IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 12, 2021, accepted September 17, 2021, date of publication September 20, 2021,
date of current version September 29, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3114161

Research and Application of Predictive Control Method Based on Deep Reinforcement Learning for HVAC Systems

CHENHUI FU AND YUNHUA ZHANG
Computer Technology Development Center, Zhejiang Sci-Tech University, Hangzhou 100191, China
Corresponding author: Chenhui Fu (201920603006@zstu.edu.cn)

ABSTRACT Energy efficiency and consumption control remain a significant topic in the area of Heating, Ventilation, and Air Conditioning (HVAC) systems. Deep reinforcement learning (DRL) is an emerging technique to optimize energy consumption. Its advantage lies in the ability to tackle the time-series nature of energy data and complexity brought by environmental factors. However, most DRL algorithms have not considered both time-of-use electricity pricing and thermal comfort. This paper proposed a hybrid approach based on twin delayed deep deterministic policy gradient algorithm and model predictive control (TD3-MPC) for HVAC systems, to mitigate function approximation errors and save cost by pre-adjusting building temperatures at off-peak times. This proposed method is compared with deep deterministic policy gradient (DDPG) algorithm under simulations of five building zones. Experiment results demonstrate that TD3-MPC outperforms DDPG algorithm and potentially saves 16% of total energy consumption cost, with better stability and robustness.

INDEX TERMS Deep reinforcement learning, energy consumption efficiency, HVAC, MPC, TD3.

I. INTRODUCTION

Global climate change is a concerning issue and people are actively exploring opportunities to reduce carbon emission and mitigate energy consumption. In China, for example, building energy consumption accounts for 21.7% of the total national energy consumption [1]. The development of low-energy buildings is one opportunity with great potential that draws a lot of attention. The convertibility of various energy sources in buildings as well as buildings' energy storage capacity make it an ideal choice for energy optimization [2]. HVAC systems are key contributors to the energy consumption within buildings. Thus, it is crucial to manage such systems with effective maintenance and operation strategies.

Researchers around the world have proposed various methods and architectures to help reduce the energy consumption of HVAC systems. Traditional feedback control such as on-off control or proportional-integral-derivative (PID) control cannot reflect external interference in time, which may result in suboptimal performance. MPC, an optimal control strategy for stochastic systems with random constraints, solves this problem by designing offline uncertainty distributions and has been proved successful in recent years [3]–[6]. For instance, Zeng and Barooah presented an autonomous adaptive MPC architecture for HVAC with periodical relearning building dynamics to maintain indoor temperatures while reducing energy usage [7]. Asvadi and Momenibuilt MPC models for each air handling unit in HVAC systems. With this approach, the mean of energy consumption reduction is about 36.94% [8]. Additionally, James et al introduced a two-layer method to decompose the economic MPC in large commercial HVAC systems hierarchically to reduce the operational costs of HVAC and improve energy efficiency [9]. The performance and reliability of above methods are highly dependent on the accuracy of models and robustness of online optimization, so mathematical tools are needed to effectively solve runtime control problems in practical applications [10]. Online optimizations would also add additional complexity to existing problems if the MPC model is nonlinear.

Recently, the rapidly evolving DRL technique starts to benefit many industries. It is often used to solve sequential decision-making and continuous control problems by

The associate editor coordinating the review of this manuscript and approving it for publication was Ahmed A. Zaki Diab.

controlling the interactions between individuals and environment. The current mainstream DRL algorithms include: Q-learning, deep Q-network and DDPG. The extensive applications of these DRL algorithms in various areas such as unmanned driving and future sales forecasting incentivize the exploration of utilizing DRL in energy consumption reduction. Compared with MPC, DRL avoids the establishment of complex models and allows the optimal controller to learn directly from real-time data. It could as a result obtain real-time dynamic systems with fewer environmental parameters and more effective controls. The core idea in utilizing DRL to predict and reduce building energy consumption is to regard HVAC systems as Markov decision processes, in which the reward function includes energy costs, impact from environmental temperatures and behavioral violations [11], [12]. For example, Jiang et al developed a deep Q-network with an action processor and reward shaping technique to overcome the issue of reward sparsity caused by the demand charge under time-varying electricity price profiles [13]. Likewise, Wei et al put forward a method based on deep Q-network to control the air conditioning systems by human expressions, while meeting the requirements of thermal comfort [14]. Du et al applied a data-driven DDPG method to minimize HVAC systems' energy consumption costs while maintaining comfort [15]. Although above methods avoided establishing complex models and took into consideration energy consumption and thermal comfort separately, the limitation was reflected in overestimation bias caused by noise when using function approximation errors to calculate the Q value in value function. The overestimation would eventually lead to sub-optimal policies.

In recent years, researchers have focused their studies on energy storage strategies, which if well applied, could have high potential in the reduction of energy consumption. In the area of energy storage, peak shaving has been proved to be able to reduce electricity cost by 10-30%, achieved through load shedding and energy storage [16], [17]. Peak shaving can reduce a consumer's power consumption quickly and in a short period of time so that it could avoid consumption spikes. Hong et al proposed a novel performance evaluation framework for deep peak shaving [18]. In this framework, operation data and reference status labels were fed into deep belief networks for dimension reduction and feature extraction in a semi-supervised way. Similarly, Mawson and Hughes utilized an analysis of thermal energy with machine learning adopted to predict spikes in energy consumption and in turn optimize HVAC systems [19]. These studies provided a new idea of combining DRL methods with peak shaving to further reduce building energy consumption.

