum of these three items, i.e. $R_t = -C_{1,t} - \alpha_1 C_{2,t} - \alpha_2 C_{3,t}$.

Hybrid-Model-Based Reinforcement Learning Framework

The hybrid-model-based RL framework is designed for the HVAC control and the main periods can be divided into the pre-training process and the online learning process. The pre-training process is very necessary and important in practice. The target of pre-training is to obtain a basic agent with a certain ability. The pre-training process can be divided into two main classes. One tendency is to learn the initial policy through imitation learning or supervised learning. These methods generally assume learning object's policy is optimal and the reward function is unknown. The other tendency is to build a 'world model' to simulate the environment and the RL agent can freely interact with the model to learn the basic policy. Then, the agent applies the basic policy in the real environment and learns the optimal policy. In the HVAC control problem, the electricity price dynamic is hard to capture, yet the physical model of building thermal can be easily built. Also, solving the optimal control result in all conditions is time-consuming. The latter one is chosen as the pre-training process.

Although the pre-training may obtain a close optimal policy, the online learning process is still necessary since the model cannot perfectly simulate the environment. To accelerate the learning process, model-based reinforcement learning methods are often used. However, the data-driven model may be inaccurate in some range due to the lack of previous data and lead to a high-cost action. The good generalization of the knowledge-driven model can solve the problem well. Here, we suppose there exists an explicit expression of the real environment and the expression is M_r. Due to measurement error, estimation error, and unknown relationship, the knowledge-based model of the real environment is built as M_k, which is a trustworthy baseline model. As the online data size increase, the data-driven model is built as M_d.

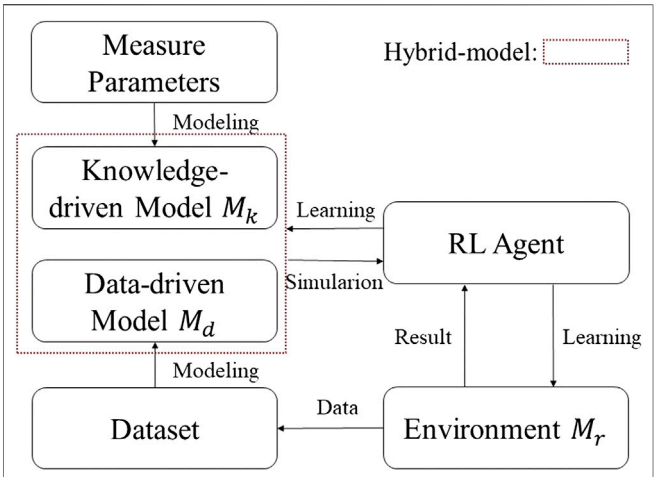

FIGURE 1 | Hybrid-model-based RL framework.

Based on the above ideas, the hybrid-model-based reinforcement learning (HMB-RL) framework is proposed and shown in **Figure 1**. The first period is the pre-training process. The knowledge-driven model is built based on the measurement parameters and the RL agent interacts with the knowledge-driven model to learn the basic policy. Then, in the second period, the agent starts the online learning process from the basic policy. Similar to the model-based RL, the data-driven model is built when the online dataset size is large enough. Between two episodes, the agent can interact with the data-driven model to do learning. The knowledge-driven model still provides simulation results, but the results are no longer used for learning. The knowledge-driven model simulation results are used to ensure the low learning cost of the RL action.

This framework is similar to the framework in Zhang et al. (2019), which also focuses on the practice of RL methods in the HVAC problem. It is worth mention that our proposed framework utilizes model-based RL in the online learning process. And both data-driven model and knowledge-driven model are used to accelerate the learning process and reduce high learning costs. This is why we call the framework as the hybrid-model-based RL framework.

Hybrid-Model-Based Deep Deterministic Policy Gradient Method

Although the agent learns the basic policy in the pre-training process using the reliable knowledge-driven model, the performance of the pre-trained policy in the environment is still uncertain. Therefore, the pre-trained policy is not suitable to directly apply to the real environment without the online learning process. Since the pre-training process is executed in the knowledge-driven model, the learning cost can be regarded as zero if we ignore the computational cost. On the other hand, in the learning period, the agent needs to consider the learning efficiency and learning cost.

Based on the HMB-RL framework and DDPG algorithm, the hybrid-model-based DDPG (HMB-DDPG) algorithm is proposed and shown in **Algorithm 1**. After initializing all the networks and replay buffer, the agent interacts with the environment to do online learning. The RL action $a_{rl,t}$ is generated according to the current state and a small noisy N_t. The protector (PM) generates the execution action $a_{e,t}$ using knowledge-driven model M_k, threshold $\overline{R}_t$, $a_{rl,t}$, and s_t. The detailed information is introduced as the protection mechanism in the next section. Then the action decided by the protection mechanism is executed in the environment. The agent observes the reward r_t and next state s_{t+1} to save the information into the replay buffer. After that, just like **Algorithm 1**, the data-driven model is updated and the agent interacts with the model to update the Q network and policy network. At the end of the episode, the target Q network and target policy network are updated.

