Modelling and energy dynamic control for a ZEH via hybrid model-based deep reinforcement learning

Yanxue Li [1, 2, 3*], Zixuan Wang [1], Wenya Xu [1], Weijun Gao [1, 4], Yang Xu [1, 4], Fu Xiao [2]

1. Innovation Institute for Sustainable Maritime Architecture Research and Technology, Qingdao University of Technology, Qingdao, 266033, China
2. Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong
3. Tongji Architectural Design Group Co., Ltd, Shanghai, 20092, China
4. Faculty of Environmental Engineering, The University of Kitakyushu, Kitakyushu, 808-0135, Japan

[*]Corresponding author: liyanxue@qut.edu.cn; Tel: 86+15689978251

Abstract: Efficient and flexible energy management strategy can play an important role in energy conservation in building sector. The model-free reinforcement learning control of building energy systems generally requires an enormous amount of training data and low learning efficiency creates an obstacle to practice. This work proposes a hybrid model-based reinforcement learning framework to optimize the indoor thermal comfort and energy cost-saving performances of a ZEH (zero energy house) space heating system using relatively short-period monitored data. The reward function is designed regarding energy cost, PV self-consumption and thermal discomfort, proposed agents can interact with the reduced-order thermodynamic model and an uncertain environment, and makes optimal control policies through the learning process. Simulation results demonstrate that proposed agents achieve efficient convergence, D3QN presents a superiority of convergence performance. To evaluate the performances of proposed algorithms, the trained agents are tested using monitored data. With learned policies, the self-learning agents could balance the needs of thermal comfort, energy cost saving and increasing on-site PV consumption compared with the baselines. The comparative analysis shows that D3QN achieved over 30% cost savings compared with measurement results. D3QN outperforms DQN and Double DQN agents in test scenarios maintaining more stable temperatures under various outside conditions.

Keywords: ZEH, Thermal comfort, Deep reinforcement learning, Energy management strategy

Introduction

To meet the Paris Agreement Goal of limiting the rise in global average temperatures to well below 2 °C above the preindustrial level [1]. Many countries, including China, U.S. and Japan, have set their ambitious carbon neutrality goal [2]. The building sector accounts for approximately 40% of global energy consumption and contributes to 38% of energy-related CO_2 emissions [3]. Consequently, decarbonizing the building sector can play a key role in achieving the carbon neutrality transition [4, 5].

In October 2020, Japan declared its commitment to achieving social carbon neutrality by 2050 [6]. To accelerate the low-carbon transition of the built environment sector, Japan's strategic energy plan set a goal of achieving average zero emissions in newly-constructed houses by 2030 [7, 8]. The Japanese government has established standards, guidelines and subsidy programs for promoting the widespread adoption of ZEB (zero energy building). ZEB concept is increasingly gaining attention, main efforts include high insulation standards, development of on-site renewable resources, high energy efficiency equipment, and advanced information and communication technologies, which have been made towards improving building energy efficiency [8-10].

Previously, buildings mainly conducted their energy demand control in passive or static approaches [11], conventional proportional-integral-derivative (PID) control presents a large delay. With the wide integration of distributed energy generations, buildings have become energy prosumers [12]. Moreover, with the wide applications of advanced metering infrastructure and information and communication technologies, prosumers can proactively and intelligently participate in managing energy consumption in

response to signals [13, 14]. Seasonal heating and cooling demands share a large percentage of building energy use, accounting for 40% of the building energy usage [15], and contribute greatly to variations in building load. The uncertainty of real-time on-site generation and load changes may bring challenges to building energy operations. Dynamic control strategy for building energy systems has emerged as a promising alternative to improve energy-saving or economic performances [16].

Literature review

The energy-saving standard of low-energy and zero energy building has been improved, building thermal inertia increases and energy losses caused by load shifting strategies such as preheating and precooling decrease [17, 18]. The energy flexible building can effectively use thermal mass as virtual energy storage to provide energy flexibility, which has drawn considerable attention [19-21]. In turn, it helps mitigate the gap between on-site intermittent renewable generation and electrical energy demand [18]. Kim, D. et, al found that discharging process of building thermal mass enabled a 4-hour load reduction during the peak period [22]. Furthermore, utilizing building thermal mass to harness energy flexibility requires nearly zero capital cost, cost-saving benefits can be achieved with actions such as load shifting or peak shaving [17, 23, 24].