Nevertheless, most of existing articles either provided no insight or showed limited focus on energy consumption, thermal comfort, and electricity prices at the same time. To address this problem and avoid overestimation, an energy consumption prediction and storage control algorithm based on TD3-MPC is proposed. The main contributions of this paper are as follows:

1) Established a thermal dynamics model to predict the future trend of HVAC systems by MPC algorithm, considering the influence of outdoor environmental factors in reality.

2) Proposed a control framework of HVAC systems based on TD3-MPC along with energy storage strategy, applicable in continuous constraint action space.

3) Compared the proposed approach with other DRL methods, given the same outdoor environment. The experiment result shows that the proposed algorithm can significantly reduce energy consumption while meeting the requirements of room temperature.

II. METHODOLOGY

A. MPC

MPC is a multivariable control strategy that can be applied to both linear and nonlinear systems with the aim to minimize the cost function [20]. It involves a dynamic model of the process loop, historical values of the control input variable, and an optimization equation in prediction horizon. The optimal control can be obtained by the above three elements. Since the MPC policy can also produce next-state predictions, the prediction result from MPC can be used as the initialization input for DRL algorithms. The control policy can be defined as a function of network parameter θ, as shown in Equation 1:

$$L(\theta) = \sum_t \lambda \left\| x_{t+1} - x'_{t+1} \right\|_2^2 + \left\| u_t - u_{t+1} \right\|_2^2 \qquad (1)$$

where x_{t+1} and x'_{t+1} are the actual next state and the next state predicted by the agent, u_t and u_{t+1} are the actions taken in the current step and the next step. The hyperparameter λ balances the relative importance of actions and next-state predictions.

After obtaining a policy, actions taken by the agent are executed under the policy. Above steps will be repeated continuously until an optimal control strategy is obtained.

B. DRL METHODS

1) DDPG

Policy gradient is an algorithm used to obtain the optimal policy which maximizes the reward expectation in continuous action space [21]. It is iteratively calculated to find an optimal Q-value. If a policy is determined, which means it can take only one action in state space, the expected Q-value is then only related to the environment and policy [22]. Gradient update can then be regarded as policy update to the Q-value gradient, described as:

$$\theta^{k+1} = \theta^k + \alpha G_t \nabla_\theta \ln \pi \left(a_t \mid s_t; \theta^k \right) \qquad (2)$$

where G_t refers to the reward by each time step; $\pi \left(a_t \mid s_t; \theta^k \right)$ represents the policy as a function of action a, state s and parameter θ^k; ∇_θ is the score function of θ which makes the gradient of log-likelihood function equal to 0; and α is a hyperparameter to balance the impact from previous steps on the current step.

DDPG establishes a Q-function and a policy-function incorporated with experience replay dual-network structure to enable a neural network in the probability algorithm to

learn with maximum efficiency [23], [24]. Convergence difficulty caused by actor-critic algorithm can be handled by random sampling from previous state transfer experience for training.

In DDPG, actor and critic networks are split into two parts, current network and target network, respectively for action selection and evaluation [25]. Actor current network is used to update network parameter θ_1, select the optimal action A according to the current state S, and interact with the environment to transfer to the next state S' and reward R. Actor target network selects an optimal next action A' according to the next state S' sampled in the experience replay buffer and periodically updates network parameter θ_1 in the current network. On the other hand, critic current network is responsible for updating network parameter θ_2 and calculating the current state-action value $Q(S, A, \theta_2)$. Critic target network then computes the target Q-value $Q'\left(S', A', \theta_2'\right)$, where network parameter θ_2' is copied from θ_2 periodically.

2) TD3

The way on which DDPG updates introduces function approximation errors to the max operator. Function approximation errors are known to lead to overestimated value estimates and suboptimal policies [26], [27]. To deal with these shortcomings of DDPG, TD3 uses delayed policy updates to reduce errors caused by each update step to further improve the performance. It also proposes to utilize two critic networks to calculate the Q-target value and Q-value to accelerate convergence, and takes the smaller one as target to update both critic networks at the same time in order to reduce overestimation bias. Additionally, parameters in actor networks are regularized to reduce noise and estimated deviations [28]. As a result, TD3 can effectively avoid the problem of overestimation.

C. ENERGY STORAGE STRATEGY

Typical energy storage technique of HVAC systems includes ice storage and chilled water storage systems, which are characterized by large energy storage capacity with high initial investments [29]. Based on the building energy storage and virtual energy storage technology initiated in advance, building envelopes are used to store a certain amount of cold energy when electricity price reaches its peak, thus achieving the purpose of energy conservation and emission reduction [30]. It is a passive coupling process with indoor temperature setting and dynamically changing load. A building's virtual storage/discharge energy can be expressed as:

$$Q_{VES}(\Delta T_{in}, \Delta T_w) = \rho_{air} \cdot c_{air} \cdot V_{zone} \cdot \frac{\Delta T_{in}}{d_t} + \rho_w \cdot c_w \cdot V_w \cdot \frac{\Delta T_w}{d_t} \quad (3)$$

where ρ_{air} is air density, c_{air} is air specific heat ratio, V_{zone} is indoor air capacity, ΔT_{in} is indoor temperature variation value, ρ_w is building envelope density, c_w is building envelope heat ratio, V_w is volume of building envelope, ΔT_w is temperature variation of building envelope.