Protection Mechanism and Reward Adjusting

In the online learning process of the real environment, judging the performance of action before executing the action is a direct way to avoid the high learning costs. The model-free RL constraint the policy divergence between each update to avoid the high changing rate of policy updating. However, these methods focus on safe exploration instead of low learning costs. On the other side, the model-based RL can directly utilize the model to accelerate the learning process and partly achieve this target. The normal model-based RL utilizes data-driven models and the accuracy of the data-driven model highly depends on the training data. Therefore, the data-driven model may be inaccurate sometimes and not stable to be the referee. In previous work, the knowledge-driven model was utilized as both protector and simulator in the wind farm control problem and provide simulation results during the learning process (Zhao et al., 2020). In that research, the knowledge-driven model contributed more to the beginning of the learning process. However, in the proposed framework, the pre-training process exhausts the potential of the knowledge-driven model as a simulator. The only contribution of the knowledge-driven model in the online process is to work as a protector.

Figure 2 shows the protection mechanism using the knowledge-driven model. Whenever an RL action needs to be executed in the real environment, the knowledge-driven model predicts the reward for the action. If the predicted reward is acceptable, i.e., $M_k(a_{rl,t}, s_t) > \overline{R}_t$, the RL action is executed in the real environment. If the predicted reward is not acceptable, the MPC result is calculated in the knowledge-driven model, i.e., $a_{MPC,t} = argmax M_k(a, s_t)$. Then the MPC action is combined with $a_{rl,t}$ to generate an acceptable action for the agent. Therefore, the worst case in the learning process can be limited. The mechanism is concluded as **Eq. 10**.

$$a_{e,t} = \begin{cases} a_{rl,t}, & If\ M_k(a_{rl,t}, s_t) > k * \overline{R}_t \\ \beta * a_{rl,t} + (1-\beta) a_{MPC,t}, & else \end{cases} \tag{10}$$

Algorithm 1 Hybrid-model-based deep deterministic policy gradient algorithm.

Input: Knowledge-driven Model M_k, threshold $\overline{R}_t$
1: Initialize Q network Q and policy network μ
2: Initialize target network Q' and μ' with the same weights
3: Initialize replay buffer RB
4: ***For*** *episode = 1,. . ., M* ***do***
5: Receive initial observation state s_0
6: ***For*** *t = 1,. . ., T* ***do***
7: Selection RL action $a_{rl,t} = \mu(s_t|\theta^{\mu}) + N_t$
8: $a_{e,t} = PM(\overline{R}_t, M_k(a_{rl,t}, s_t))$
9: Execute $a_{e,t}$, observe reward r_t and next state s_{t+1}
10: Store transition $(s_t, a_{e,t}, r_t, s_{t+1})$ in RB
11: Update Data-driven model M_k with RB
12: ***For*** *repeat times = 1,. . .,N* ***do***
13: Update Q using the sampling data from M_k
14: Update μ using the sampled policy gradient
15: Update the target networks Q' and μ'

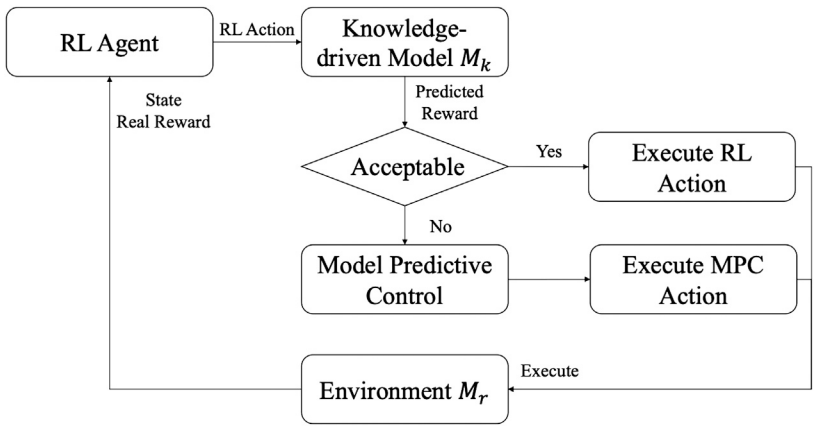

FIGURE 2 | Protection mechanism.

Solving MPC control results in an accurate knowledge-driven model is trustworthy but time-consuming. This is one of the main reasons for using model-free methods. For the proposed method, it is necessary to choose a low-fidelity knowledge-driven model that can be solved quickly. Although the result is not as accurate as the high-fidelity knowledge-driven model, it is sufficient to provide protection information for the RL agent.