Building thermal load control mainly aims to minimize energy usage while maintaining occupant thermal comfort [25, 26]. In fact, stochastic occupancy, uncertainties of on-site generation and weather conditions have a significant impact on the operational performances of building energy systems [27, 28]. Building heating and cooling systems are typical nonlinear time-dependent multivariable systems [29, 30]. The

conventional rule-based controllers use on/off and proportional-integral-derivative (PID) control to maintain given set point temperatures, presenting a less efficient performance in reducing operational cost of building HVAC system [31], a large time delay of PID can lead to room comfort violation [32]. Model predictive control considering various operational constraints and prediction of disturbances, is usually adopted to optimizes the operation of the system [33]. To achieve an efficient building energy system management, prediction model combined physical model is usually used to determine optimal control actions. Typical building RC (resistances and capacitances) models include white-box model, black-box model and gray-box model. White-box model presents a good prediction accuracy. However, the construction of white-box model needs a full understanding of the system's dynamics, the number of parameters is huge and process is time-consuming [34]. Black-box model is a data-driven model, the accuracy and reliability of the modelling approach are significantly affected by the quality of collected monitored data, long training period limits the application of this model [32]. The accurate scheduling models require many variables and constraints, and reducing the complexity of mathematical programming modeling embedded in energy management is attracting research interest [35]. The reduced-order grey-box modeling approach takes advantage of black-box and white-box models, presenting a trade-off between physics and historical data. It has been widely applied to load estimation, control and optimization of building HVAC systems [36].

MPC faces more challenges when dealing with the stochastic process, the applications of MPC are limited due to high reliance on the accuracy of the built model, and the long

prediction horizon may result in an infeasible computational burden for MPC controller [37]. The Reinforcement Learning (RL) algorithm is a class of machine learning algorithms based on a trial-and-error learning approach. RL shows an advantage in that accurate physical models of the control environment are not required, and it presents a more robust performance in comparison with the MPC approach [38]. RL strategy as an adaptive feedback controller is well suited to handle sequential decision-making tasks, RL agent interacts with a stochastic environment experientially, and learns to decide based on a reward mechanism [39, 40].

Due to the curse of dimensionality as the states and actions spaces increase exponentially, approximations must be made [41]. In recent years, reinforcement learning algorithm combined with artificial neural network is becoming a solution for complex system control [42, 43]. The deep reinforcement learning (DRL) is a powerful approach that offers promising dynamic energy system management through interacting with the control objective and the environment [39, 44]. Generally speaking, RL methods can be classified into two categories: model-based and model-free, the model free RL is a data-driven method [45]. Model-free RL can be trained to work without prior knowledge [46]. However, the training process of model-free approach is time-consuming to learn an optimal control policy. And the approach requires an enormous amount of training samples for the policy function to converge to a high-quality optimum [47]. In [48], a model-based RL was proposed for eco-driving strategies, achieving optimal energy-saving performance of 93.8% of exact dynamic programming result. Authors [49] adopted a new approach that combined model-free

and model-based RL controller for autonomous underwater vehicles, results demonstrate the robustness and effectiveness of the approach. Uncertainty of renewable generation and the stochastic nature of electricity load challenge the efficient control of isolated microgrid, the result verified the effectiveness of model-based RL algorithms [50]. The model-based RL has proved to be an effective way to manage complex energy systems in terms of stability, generality and adaptability [51]. Considering the stochastic parameters, such as on-site renewable energy generation, weather conditions and energy consumption, the environments of the building energy system are heterogeneous and difficult to find an optimal control strategy [52, 53]. To address the scalability issues, this work proposes a model-based RL framework, the model of system dynamics is learned by interacting with the environment, then offers optimal control policy. The main contributions of this work are summarized as follows:

- We proposed a novel controller through synthesizing model-based and data-driven learning methods for ZEH heating supply system management, the approach learns the system dynamics and a reward function to offer optimal control policies, improve the interpretability.
- We formulate the room heating supply as a Markov decision process-based task problem, with a goal of exploring the demand-side energy flexibility, optimizing the space heating cost and local PV consumption in a real ZEH.
- Design and test different deep RL agents in managing building heating system, assess their performances in comparison with a baseline using monitored data.