The advance startup time of HVAC systems is the key to building energy storage [30]. If started too early, there would be no insulation between the building and outdoor environment, and the energy consumption would be large. Alternatively, if started too late, it cannot meet the goal of energy storage. Back propagation method is used to predict the advance startup time of HVAC systems according to the outdoor hourly temperature. The learning process of this approach is divided into forward propagation and back propagation. The cooling load of a day is predicted by the hourly outdoor temperature to obtain the relationship between the cooling load and start time of HVAC, in order to determine the advance startup time. In the simulation, the advanced startup time is set at 5 am. That's because time-of-use electricity price is lower before 6 am. Electricity prices for energy cost calculations are shown in Fig.1, based on local time-of-use electricity prices.

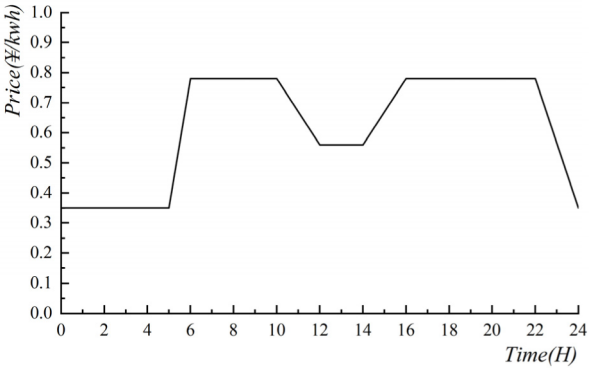


FIGURE 1. Time-of-use electricity prices.

III. FRAMEWORK OF ALGORITHM

This section introduces the whole framework and process of TD3-MPC algorithm, and lays out the definitions of state space, action space and value function in the algorithm. There are three parts within the algorithm framework: energy consumption prediction, action selection, and update. At last, this TD3-MPC algorithm is further optimized to incorporate energy storage strategy.

A. DEFINITIONS IN HVAC

Building HVAC systems' operation is based on current temperatures and outdoor environmental disturbances to ensure each region reaches desired temperatures with minimized energy costs while achieving temperature comfort [31]. The regional temperature of the next time step is only determined by the current system state, environmental interference, and the air conditioning input of HVAC systems, independent of previous state of the building. A building with five temperature zones equipped with a HVAC system to provide constant temperature with air flow is simulated. Five zones are created in order to verify the robustness of the proposed algorithm.

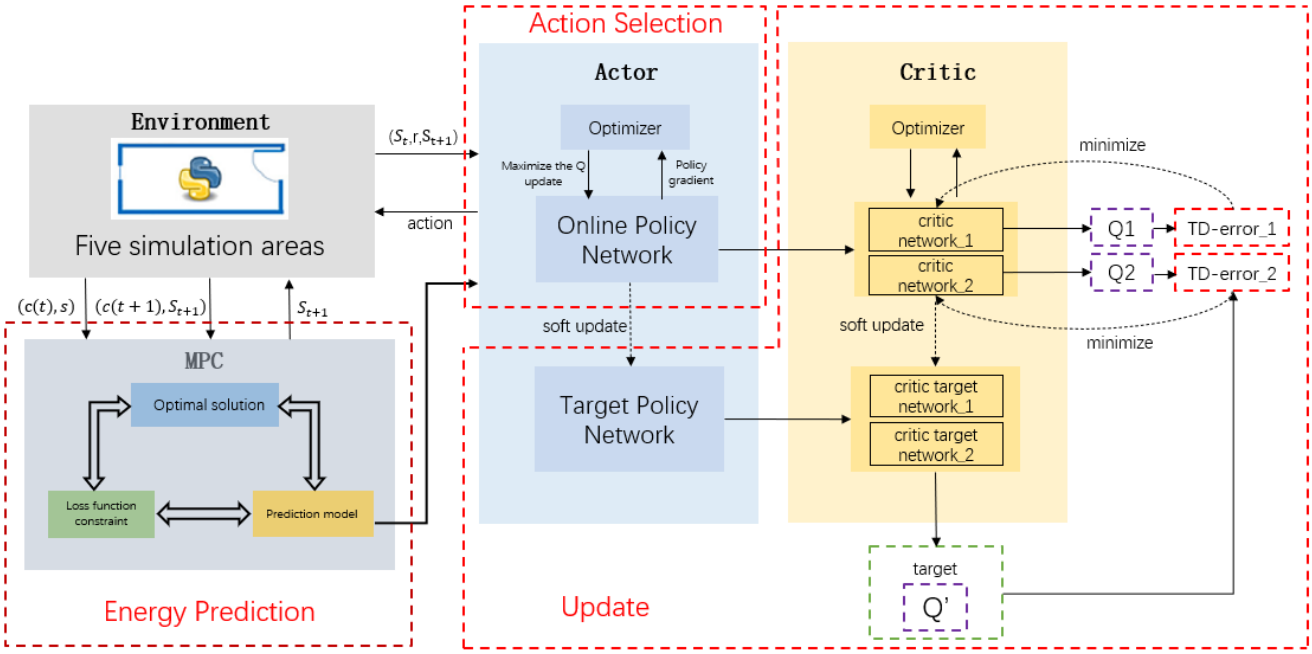


FIGURE 2. Framework of TD3-MPC algorithm.