Setting the action violation as a punishment in the reward function makes the agent approaching the available action range at the beginning of the learning process. Adding the penalties prevent the agent from falling into the local optimal outside the range of available action. However, this makes the agent being afraid to take action near the boundary of the available action range and reduce the RL's performance. The value of the reward function parameter influences the final result of the learned policy. Generally, the smaller value of the penalty parameter, the less important the violation, and the greater possibility of violating this item. On the other hand, the importance of other items has increased relatively. Separate the learning process into two periods also allows the framework to adjust the reward function between the two periods. Since the agent have learned the basic policy to stay within the available range, the weight of the action violation item can be reduced or removed during the online learning process, i.e. $R_{t,\,new} = -C_{1,t} - \alpha_1 C_{2,t} - \alpha_{2,\,new} C_{3,t},\ 0 \le \alpha_{2,new} \le \alpha_{2,old}$.

TABLE 1 | Environment parameters.

Symbol	Quantity	Symbol	Quantity
A_w	1000 m^2	ρ_a	1.25 kg/m^3
C_a	1.005 kJ/kgC^o	Δt	1 s
k_f	1.675 kWs^2/kg^2	f_z^{max}	0.45 kg/s
U_w	10^{-3} kW/m^2C^o	T_{zone}^{min}	19 C^o
T_s	16 or 30 C^o	T_{zone}^{max}	25 C^o

SETUPS

Heating, Ventilation, and Air-Conditioning System and Zone Model

In this paper, a simulation building (25 m*25 m*10 m) with a VAV system and fixed strategy cooling/heating system is implemented to test the proposed method. To investigate the effects in the simulation environment, we maintain two versions of the HVAC systems, zone models, and dynamic models. The

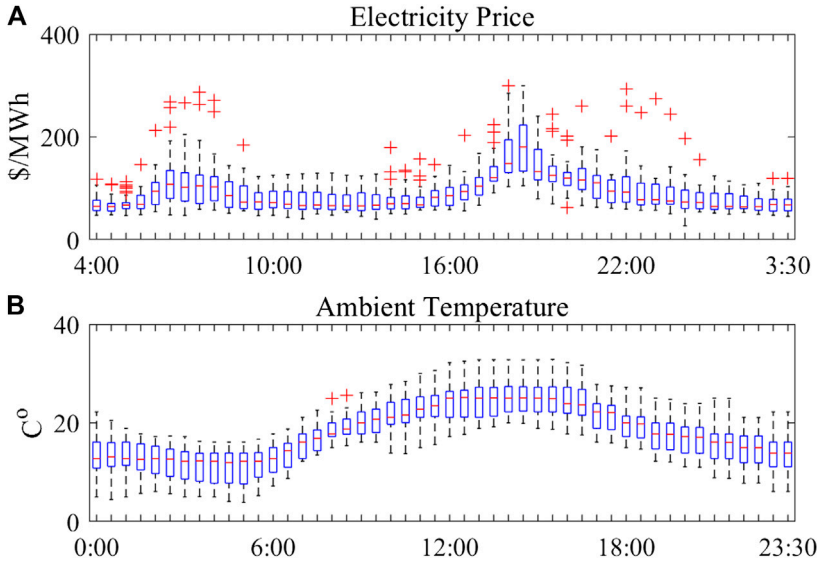

FIGURE 3 | Distribution of **(A)** electricity price data **(B)** ambient temperature data.

first version is built as the real environment M_r and not available to know in practice. The second version is built to represent the knowledge-driven model M_k. M_k is an inaccurate translation of the M_r.

The setup of the HVAC system and zone model of M_r is summarized in **Table 1**. In the test, we suppose the degradation of M_r to M_k is majorly caused by an unknown dynamic of electricity price and ambient temperature. The setup of the dynamic model of M_r and M_k is introduced in the next part. The only difference between M_r and M_k in the HVAC system and zone model is the k_f. In the M_k, the value of k_f is 1.5.

The discount factor is related to how much the agent cares about the future. The range of the discount factor is from 0 to 1. The larger the value, the greater the future impact. In the test, we set the discount factor as 0.9.

Electricity Price and Ambient Temperature

To simulate the electricity price and ambient temperature dynamics, the day-ahead electricity market price data from the Australian Energy Market Operator (AEMO) and the ambient temperature in New South Wales are utilized. The interval of these data is half an hour and the data from 2018/09/01 to 2018/09/30 are chosen. The distribution of electricity price and ambient temperature are shown in **Figure 3**. The red '+' in the box plot is the outlier of each period.