Methodology

Reinforcement learning algorithm

Reinforcement learning is a branch of machine learning algorithms based on a trial-

and-error learning process[38]. The interactions between the learning agent and environment can be formulated as Markov decision process (MDP) in discrete time-space. And the stochastic control can be described by a four elements tuple $\langle S, A, P, R, \gamma \rangle$. where S represents state space and A represents a set of actions that the agent can perform, P is the state-transition probability, which means the probability of the agent taking a random action a may lead the current state s to another state s', R represents the reward function, which is usually the objective function in a control problem. And finally, the discount factor $\gamma \in [0,1]$ decides how much importance we give to future rewards and immediate rewards depending on the specific case. If $\gamma = 0$, the agent is only concerned on maximizing immediate rewards. As γ increases, the agent will be increasingly focused on future rewards.

At each time step t, the agent observes the environmental state $s_t \in S$ and chooses an action $a_t \in A$ based on policy π. Then the agent will receive the reward $r(s_t, a_t)$, and the system will evolve to another state $s_{t+1} \in S$. The *policy* π is probability distribution for each possible action $a \in A$ been selected in a state $s \in S$, and it can be performed deterministic or stochastic based on the specific algorithm.

State-value function and State-action value function (Q-function)

To solve sequential decision questions, the purpose of RL algorithms is to learn a value function $v_\pi(s)$ or state-action value function $Q^\pi(s,a)$, it is also called Q-function. The first refers to the value of a state s under a policy π, which indicates the expected return when starting in the state s and continuing with policy π. While the second refers to the expected return when starting in the state s with action a according to

policy π , the *state-value function* is defined as Eq. (1):

$$V_\pi(s) = E_\pi\left[\sum_{k=0}^{N}\gamma^k R_{t+k+1}\middle|s_t = s\right] \tag{1}$$

Where, E_π denotes the expected value under a policy π in the state s_t . N is the final step in an episode and t is any time step, the value of the discount factor γ need to be tuned to balance the future and immediate reward.

The *action-value function* is strong correlation with *state-value function*, it can be defined as follows (see [38] for more details):

$$Q_\pi(s,a) = E_\pi\left[\sum_{k=0}^{N}\gamma^k r_{t+k+1}\middle|s_t = s, a_t = a\right] \tag{2}$$

Where, E_π is the expected value that the agent chooses an action a_t based on the policy π in the state s_t .

Bellman Equation

Given the MDP problem, the learning agent needs to find the optimal policies π^* and value functions. Value functions are different according to the various policies, the optimal value function is the one that gets the maximum value compared to all the other value functions. It can be easily computed by taking the maximum of the Q-function as follows Eq. (3):

$$Q_\pi^*(s,a) = \left[r(s,a) + \gamma \max Q^\pi(s_{t+1}, a_{t+1})\right] \tag{3}$$

The above equation is called a *Bellman optimality equation*, and it indicates the recursive relation between a value of state s_t performed an action a_t under the policy π and its subsequent state and the average overall possibilities [54].

RL algorithms can be categorized into different groups: Firstly, an agent can be divided

into model-based and model-free algorithm; the model-free approach learn the policy from historical data while the model-based method requires estimating the environment's transition model and learning the policy based on this model [55]. Secondly, it can also be divided into two categories, on-policy, and off-policy. This means we utilize a greedy method to select the best choice available from the current policy even though this option may not be the best during the overall phase. In the off-policy, there have two policies, behavior policy and target policy. We use behavior policy to explore any states and actions and store them in memory, whereas a target policy is a greedy policy that has a maximal value derived from the memory. On-line learning implies that the agent learns from interacting with the environment, whereas off-line learning is applied until it has been sufficiently trained. In real-world scenarios where data is difficult to collect and slow to generate (e.g., HVAC system), the off-policy method can use historical data to generate an optimal policy to achieve specific goals, while on-policy learning has a high learning cost characteristic in control systems. Therefore, we choose off-policy as the main research method in this paper.

Q-learning Algorithm

Q-learning is a very popular off-policy TD (temporal-difference) control algorithm based on Q-function without model dynamics in advance, the RL agent policy is decided by a state-action table, sometimes called *Q-table*, it uses the epsilon-greedy policy to choose the action in a state while updating the *Q-value* of the previous state based on the following equation Eq. (4):

$$Q_\pi(s_t,a_t) \leftarrow Q_\pi(s_t,a_t) + \alpha\left[r(s_t,a_t) + \gamma \max Q_\pi(s_{t+1},a') - Q_\pi(s_t,a_t)\right] \quad (4)$$

Where $Q_\pi(s_t, a_t)$ presents the Q-value of the previous state what we expected, $r(s_t, a_t) + \gamma \max Q_\pi(s_{t+1}, a')$ is the present reward and discount future reward, $\alpha \in [0,1]$ is the learning rate.