The simulated environment is then used to train a DRL agent to evaluate performances in energy consumption and thermal comfort. Finally, time-of-use electricity price is used to calculate the energy cost of HVAC.

1) STATE SPACE

States are a DRL agent's observation of each control step, representing valuable information that the agent can obtain before making decisions to help evaluate situations. In order to equip the agent with a better understanding of the changing environment, states are divided into current environment state superposed upon history and predictive state of environment. The state value can be expressed as $S = \{s_t, s_{t-1}, \ldots, s_{t-n}\}$.where t is measured time step, n represents number of historical control time steps. Each item s consists of the following factors: regional air temperature T, outdoor air temperature T_{out}, outdoor air flow rate V, desired temperature T_{set}, time-of-use electricity price p_t, electrical power load l_0, and regional cooling load l_t. The total energy consumption of HVAC can be calculated as follows, where η is a constant coefficient:

$$W_t = \eta l_t + l_0 \tag{4}$$

The regional cooling load l_t is directly controlled by the agent. Electrical power load l_0 is a fixed constant, which is not affected by other factors in the region and is the basic operational cost of building HVAC systems. The total energy consumption is as follows, where δ is a constant coefficient, and p_t refers to the time-of-use electricity price:

$$E = \delta p_t W_t \tag{5}$$

2) ACTION SPACE

Actions are what agents take to control the environment according to a specified strategy. In building HVAC systems, actions are the temperature settings of regions which agents adjust. The adjustments are based on observed external feedback conditions, and the values are discrete, such as: $A = \{20, 21, \ldots, 32\}$. The temperatures of a controller range from 20 to 32 degrees Centigrade.

3) REWARD FUNCTION

The formulation of reward function in DRL algorithm determines the objective of control optimization [32]. The goal of the study is to minimize energy consumption while satisfying thermal comfort requirements. Based on the total energy consumption cost (captured by the output of the simulation model) in Section 2 Part C, timely reward can be obtained as follows:

$$r = -\beta E - (1-\beta)(T_{diff})^2 \tag{6}$$

where T_{diff} is the difference between indoor and outdoor temperatures, and β is a constant weight introduced to avoid biased results. In this simulation, β is 0.05. Within the above equation, negative incentives are used to get a maximum cumulative reward. During the operation of HVAC systems, the cumulative reward can be defined as $R = \sum_{i=1}^{n} \gamma^{i-1} r_{t+i}$, where γ is attenuation factor.

B. ALGORITHM FRAMEWORK

The DRL control framework of HVAC systems is shown in Fig.2, including three steps: energy consumption prediction, action selection, and update. MPC is used to predict

TABLE 1. Pseudocode of proposed TD3-MPC algorithm.

PSEUDOCODE OF TD3-MPC ALGORITHM

/* Step 1: Model predictive control */
Initialize control policy $\boldsymbol{\theta} \leftarrow \mathbf{0}$, $\boldsymbol{D} = \emptyset$
for $\boldsymbol{t}$ = 1: T **do**
  Sample HVAC environment state $\boldsymbol{s}$
  Obtain optimal action sequence $\boldsymbol{A_t^H}$ with $\boldsymbol{H}$ steps prediction based on prediction model
  Execute first action $\boldsymbol{a_t}$ from selected sequence $\boldsymbol{A_t^H}$
  Store $(\boldsymbol{s_t}, \boldsymbol{a_t}, \boldsymbol{r_t}, \boldsymbol{s_{t+1}})$ to $\boldsymbol{D}$, $\boldsymbol{a_{t+1}} = \boldsymbol{\pi_\theta(a_t)}$
end for
Train $\boldsymbol{\theta}$ on optimal weights using $\boldsymbol{D}$
/*Step 2: Energy consumption predict by TD3*/
Input the deterministic policy $\boldsymbol{\pi_\theta}$
Initialize actor network and critic network with weights $\boldsymbol{\theta_1}, \boldsymbol{\theta_2}$
Initialize target network $\boldsymbol{\theta_1}' \leftarrow \boldsymbol{\theta_1}$, $\boldsymbol{\theta_2}' \leftarrow \boldsymbol{\theta_2}$ and replay buffer B
for episode $\boldsymbol{t}$ = 1: T **do**
  Initialize a random noise $\boldsymbol{\epsilon \sim clip(N(0, \sigma), -c, c)}$ for action exploration and observe reward $\boldsymbol{r}$ and new state $\boldsymbol{s'}$
  Store transition tuple $(\boldsymbol{s, a, r, s'})$ in B
  Sample minibatch of transitions $(\boldsymbol{s, a, r, s'})$ from B
  $\boldsymbol{a' \leftarrow \pi_\theta(a) + \epsilon}$
  $\boldsymbol{y \leftarrow r + \gamma(1-d) \min_{i=1,2} Q_{i,arg}(s', a'(s'))}$
  Update critic by minimizing loss $\boldsymbol{L(\theta_{i=1,2})}$ in Eq. 8
  If $\boldsymbol{t}$ mod d **then**
    Update actor policy using deterministic policy gradient $\boldsymbol{\theta^{k+1}}$ in Eq. 2
    Update target actor network: $\boldsymbol{\theta_1}' \leftarrow \boldsymbol{\tau\theta_1 + (1-\tau)\theta_1}'$
    Update target critic network: $\boldsymbol{\theta_2}' \leftarrow \boldsymbol{\tau\theta_2 + (1-\tau)\theta_2}'$
  end if
end for