The electricity prices in the data follow a non-Gaussian distribution and the price fluctuations in the data $\Delta\lambda^D_{t,d}$ of all periods t in all days d is calculated to simulate the environment dynamic M_r, i.e., $\Delta\lambda^D_{t,d}\lambda^D_{t,d} = \lambda^D_{t+1,d} - \lambda^D_{t,d}$. The $(.)^D$ is the value of data. For the knowledge-driven model M_k, detailed information is ignored. The change expectations of all periods $E(\Delta\lambda^D_t)$ are the model prediction outputs, i.e., $E(\Delta\lambda^D_t) = \frac{\sum_{d=1}^{30}\Delta\lambda^D_{t,d}}{30}, t = 1, \ldots, 48$. The electricity price prediction in the knowledge-driven model is $\lambda_{t+1} = \lambda_t + E(\Delta\lambda^D_t)$. For the environment model, detailed information is included. The price update in the knowledge-driven model is $\lambda_{t+1} = \lambda_t + \Delta\lambda^D_{t,d}$ and d is randomly selected from all days with uniform distribution. To ensure the stability of the process, the maximum price and minimum price of each period are used to constraint the generated result. The same method is applied to generate the ambient temperature using the temperature data.

RESULTS

Optimal Results

To verify the effectiveness of the proposed method, we design a testing process in the simulation environment that uses the previous setups. Firstly, the agent is pre-trained by interacting with the knowledge-driven model M_k for 2,880 iterations (60 days). The pre-trained policy is tested in the environment model M_r for 480 iterations (10 days). Then, to simulate the online learning process, the agent interacts with the M_r using the proposed algorithm and methods. The online learning process also has 2,880 iterations. After that, the final policy is tested with the MPC method in the M_r for 480 iterations. The MPC method solves the one-period optimal result of **Eqs. 1–9** with M_k. Since the data-driven model is hard to be trained well in such a short number of iterations, the replay buff is used to simulate an accurate data-driven model M_d. This testing process is repeated ten times to ensure generalization.

The total energy cost and temperature violation are two main aspects of a good HVAC control policy. The energy cost and temperature violations are as low as possible. To compare the performance of the RL method and the traditional MPC method, the average energy costs and temperature violations are compared. On the other hand, the target of RL is to utilize

TABLE 2 | Testing results conclusion.

	Pre-trained policy	Final policy	MPC
Average energy cost	219.5155	247.8671	339.5064
Average energy cost (peak periods)	369.3853	431.5970	636.3084
Average temperature violation	0.2095	0.0802	0.0032
Average temperature violation (peak periods)	0.5744	0.1619	0.0082

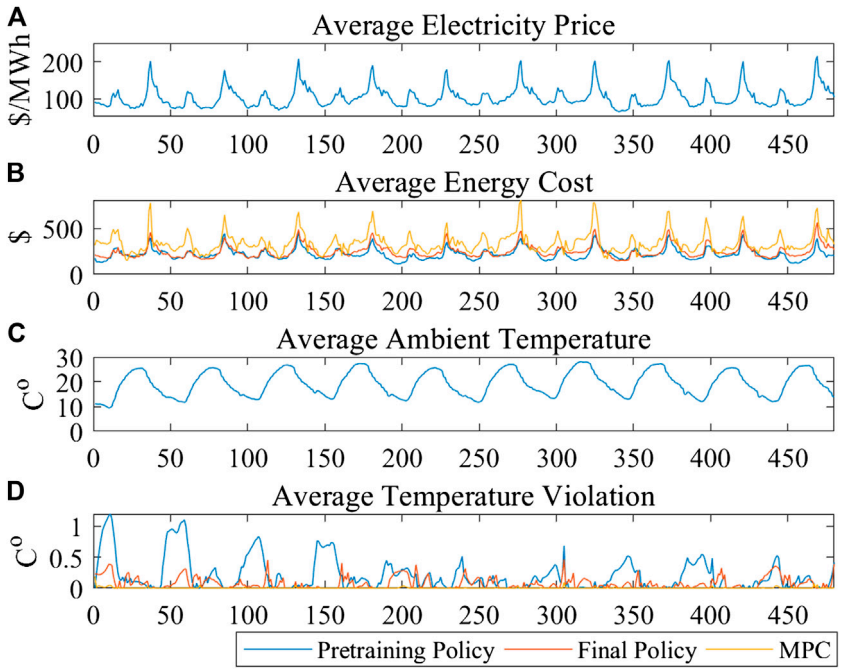

FIGURE 4 | The profile of average testing results **(A)** average electricity price **(B)** average energy cost **(C)** average ambient temperature **(D)** average temperature violation.

long-term rewards to reduce the energy cost during periods of high electricity prices, which are referred to as the peak price periods in this paper. According to the historical data, the periods 36 and 37 are the peak price periods, and the time of these periods corresponds to 18:00 to 19:00. Also, the temperature violation during the high/low periods, which are referred to as the peak temperature periods, are checked. The low temperature is further to the acceptable zone temperature than the high temperature. Therefore, periods 8, 9, and 10 are the peak temperature (lowest) periods, and the time corresponds to 4:00 to 5:30.