In order to receive the greatest cumulative rewards. The *Bellman equation*, which is a sum of the present reward and the maximum discounted future benefits, can be used to define the optimal *Q-value* for action a_t that was chosen at state s_t :

$$Q_\pi^*(s_t, a_t) = r(s_t, a_t) + \gamma \max Q_\pi(s_{t+1}, a_{t+1}) \tag{5}$$

Deep Q-learning Algorithm

As the name suggests, deep reinforcement learning combines deep learning with reinforcement learning. It is bleeding the technique of getting rewards based on the action interacting with the environment, and the ability to utilize a neural network for learning feature representations from deep learning. A *Q-table* is typically used to store the value of states and actions where the environment can be described by a limited range of discrete states and actions, but when the state and action space is infinite, starting in about 2013 [56], Google Deepmind has proposed a variant of Q-learning algorithm which is the first deep learning model to successfully learn control policies directly from high-dimensional sensory input using Q learning estimate future rewards through value function and achieved human-level results when playing six of Atari games. In addition, the DQN algorithm was improved in 2015 [57] to use a replay buffer and two neural networks to overcome the instability issues of previous approaches.

The pseudo-code of DQN is given in **Algorithm 1**. The main idea of a replay buffer is that we save each transition information as (s_t, a_t, s_{t+1}, r_t) in a buffer $\mathcal{B}$ and train the deep

Q network with a random batch of m transitions sampled from buffer instead of the latest transition, to reduce the overfit with correlated experiences at each iteration. At the beginning of training, memory $\mathcal{R}$, weight ω and ω^- of neural networks should be initialized, agent obtains the initial observation of state s_1 and preprocessed sequenced $\psi_1 = \psi(s_1)$, the $\varepsilon - greedy$ policy is used for agent to trial-and-error, selecting the action a_t in the possibility range of ε with the maximum *Q-value* output by the current Q network and randomly choose an action within the $(1-\varepsilon)$ possibility, then execute the action, compute reward r_t and observe the next station, We then calculate the squared error between the target and the predicted value of the neural network in the loss function and update the parameters of the current neural network using the gradient descent method, then the weight ω^- will be frozen for several time step C and replaced by copying the actual Q network weight ω stabilizes the training.

Algorithm 1: Deep Q-learning

Initialize replay memory $\mathcal{R}$ to capacity $\mathcal{B}$
Initialize current network $Q_c(\chi(s_t), a; \omega)$ and target network $Q_t(\chi(s_t); \omega^-)$ with random weights ω and ω^-
for $episode = 1, M$ **do**
  Obtain initial observation of state s_1 and preprocessed sequenced $\psi_1 = \psi(s_1)$
  for $t = 1, T$ **do**
    With probability ϵ select a random action a_t
    therwise select $a_t = \max_a Q_c(\psi(s_t), a; \omega)$
    Execute action a_t and compute reward r_t and observe the new state s_{t+1}
    Store transition $(\psi_t, a_t, r_t, \psi_{t+1})$ in $\mathcal{B}$
    Randomly sample a mini-batch of m transitions $(\psi_i, a_i, r_i, \psi_{i+1})$ from $\mathcal{R}$
    Set $y_i = \begin{cases} r_i & \text{for terminal } \psi_{i+1} \\ r_i + \gamma \max_{a'} Q_t(\psi_{i+1}, a'; \omega^-) & \text{otherwise} \end{cases}$
    Update ω by minimizing the loss:
    $L_i(\omega_i) = E\left[\left(y_i - Q_c(\psi_i, a_i; \omega_i)\right)^2\right]$
    Update ω^- using the sampled policy gradient:
    $\nabla_{\omega_i} L_i(\omega_i) = E\left[\left(r_i + \gamma \max_{a'} Q_t(\psi_{i+1}, a_{i+1}; \omega^-) - Q_c(\psi_i, a_i; \omega_i)\right) \nabla_{\omega_i} Q_c(\psi_i, a_i; \omega_i)\right]$
    Every C steps reset $\omega^- = \omega$
  end
end

The framework of DDQN, Dueling DQN and D3QN

According to the above description, Double DQN, Dueling DQN, and D3QN are the derivative method of DQN. The detailed differences between these algorithms are described as follows.