the energy consumption of building HVAC systems. After the phase of constructing the prediction model, the goal of reinforcement learning is to update control policy by getting the optimal action. The balance and utilization of the relationship between actions is related to the formulation of optimal strategies and learning speed [33].

Table 1 presents the structure of TD3-MPC algorithm. The interactions between the agent and environment begin with the initialization of control policy θ and empty replay buffer D. It takes current environmental state s and outdoor environmental temperature as inputs. Then the algorithm calculates an optimal action sequence A_t^H with H steps prediction. The agent executes first action a_t from A_t^H and stores to replay buffer, and later uses the replay buffer D to optimize parameter θ. After getting the optimized parameters for control policy, the algorithm reaches the action selection stage, which determines the accuracy of the prediction. TD3 takes the current environmental state as input to initialize actor network and critic network with weights θ_1 and θ_2, and target network with weights θ_1', θ_2'. To avoid overfitting, a truncated normally distributed noise $\in \sim clip\left(N\left(0, \tilde{\sigma}\right), -c, c\right), c > 0$ is added to each action.

Since the actor and critic networks have randomly initialized parameters, the output Q-values from the networks must be different, respectively denoted as Q1 and Q2. TD3-MPC takes the clipped minimum as Q-target to offset the overestimation of Q-value and substitutes it into the Bellman equation to calculate temporal difference (TD) error and the expectation loss function, where γ and d are setting parameters and r is the reward by taking specific action:

$$y\left(r, s', d\right) = r + \gamma\left(1 - d\right) \min_{i=1,2} Q_{\theta i}\left(s', a'\left(s'\right)\right) \tag{7}$$

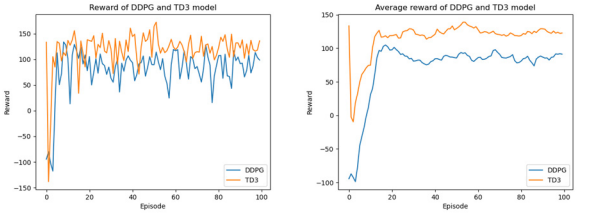
$$L\left(\theta_i\right)_{i=1,2} = E_{(s,a,r,s',d)}\left[\left(Q_{\theta_i}\left(s, a\right) - y\left(r, s', d\right)\right)^2\right] \tag{8}$$

Although this update rule for Q-value may induce underestimation bias over standard Q-learning method, the underestimated actions will not be explicitly propagated through the policy update. In the end, θ_1' and θ_2' are respectively used to update target actor and critic networks. The whole modeling process takes into account influences of outdoor temperatures and electricity prices on energy consumption cost, which is considered as an improvement over traditional algorithms.

IV. EXPERIMENTS

A. SIMULATION

In order to verify the effectiveness of TD3-MPC algorithm, a building with five regions was simulated and set with the same expected indoor temperature in these regions (22°C). These regions also had the same outdoor temperature derived from the actual local temperature. Controllable environment was rapidly modeled based on MATLAB. MATLAB was chosen as the modeling tool because it had been widely used in many experiments and achieved effective results [34]. It enabled the ability to present the complex building HVAC systems in an abstract code structure, allowing agents to control the building regions directly. Table 2 describes the indoor conditions set in 5 scenes. The only difference between these scenes was the indoor air capacity, because it was one important indicator of the indoor thermal environment [35]. All environment used 24 hours as a time step to simulate.



(a) Reward of DDPG and TD3 Models (b) Average Reward of DDPG and TD3 Models

FIGURE 3. Reward and average reward of DDPG and TD3 models.

In the first step of the simulation, TD3-MPC algorithm modeled the energy consumption prediction problem as a MPC prediction model. After a series of processing, the final energy consumption prediction vector and current environmental state vector were input into the DRL model. Secondly, the algorithm assumed that only the regional temperature was observable and controlled other irrelevant factors such as wind velocity and humidity. In the process of offline pretraining, the imitation loss was minimized to directly evaluate the performance by allowing the agent to control the environment. After 100 episodes of training, the reward and average reward of DDPG and TD3 model were shown in Fig.3.

TABLE 2. Comparison of indoor conditions in five scenes.