The results are shown in **Table 2**. The pre-trained policy is the DRL policy, which is only pre-trained in the knowledge-driven model. The final policy is the DRL policy, which is pre-trained in the knowledge-driven model and trained in the environment. The MPC method solves the one-period optimal result of **Eqs. 1–9** with M_k. Generally, the MPC can satisfy the temperature requirement at a relatively high cost. The temperature violation of MPC is almost zero. In comparison, the pre-trained policy saves about 1/3 energy cost on average but the temperature violation is unacceptable. The final policy of proposed method saves an average of 26.99% energy cost in all periods and 32.17% energy cost in peak price periods comparing to one-period MPC. The temperature violation of the final policy is very small too. The temperature randomness in the environment causes a large temperature violation of the pre-trained policy. Compare to the pre-trained policy, the final policy reduces the influence of temperature randomness at an average of 61.71% in all periods, and 71.81% in peak temperature periods. The detailed profile of the testing results is shown in **Figure 4**. The figure shows the average electricity price, the average energy cost, the average ambient temperature, and average temperature violation for 10 trails.

Learning Efficiency and Learning Costs

Comparison of Deep Deterministic Policy Gradient and HMB-Deep Deterministic Policy Gradient

In the first test, we compare the policy of DRL with the one-period MPC policy to show the improvement of DRL methods in the HVAC control problem. In this test, the improvement of learning efficiency and learning cost is tested compared to the DDPG algorithm. Since the learning efficiency and learning cost of the pre-training process is not as important as the online

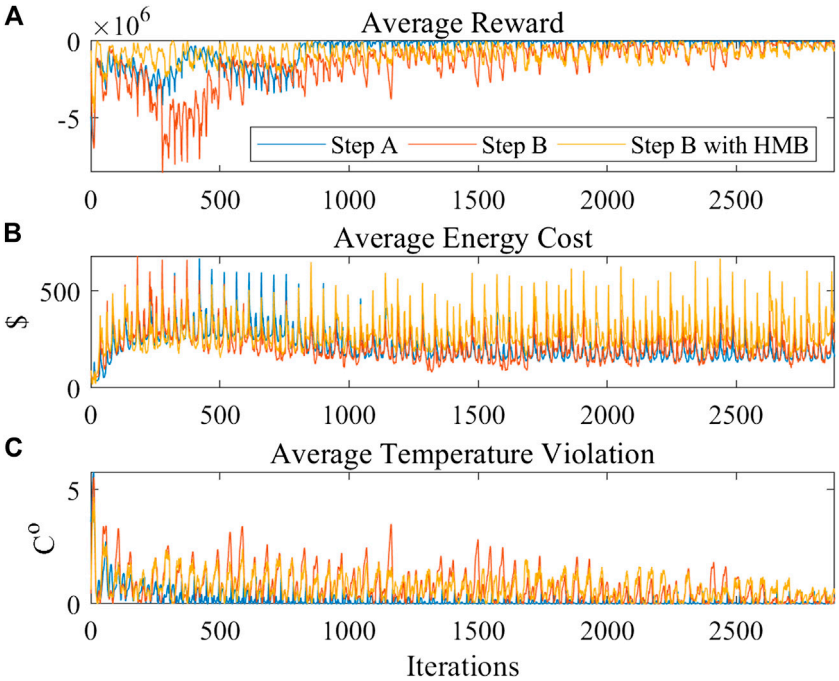

FIGURE 5 | The comparison of DDPG and HMB-DDPG **(A)** average reward **(B)** average energy cost **(C)** average temperature violation.

learning process, only the process of online learning is compared. The other settings are the same as the previous test.

The average reward, average energy cost, and average temperature violation profile results are shown in **Figure 5**. Step A represents the pre-training process using DDPG in the knowledge-driven model. Step B represents the training process using DDPG in the environment. Step B with HMB represents the training process with HMB-DDPG in the environment. Both DDPG and HMB-DDPG use the same pre-trained policy from step A for further training. The HMB-DDPG shows less learning cost and higher learning efficiency compared to DDPG. The low learning rewards of DDPG do not occur in the HMB-DDPG. The next two tests show the specific improvements in the protection mechanism and adjusting reward methods.

Comparison of Deep Deterministic Policy Gradient and Deep Deterministic Policy Gradient With Protection Mechanism

The target of the protection mechanism is to avoid the high learning cost during the learning process. To verify the performance of the protection mechanism, the DDPG with the protection mechanism is compared with DDPG during the online learning process in the environment. The threshold of the protection mechanism is set as the temperature violation of RL action in the knowledge model exceeds 7°, and the executed action is the average of the MPC action and the RL action.

The average reward, the average energy cost, the average temperature violation, and the average percentage of use protection during the training process in 10 trails are shown in **Figure 6**. Similarly to the former test, step A represents the pre-training process using DDPG in the knowledge-driven model. Step B represents the training process using DDPG in the environment. Step B with PM represents the training process using DDPG with the protection mechanism method in the environment. Both step B and step B with PM use the same pre-trained policy from step A for further training. In the upper picture of **Figure 6**, when the agent detects high learning cost periods, the protection mechanism has a higher activation probability in the corresponding periods in the second picture. Since the executed actions change from the unacceptable RL actions to the new actions which combined with the MPC actions, the energy costs and temperature violations are reduced.