Double DQN Algorithm

The max operator in standard DQN uses the same Q networks to select and evaluate actions, this makes it more likely to select overestimated values [57]. To fix this, Ref [58] describes the kernel idea behind the DDQN (Double DQN) which decouples the action selection from the action evaluation. The main neural network determines the best next action among all the available next actions, and then the target neural network evaluates this action to know its Q-value, the target Q value is defined as Eq. (6):

$$TargetQ = r(s_t, a_t) + \gamma \cdot Q(s_{t+1}, argmaxQ(s_{t+1}, a'; \omega_t); \omega_t^-) \quad (6)$$

Where, $r(s_t, a_t)$ is reward; γ is attenuation coefficient and two Q functions each with different weights, a Q function with weights ω_t to select the action in the argmax while the other function with a set of weights ω_t^- to evaluate the action.

Dueling DQN Algorithm

The network structure of Dueling DQN is identical to DQN, but it produces results in different ways, DQN outputs Q value directly, while Dueling DQN network produces separate the predictive state value function $V(s_t; \omega, \beta)$ and relative action advantage function $\mathcal{A}(s_t, a_t; \omega, \alpha)$. $V(s_t; \omega, \beta)$ presents the value of a state and $\mathcal{A}(s_t, a_t; \omega, \alpha)$ is the advantage of action a in a specific state s at time t. The two functions are combined by using the aggregate layer to address the issue of overestimating and accelerate the training rate of the network. Therefore, the Q-

function can be described as:

$$Q(s_t, a_t) = \mathcal{A}_{\pi^*}(s_t, a_t) + V_\pi(s_t) \tag{7}$$

For the optimal policy $_{a_{t+1} \in A} Q(s_t, a_t)$, $\mathcal{A}(s_t, a_t) = 0$, then $Q(s_t, a_t) = V_\pi(s_t)$, and Dueling DQN network outputs is described by:

$$Q(s_t, a_t; \omega, \alpha, \beta) = V(s_t; \omega, \beta) + \mathcal{A}(s_t, a_t; \omega, \alpha) \tag{8}$$

Where, Q is the value of current network, ω is Q network parameters, S is the current state, a is the current action, while α and β are the fully connected layer parameters of the two streams.

In order to solve the problem that is difficult to map from Q values to unique $V(s_t; \omega, \beta)$ and $\mathcal{A}(s_t, a_t; \omega, \alpha)$ values, in [59] proposed an approach for making the advantage function estimator has zero advantage while making a decision. This can be achieved by subtracting the average $\bar{\mathcal{A}}$ from the action value, through training V and $\mathcal{A}$ is more effective and robust than the standard DQN network structure, the Q function can be expressed as Eq. (9):

$$Q(s_t, a_t; \omega, \alpha, \beta) = V(s_t; \omega, \beta) + \left(A(s_t, a_t; \omega, \alpha) - \frac{1}{|\mathcal{A}|} \sum_{a'} A(s_t, a'; \omega, \alpha) \right) \tag{9}$$

Dueling Double-Deep Q Networks Algorithm

D3QN (Dueling Double-Deep Q Networks) algorithm has combined Double-DQN and Dueling-DQN [60], this technique can effectively address several shortcomings of the DQNs, such as the overestimation problem. Thus, the target Q value of Q network Y_t^{D3QN} is the same as DDQN, it can be denoted by Eq. (6) before and appropriate parameters should be trained by minimizing the loss function, which can be expressed as following:

$$L^{D3QN}(\omega_t) = E[(Y_t^{D3QN} - Q_t(s_t, a'; \omega_t, \alpha, \beta))^2] \tag{10}$$

D3QN updates the training parameters ω_t of the Q network with stochastic gradient descent and copies ω_t to the target network's parameters every fixed-step. Update parameters in the training process can be formulated as:

$$\omega_{t+1} = \omega_t + \alpha \cdot E[(Y_t^{D3QN} - Q_t(s_t, a'; \omega_t, \alpha, \beta)) \cdot \frac{\partial Q_t(s_t, a'; \omega_t, \alpha, \beta)}{\partial \omega_t}] \quad (11)$$

DRL agent design

In this section, the indoor thermal management problem is formulated as an MDP, which typically involves the agent (e.g., HVAC system) interacting with the environment that contains lots of states (e.g., indoor or ambient temperature by performing an action (e.g., HVAC system power consumption) from action space to another state and receive an immediate reward based on the reward function (e.g., indoor thermal comfort), the agent will understand whether the action was good or bad, the elements for the MDP formulation are described as follows:

System state

In this work, the state space utilizes a set of physical parameters that represent the realistic environment, in the process of managing the indoor environment, the state space observed by the agent includes the time of day T^{info}, indoor temperature T^{in}, envelope temperature T^{en}, outdoor temperature T^{o}, on-site PV generation P^{gen} and solar irradiance. The time information enables the DRL algorithm to learn the patterns of time-related activities, such as occupancy behavior, electricity price and photovoltaic power generation. Therefore, the state at each time t can be described as:

$$S = [T_t^{info}, T_t^{in}, T_t^{en}, T_t^{o}, P_t^{gen}, price_t] \quad (12)$$

Control actions

The agent interacts with a controlled environment and chooses an action based on

various states. In the process of managing the indoor environment, in order to properly control the air conditioner, we discrete action space A of the agent as:

$$A = [0, 25\%Q_{max}, 50\%Q_{max}, 75\%Q_{max}, Q_{max}] \tag{13}$$

Where, Q_{max} is rated power of air conditioner in heating mode, kW.

Reward function design

The goal of the proposed deep reinforcement learning controller is to identify an optimal policy that can minimize electricity costs while maintaining indoor thermal comfort. That is to say, the objective is to increasing directly local PV consumption while maintaining optimal indoor temperature. In order to achieve the goal, the developed reward function was composed of three components: (1) the penalty for the space heating electricity cost C_t, (2) penalizing the indoor discomfort when the indoor temperature is out of comfort range $T_{a_t, s_{t+1}}$, (3) the reward for the local consumption of PV generation P_t^{con}. The reward function is defined as follows:

$$R_t = -\alpha \cdot C_t - \delta \cdot T_{a_t, s_{t+1}} + \beta \cdot P_t^{con} \tag{14}$$

$$T_{a_t, s_{t+1}} = 1 - exp(-0.5(T_t^{in} - T_m)^2) + penalty(T_t^{in}) \tag{15}$$

$$penalty(T_t^{in}) = \begin{cases} -\lambda(T_t^{in} - T_{lower\_bound})^2 & if \quad T_t^{in} < T_{lower\_bound} \\ -\lambda(T_{upper\_bound} - T_t^{in})^2 & if \quad T_t^{in} > T_{upper\_bound} \\ 0 & if \quad T_{lower\_bound} \le T_t^{in} \le T_{upper\_bound} \end{cases} \tag{16}$$

Where, the negative reward means the agent is penalized, α, δ and β are the hyperparameters, their value is set to ensure the equation is homogeneous, thus improving the robustness of the RL agents. $T_{a_t, s_{t+1}}$ is the penalty value of indoor thermal discomfort which consists of two parts, T_t^{in} is room temperature at each time step t, T_m is the average desired temperature, Eq. (11) used the method has been proposed in [61].

Case Study

In order to examine the feasibility and effectiveness of proposed RL algorithms, this work selected a two-story ZEH in Kyushu, Japan. The total floor area is 105m^2, the envelope of the selected ZEH is well insulated with infilled glass wool and presents a high thermal insulation performance and the average overall heat loss rate of the house envelope U_a is $0.58W/(m^2 \cdot K)$, presenting a high potential of thermal flexibility, detail design information of the ZEH is described in Table 1. The collected operational data of a ZEH at 30-minitute interval that lasts from 1st, January to 30st, March in 2020 as shown in Fig.1. Rooftop PV generation (a 4.8kWp photovoltaic system) and on-site consumption are demonstrated in Fig.1 (a), indoor temperature, outdoor temperature, and power consumption profiles of the air conditioner are shown in Fig.1 (b). In this work, we set the coefficient of performance of the air conditioner in space heating mode to 2.5.