#	Hyperparameters	
	Air Capacity(m^3)	ρ_{air}, c_{air},V_w in Eq.(3)
Scene 1	118	1.3 kg/m^3,1.006 kJ/(kg • K),20m^3
Scene 2	80	1.3 kg/m^3,1.006 kJ/(kg • K),20m^3
Scene 3	192	1.3 kg/m^3,1.006 kJ/(kg • K),20m^3
Scene 4	140	1.3 kg/m^3,1.006 kJ/(kg • K),20m^3
Scene 5	132	1.3 kg/m^3,1.006 kJ/(kg • K),20m^3

B. EXPERIMENT RESULTS

At the beginning of training, reward of DDPG is very small because it frequently violates the temperature requirement. In other words, it cannot reach the expected value which results in great punishment [32]. In comparison, TD3 has higher reward at the beginning but the policy network is affected by valuation network as the training process goes on, resulting in a significant decline in the reward shortly after episode 0 and then volatility along the curve. This is because the strategy function is not fixed during training and it can be learned together with the valuation function. With the deepening of the learning process, these two algorithms show almost the same convergence, and the reward value gradually increases. This illustrates the agent can organically adjust policies according to changes in environmental conditions. Finally, expectations gradually converge to a fixed constant, which indicates that the agent can make the right decision in all states to adjust the temperature in order to minimize the energy consumption cost while satisfying thermal comfort requirements. Due to changes in outdoor temperatures and time-of-use electricity prices, the reward curve fluctuates slightly. Overall, TD3-MPC model has better performance than DDPG, which is reflected in the higher reward of TD3-MPC.

The trained models are then put into set environment for evaluation. DRL controller receives the state parameters transmitted from MATLAB by socket connection. The accuracies of DDPG and TD3-MPC models in energy consumption control strategy are measured through a comparative experiment. Fig.4 shows the prediction results randomly sampled on a day in five regions. Three lines represent predicted energy consumption results of DDPG, TD3-MPC, and TD3-MPC with energy storage strategy respectively. Three columns respectively represent the root mean square error (RMSE) of DDPG, TD3-MPC, and TD3-MPC with energy storage strategy at indoor temperature. This performance metric reflects the difference between indoor and set temperatures. The greater the RMSE value is, the greater error between indoor and set temperatures.

The results suggest that there is certain similarity between three models in various aspects. Curves of total energy consumption cost rise over time during 12 to 18 pm and peak at 17 to19 pm. This is due to the afterheat of high temperatures in the afternoon and the rise of electricity price in the evening. When the outdoor temperature is the highest in a day, the external interference is the strongest as well.

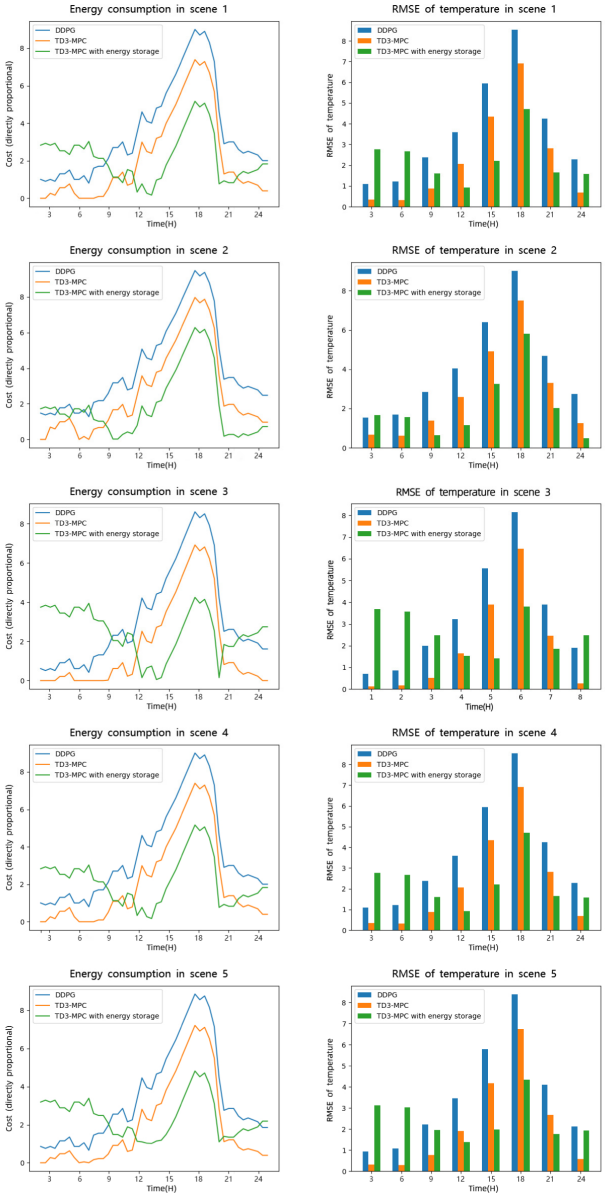


FIGURE 4. Experimental results.