Comparison of Deep Deterministic Policy Gradient and Deep Deterministic Policy Gradient With Adjusting Reward

The target of the reward function is to minimize the energy cost within a satisfactory temperature range. For the continuous action space problem, the first challenge is to find the available action range so the agent can exploit the range. The common way is to add the action violation to the reward function as a penalty. Directly ignore the violation of action will cause the policy hard to converge to the feasible range. However, when the parameter value of the violation item is large, exploration at the boundary of the available range is hard because the policy is influenced by the boundary of action.

The adjusting reward method is tested using the DDPG algorithm that activating the penalty item during the pre-training process and removing the penalty item during the

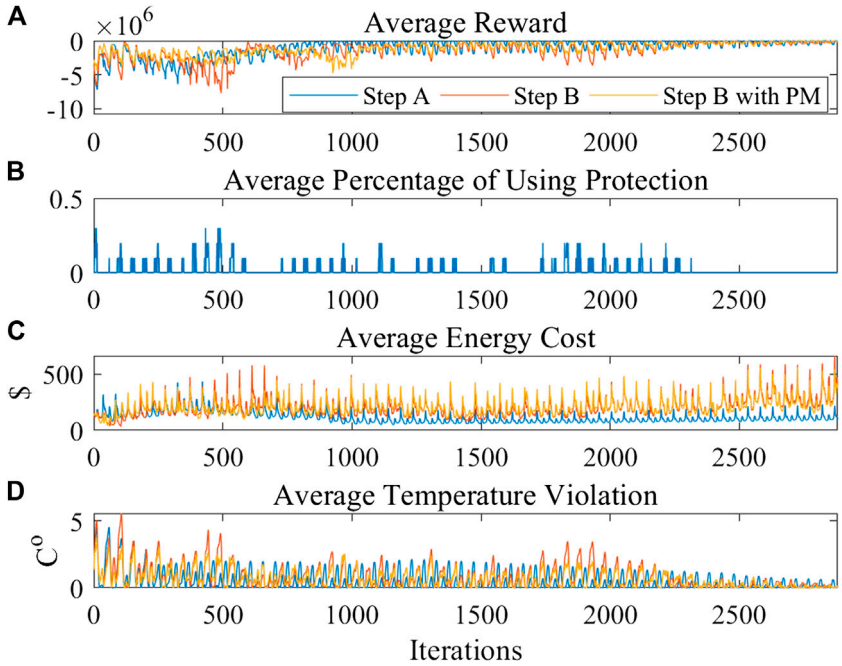

FIGURE 6 | The comparison of DDPG and DDPG with protection mechanism **(A)** average reward **(B)** average percentage of using protection **(C)** average energy cost **(D)** average temperature violation.

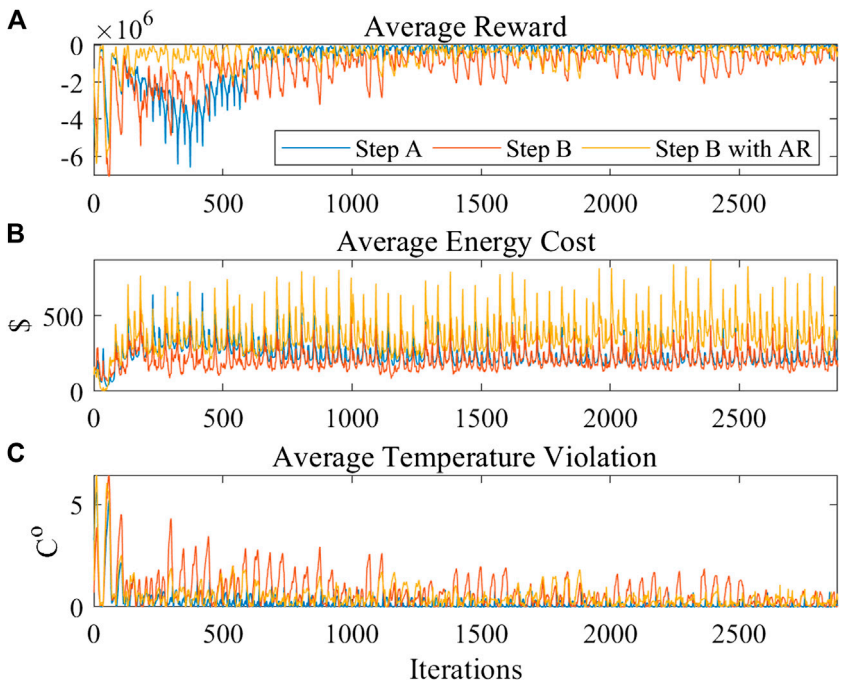

FIGURE 7 | The comparison of DDPG and DDPG with adjusting reward **(A)** average reward **(B)** average energy cost **(C)** average temperature violation.

online learning process. The average reward, average energy cost, and average temperature violation profile results are shown in **Figure 7**. Step A represents the pre-training process using DDPG in the knowledge-driven model. Step B represents the training process using DDPG in the environment. Step B with AR represents the training process using DDPG with adjusting the

reward method in the environment. Both step B and step B with AR use the same pre-trained policy from step A for further training. Comparing to the training process of DDPG without AR (step B), the DDPG with AR (Step B with AR) converge faster.