Table 1 Main characteristics of ZEH

Feature	Details
Established Year	2017
Envelope heat loss coefficient	$U_a: 0.58W/(m^2 \cdot K)$
Thermal insulation material and characteristic	Wall: glass wool 120mm and glass board 12 mm; Roof: glass wall 100mm; Floor: glass wall 67mm; Thermal conductivity $0.04W/(m \cdot K)$
Ventilation rate	Mechanical ventilation, $0.5ac/h$
Window	Low-E pair glass with plastic combined aluminum sash
Window-to-wall ratio	0.16

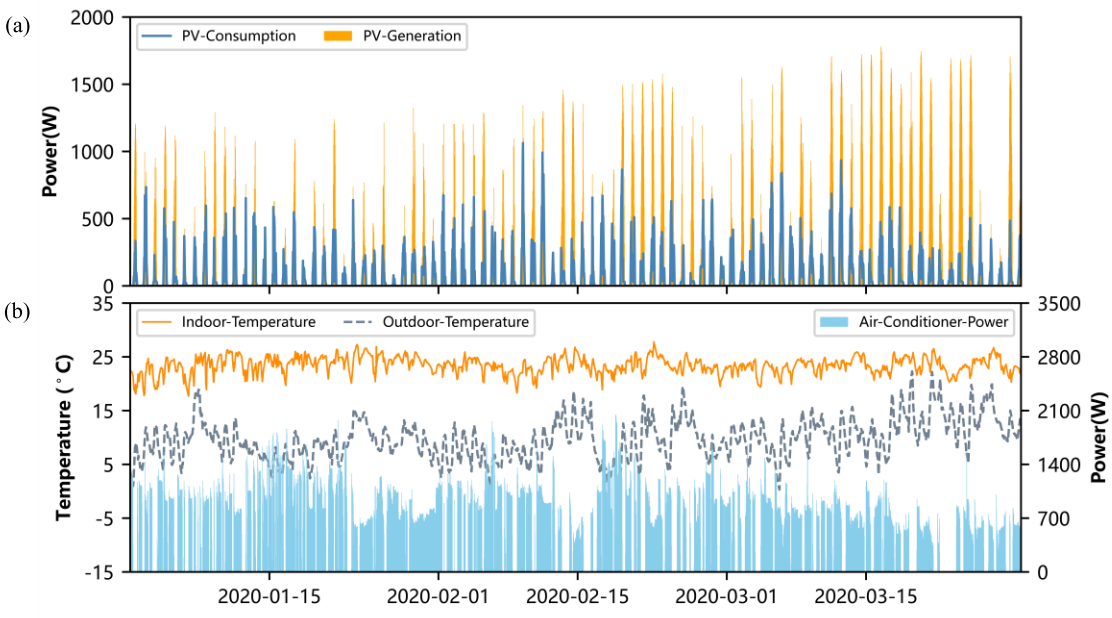

Fig. 1. Historical data of the ZEH at 30-min interval from 1st January to 30st March in 2020

Building dynamic model

Establishing a thermodynamic building model can be quite complex depending on the considered components and required accuracy degree, and a complex model would increase the computational load of the RL iterative learning process. To manage the space heating energy system, the physical model adopts a reduced-order RC network. The building physical information and measured data could be used to calculate the model parameters and assess the accuracy by the RMSE (Root Mean Squared Error) indicator, which has been used in previous works [18, 62]In this model, the following presumptions are made:

- The whole building is modeled as a single zone, air conditioning system operates to heat the room.
- The heat transfer between the whole building and the ground is negligible due to the well-insulated ZEH, which shows the excellent ability to retain indoor heat energy.
- Assuming that there is no difference between the temperature of the outside surface

of the wall with ambient temperature.

The considered components of the thermodynamic model are shown in Fig. 2. R and C present the equivalent thermal resistance and capacitance, respectively, T presents the temperature, $°C$. We adopt a simple 2R2C network model, the values of parameters such as R and C, are determined according to available design information and measured data, the accuracy has been verified in our previous work [18]. The electricity price information is obtained from (https://www.kyuden.co.jp/user_menu_plan_denka-de-night.html), and the price scheme is shown in **Table 2.**