After 19 pm, the fall of outdoor temperatures also directly leads to lower energy costs. Since the proposed TD3-MPC algorithm incorporates energy storage strategy, it is based on time pre-start after 22 pm when the electricity prices get down to avoid electricity peak. This also led to an increase in energy consumption from 22 pm to 9 am in the following morning. However, the cold air stored in the previous period can be effectively saved, to achieve the purpose of reducing energy consumption. The total energy consumption is still a lot lower than other models. Although electricity prices are at peak during 6 to 10 am, in order to store a certain amount of cold air before the high temperatures in the afternoon, some energy consumption during this period is needed to satisfy the thermal comfort requirements. In general, TD3-MPC algorithm can significantly reduce the energy. Compared with

DDPG, TD3- MPC is able to reduce energy consumption by about 16% at the peak of energy consumption. Moreover, the distribution diagrams of RMSE suggest that TD3-MPC algorithm is able to further reduce RMSE by 0.4 compared with the other two methods. This additional reduction in RMSE implies that this algorithm can maintain a lower temperature violation rate with better robustness.

V. CONCLUSION

This paper proposes an energy consumption prediction control algorithm combined with relevant strategies of energy storage for HVAC systems. The algorithm is named as TD3-MPC. This approach greatly reduces the uncertainty brought by the outdoor environment and it is suitable under different indoor air capacity settings. The experiment results demonstrate that

TD3-MPC algorithm has the following advantages:

1) TD3-MPC algorithm can effectively improve the accuracy of energy consumption prediction and control of building HVAC systems. It has better performance than the traditional control strategy DDPG.
2) It is applicable to different scenarios. Five scenes are enumerated in the experiment and TD3-MPC algorithm is the most efficient method in all five scenarios. This illustrates the robustness of the algorithm.
3) TD3-MPC is able to store energy in building area at off-peak periods to further reduce the energy consumption cost.

Compared with DDPG, TD3-MPC is more suitable to solve the continuous state space problem of action control, which minimizes the influence of overestimated Q-value on the control process. Results show that TD3-MPC reduces energy consumption cost by 16% and thermal comfort RMSE by 0.4 respectively.

REFERENCES

[1] Building Energy Efficiency Research Center of Tsinghua University, "China building energy research report 2020," *Building Energy Efficiency*, vol. 49, no. 2, pp. 1-6, 2021.
[2] Z. Cai, C. Lai, S. Chen, and Y. Xu, "A virtualized network experimental design scheme for energy-saving storage," *Comput. Times*, vol. 7, no. 4, pp. 11–14, 2019.
[3] Y. Long and L. Xie, "Iterative learning stochastic MPC with adaptive constraint tightening for building HVAC systems," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 11577–11582, 2020.
[4] B. Chen, Z. Cai, and M. Bergés, "Gnu-RL: A practical and scalable reinforcement learning solution for building HVAC control using a differentiable MPC policy," *Frontiers Built Environ.*, vol. 6, p. 174, Nov. 2020.
[5] A. Afram, F. Janabi-Sharifi, A. S. Fung, and K. Raahemifar, "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 Buildings*, vol. 141, pp. 96–113, Apr. 2017.
[6] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, D. D. G. Van, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.
[7] T. Zeng and D. Barooah, "An adaptive MPC scheme for energy-efficient control of building HVAC systems," *J. Eng. Sustain. Buildings Cities*, vol. 2, no. 3, pp. 1–12, 2021.
[8] O. Asvadi-Kermani and H. Momeni, "A constrained distributed time-series neural network MPC approach for HVAC system energy saving in a medium-large building," *J. Building Perform. Simul.*, vol. 14, no. 4, pp. 383–400, Jul. 2021.
[9] J. B. Rawlings, N. R. Patel, M. J. Risbeck, C. T. Maravelias, M. J. Wenzel, and R. D. Turney, "Economic MPC and real-time decision making with application to large-scale HVAC energy systems," *Comput. Chem. Eng.*, vol. 114, pp. 89–98, Jun. 2018.
[10] L. Yang, Z. Nagy, P. Goffin, and A. Schlueter, "Reinforcement learning for optimal control of low exergy buildings," *Appl. Energy*, vol. 156, pp. 577–586, Oct. 2015.
[11] Y. Wang, V. Kirubakaran, and H. Biao, "A long-short term memory recurrent neural network based reinforcement learning controller for office heating ventilation and air conditioning systems," *Processes*, vol. 5, no. 3, p. 46, Sep. 2017.
[12] Z. Wan, H. Li, and H. He, "Residential energy management with deep reinforcement learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–7.
[13] Z. Jiang, M. J. Risbeck, V. Ramamurti, S. Murugesan, J. Amores, C. Zhang, Y. M. Lee, and K. H. Drees, "Building HVAC control with reinforcement learning for reduction of energy cost and demand charge," *Energy Buildings*, vol. 239, May 2021, Art. no. 110833.
[14] Q. Wei, T. Li, and D. Liu, "Learning control for air conditioning systems via human expressions," *IEEE Trans. Ind. Electron.*, vol. 68, no. 8, pp. 7662–7671, Aug. 2021.
[15] Y. Du, F. Li, J. Munk, K. Kurte, O. Kotevska, K. Amasyali, and H. Zandi, "Multi-task deep reinforcement learning for intelligent multi-zone residential HVAC control," *Electr. Power Syst. Res.*, vol. 192, Mar. 2021, Art. no. 106959.
[16] X. Chen, L. Huang, J. Liu, D. Song, and S. Yang, "Peak shaving benefit assessment considering the joint operation of nuclear and battery energy storage power stations: Hainan case study," *Energy*, vol. 239, Jan. 2022, Art. no. 121897.
[17] Y. Wang, Y. Tong, M. Huang, L. Yang, and H. Zhao, "Research on air conditioning load virtual energy storage model based on demand side response," *Power Syst. Technol.*, vol. 41, no. 2, pp. 394–401, 2017.
[18] F. Hong, R. Wang, J. Song, M. Gao, J. Liu, and D. Long, "A performance evaluation framework for deep peak shaving of the CFB boiler unit based on the DBN-LSSVM algorithm," *Energy*, vol. 238, Jan. 2022, Art. no. 121659.
[19] V. J. Mawson and B. R. Hughes, "Optimisation of HVAC control and manufacturing schedules for the reduction of peak energy demand in the manufacturing sector," *Energy*, vol. 227, Jul. 2021, Art. no. 120436.
[20] J. Xue, X. Kong, B. Dong, and M. Xu, "Multi-agent path planning based on MPC and DDPG," 2021, *arXiv:2102.13283*. [Online]. Available: http://arxiv.org/abs/2102.13283
[21] S. Han, W. Zhou, S. Lü, and J. Yu, "Regularly updated deterministic policy gradient algorithm," *Knowl.-Based Syst.*, vol. 214, Feb. 2021, Art. no. 106736.
[22] T/ P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," 2016, *arXiv:1509.02971*. [Online]. Available: https://arxiv.org/abs/1509.02971
[23] T. Schreiber, S. Eschweiler, M. Baranski, and D. Müller, "Application of two promising reinforcement learning algorithms for load shifting in a cooling supply system," *Energy Buildings*, vol. 229, Dec. 2020, Art. no. 110490.
[24] H. Wu, "Control method of traffic signal lights based On DDPG reinforcement learning," in *Proc. J. Phys., Conf.*, 2020, vol. 1646, no. 1, Art. no. 012077.
[25] R. He, H. Lv, S. Zhang, D. Zhang, and H. Zhang, "Lane following method based on improved DDPG algorithm," *Sensors*, vol. 21, no. 14, p. 4827, Jul. 2021.
[26] S. Fujimoto, H. V. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. ICML*, 2018, pp. 1587–1596.
[27] J. Zhou, S. Xue, Y. Xue, Y. Liao, J. Liu, and W. Zhao, "A novel energy management strategy of hybrid electric vehicle via an improved TD3 deep reinforcement learning," *Energy*, vol. 224, Jun. 2021, Art. no. 120118.
[28] X. Shi, Z. Guo, J. Huang, Y. Shen, and L. Xia, "A distributed reward algorithm for inverse kinematics of arm robot," in *Proc. 5th Int. Conf. Autom., Control Robot. Eng. (CACRE)*, Sep. 2020, pp. 92–96.