CONCLUSION

Energy consumption caused by the HVAC systems accounts for a large proportion of the entire building. Reducing the energy costs while maintaining temperature satisfaction is the main target of the HVAC system, but the performance is limited by the dynamic environment and system modeling accuracy. DRL is a model-free method that interacts with the environment to learn the optimal policy. Learning efficiency and learning cost are the main obstacles to the implementation of the DRL method. Therefore, we proposed a new hybrid model-based RL framework for the HVAC control problem. The model-based RL framework can learn the policy efficiently and the knowledge-driven model can provide additional information for the agent to avoid low reward actions. The simulation results show that the final policy of the proposed method saves an average of 26.99% energy cost in all periods and 32.17% energy cost in peak price periods comparing to one-period MPC. The hybrid-model-based method reduces the online learning cost by using the knowledge-driven model.

Although the simulation results show the reliability of the method, it still needs to set many hyperparameters such as neural network structure and learning rate to obtain a good online learning process. In real cases, repeating the learning process will also cause the learning cost. Further works will focus on how to automatically adjust these hyperparameters during the pre-training process to the reduce online learning cost. Also, the dynamic human activity need to be considered in the framework which can further reduce the energy consumption. By appending the information about whether humans are acting in the area or not to the state, the factor can be integrated into the proposed method. The data-driven model like RNN will be studied to model the dynamic human activities in the future.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

HZ developed the idea and wrote the initial version of the paper. JZ provided the idea and performed the full edit of the paper. TS provided the theory of thermal and weather models. ZP implemented the thermal model. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of the Shenzhen Municipal Science and Technology Innovation Committee (ZDSYS20170725140921348, JCYJ20160510153103492).

REFERENCES

Afram, A., and Janabi-Sharifi, F. (2014). Theory and applications of hvac control systems–a review of model predictive control (mpc). *Build. Environ.* 72, 343–355. doi:10.1016/j.buildenv.2013.11.016

Afram, A., Janabi-Sharifi, F., Fung, A. S., and Raahemifar, K. (2017). Artificial neural network (ann) based model predictive control (mpc) and optimization of hvac systems: a state of the art review and case study of a residential hvac system. *Energy Build.* 141, 96–113. doi:10.1016/j.enbuild.2017.02.012

Amasyali, K., and El-Gohary, N. M. (2018). A review of data-driven building energy consumption prediction studies. *Renew. Sustain. Energy Rev.* 81, 1192–1205. doi:10.1016/j.rser.2017.04.095

Barrett, E., and Linder, S. (2015). "Autonomous hvac control, a reinforcement learning approach," in Joint european conference on machine learning and knowledge discovery in databases. Dublin, Ireland, September 10-14. *Lecture notes in computer science*. Editors A. Bifet, M. May, B. Zadrozny, R. Gavalda, D. Pedreschi, F. Bonchi, et al. (Cham: Springer International Publishing), Vol. 9286, 3–19.

Belic, F., Hocenski, Z., and Sliskovic, D. (2015). HVAC control methods - a reviewInternational conference on system theory, control and computing, ICSTCC 2015 - joint conference SINTES 19, SACCS 15. *SIMSIS* 19, 679–686. doi:10.1109/ICSTCC.2015.7321372

Chen, Y., Norford, L. K., Samuelson, H. W., and Ali, M. (2018). Optimal control of hvac and window systems for natural ventilation through reinforcement learning. *Energy Build.* 169, 195–205. doi:10.1016/j.enbuild.2018.03.051

Chua, K., Calandra, R., McAllister, R., and Levine, S. (2018). "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," in Advances in neural information processing systems 2018-Decem (NeurIPS), Montréal, Canada, December 2-8 , 4754–4765.

Deisenroth, M. P., and Rasmussen, C. E. (2011). PILCO: a model-based and data-efficient approach to policy search. Icml, June 28th to July 2nd. Available at: http://eprints.pascal-network.org/archive/00008310/, 465–472.

Ding, X., Du, W., and Cerpa, A. (2019). "OCTOPUS: deep reinforcement learning for holistic smart building control," in BuildSys 2019–proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation, New York, USA, November 13-14, 326–335.

Fazenda, P., Veeramachaneni, K., Lima, P., and May O'Reilly, U. (2014). Using reinforcement learning to optimize occupant comfort and energy usage in hvac systems. *J. Ambient Intell. Smart Environ.* 6 (6), 675–690. doi:10.3233/AIS-140288

Gao, G., Li, J., and Wen, Y. (2019). Energy-efficient thermal comfort control in smart buildings via deep reinforcement learning. ArXiv. 10.1109/jiot.2020.2992117

Gomez-Romero, J., Fernandez-Basso, C. J., Victoria Cambronero, M., Molina-Solana, M., Campana, J. R., Ruiz, M. D., et al. (2019). A probabilistic algorithm for predictive control with full-complexity models in non-residential buildings. *IEEE Access*. 7, 38748–38765. doi:10.1109/ACCESS.2019.2906311

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature* 518 (7540), 529–533. doi:10.1038/nature14236

Paone, A., and Bacher, J. P. (2018). The impact of building occupant behavior on energy efficiency and methods to influence it: a review of the state of the art. *Energies* 11 (4). 953. doi:10.3390/en11040953

Sutton, R. S. (1990). "Integrated architectures for learning, planning, and reacting based on approximating dynamic programming," in Machine learning proceedings, Austin, Texas, June 21–23, Vol. 02254, 216–224.

Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning : an introduction*. Cambridge, MassachusettsMassachusetts London, England: The MIT Press Cambridge.

Vázquez-Canteli, J., Kämpf, J., and Nagy, Z. (2017). Balancing comfort and energy consumption of a heat pump using batch reinforcement learning with fitted q-iteration. *Energy Procedia* 122, 415–420. doi:10.1016/j.egypro.2017.07.429

Wei, T., Wang, Y., and Qi, Z. (2017). "Deep reinforcement learning for building hvac control," in Proceedings–design automation conference part, Austin, TX, USA, June 2017, 12828.

Xie, D., Liang, Y., Jiang, T., and Zou, Y. (2018). Distributed energy optimization for hvac systems in university campus buildings. *IEEE Access* 6, 59141–59151. doi:10.1109/ACCESS.2018.2872589

Yan, Z., and Xu, Y. (2019). Data-driven load frequency control for stochastic power systems: a deep reinforcement learning method with continuous action search. *IEEE Trans. Power Syst.* 34 (2), 1653–1656. doi:10.1109/TPWRS.2018.2881359

Yan, Z., and Xu, Y. (2020). Real-time optimal power flow: a lagrangian based deep reinforcement learning approach. *IEEE Trans. Power Syst.* 35 (4), 3270–3273. doi:10.1109/TPWRS.2020.2987292

Ye, Y., Qiu, D., Jing, L., and Strbac, G. (2019). Multi-period and multi-spatial equilibrium analysis in imperfect electricity markets: a novel multi-agent deep reinforcement learning approach. *IEEE Access* 7, 130515–130529. doi:10.1109/ACCESS.2019.2940005

Yu, L., Sun, Y., Xu, Z., Shen, C., Dong, Y., Jiang, T., et al. (2020a). Multi-agent deep reinforcement learning for hvac control in commercial buildings. *IEEE Trans. Smart Grid* 1. doi:10.1109/tsg.2020.3011739

Yu, L., Xie, W., Xie, D., Zou, Y., Zhang, D., Sun, Z., et al. (2020b). Deep reinforcement learning for smart home energy management. *IEEE Internet Things J.* 7 (4), 2751–2762. doi:10.1109/JIOT.2019.2957289

Zhang, D., Han, X., and Deng, C. (2018). Review on the research and practice of deep learning and reinforcement learning in smart grids. *CSEE J. Power Energy Syst.* 4 (3), 362–370. doi:10.17775/cseejpes.2018.00520

Zhang, Z., Chong, A., Pan, Y., Zhang, C., and Lam, K. P. (2019). Whole building energy model for hvac optimal control: a practical framework based on deep reinforcement learning. *Energy Build.* 199, 472–490. doi:10.1016/j.enbuild.2019.07.029

Zhao, H., Zhao, J., Qiu, J., Liang, G., and Zhao Dong, Y. (2020). Cooperative wind farm control with deep reinforcement learning and knowledge assisted learning. *IEEE Trans. Ind. Inf.* 16 (11), 6912–6921. doi:10.1109/tii.2020.2974037

Zhao, J., Zhu, N., and Wu, Y. (2009). The analysis of energy consumption of a commercial building in tianjin, china. *Energy Pol.* 37 (6), 2092–2097. doi:10.1016/j.enpol.2008.11.043

Zheng, G., Zhang, F., Zheng, Z., Xiang, Y., Yuan, N. J., Xie, X., et al. (2018). "DRN: a deep reinforcement learning framework for news recommendation," in The web conference 2018–proceedings of the world wide web conference, WWW 2018, Lyon, France, April 23-27, Vol. 2, 167–176.

Zou, Z., Yu, X., and Ergan, S. (2020). Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network. *Build. Environ.* 168, 106535. doi:10.1016/j.buildenv.2019.106535

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.