[29] Q. Meng, Y. Li, X. Ren, C. Xiong, W. Wang, and J. You, "A demand-response method to balance electric power-grids via HVAC systems using active energy-storage: Simulation and on-site experiment," *Energy Rep.*, vol. 7, pp. 762–777, Nov. 2021.
[30] X. Du, H. Wang, Z. Du, and X. Jin, "Optimal operation strategy of VAV system based on early start-up and building energy storage," *Building Thermal Ventilation Air Conditioning*, vol. 5, pp. 7–10, Mar. 2007.
[31] Y. Du, H. Zandi, O. Kotevska, K. Kurte, J. Munk, K. Amasyali, E. Mckee, and F. Li, "Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning," *Appl. Energy*, vol. 281, Jan. 2021, Art. no. 116117.
[32] T. Liu, C. Xu, Y. Guo, and H. Chen, "A novel deep reinforcement learning based methodology for short-term HVAC system energy consumption prediction," *Int. J. Refrig.*, vol. 107, pp. 39–51, Nov. 2019.
[33] E. Diederichs, "Reinforcement learning–A technical introduction," *J. Auto. Intell.*, vol. 2, no. 2, p. 25, Aug. 2019.
[34] C.-E. Huang, C. Li, and X. Ma, "Active-disturbance-rejection-control for temperature control of the HVAC system," *Intell. Control Autom.*, vol. 9, no. 1, pp. 1–9, 2018.
[35] B. Wang, F. Zhu, W. Ji, and Y. Cao, "Modeling of load reduction potential of central air conditioning and analysis of influencing factors," *Autom. Electr. Power Syst.*, vol. 40, no. 19, pp. 44–52, 2016.

CHENHUI FU received the bachelor's degree in computer science and technology from Zhejiang Sci-Tech University, where he is currently pursuing the master's degree in software engineering. His research interests include machine learning, deep learning, and intelligent information processing.

YUNHUA ZHANG received Ph.D. degree in chemical engineering from Zhejiang University. He has been working as a Professor with Zhejiang Sci-Tech University, since 2003. His recent work focuses on the promotion and development of intelligent medicine through automated information processing. His research interests include software architecture, software engineering, and intelligent information processing.