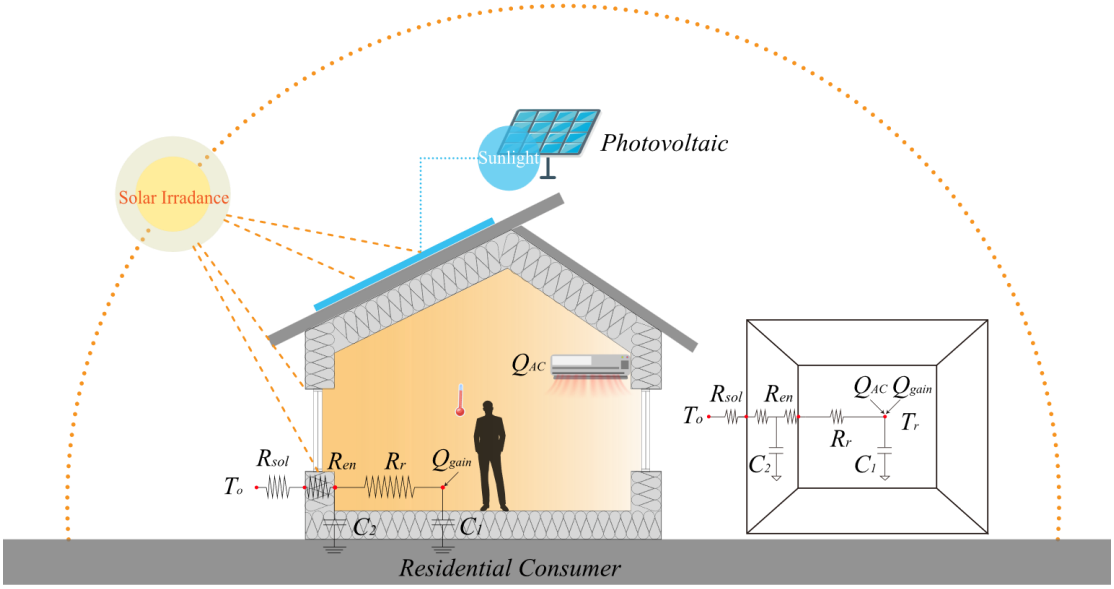

Fig. 2. The reduced-order RC network for the ZEH thermodynamic model

Table 2. Dynamic electricity price

Days	Periods	Midday (Yen/kWh)	Nighttime (Yen/kWh)
Weekday	Summer, winter	26.84	13.21
	Spring, autumn	23.95	13.21
Weekend	Summer, winter	21.22	13.21
	Spring, autumn	17.82	13.21

Agent training setup

These agents were implemented with the Pytorch 1.11, and the simulation framework was utilized to generate an optimal policy for the agent to choose actions based on the following hyperparameters:

- mini-batch size: we train the deep neural network with transitions (i.e (s_t, a_t, s_{t+1}, r_t)) which select a random batch from the replay buffer.
- learning rate: also called step size, α represents the convergence speed during the training process.
- neurons number: each network contains three layers, each containing a group of neurons.
- discount factor: the parameter indicates that the return at time t in the loss function will depend on how much future reward there will be.
- reward function weights: weights access the value of electricity cost vs. thermal comfort vs. consumption rate of renewable energy.

The hyperparameters can significantly affect the behaviors of DRL agent. As described in Table 3, the value of input parameters was appropriately determined in accordance to researcher's experience and target (energy cost saving, local PV consumption ratio and thermal comfort control). The framework for simulation training was described above. Given the available data, 10-week data (3628 control steps, at 30-minute intervals) were utilized for training the DRL agents, and we selected two weeks to test the adaptability of the trained model, the data of a cold week is chosen from 17th to 23rd in February, the data of warm week is collected from 16th to 22nd in March in 2020. The Simulation is carried out on a computer with an Intel® Core™ i7-9700 CPU @ 3.00GHz and 16.0 GB RAM, each training event was iterated 500 episodes for the configuration of the hyperparameters.

Table 3 Hyperparameters of the DRL agent

Category	Parameters	Values
Environment	λ	0. 2
	α	0.00008
	δ	0.01
	β	0.00001
	T_m	22.5
	T_{upper_bound}	24
	T_{lower_bound}	20
Algorithm Parameters	DNN architecture	1 layer
	Neurons per hidden layers	32
	DNN optimizer	Adam
	Optimizer learning rate	0.0001
	Batch size	48*7
	Episode length	3628 Control Steps
	Memory size	48*60
	Target model update	100 Steps
	Training episodes	500
	Discount factor	0.9
	ϵ Start	0.9
	ϵ End	1.0

This work proposes four deep reinforcement learning algorithms: DQN, Double DQN, Dueling DQN, Dueling-Double DQN. The control logic has been described in Fig. 4. Firstly, by taking a full observation of the data and evaluating the trade-off energy consumption, indoor thermal comfort, and photovoltaic local consumption ratio, we set actions as a discrete space of the air conditioner power consumption, making it easy to realize in engineering programs [38]. To explore the energy flexibility of building thermal mass, the designed controllers maintain a thermal comfort range and the chosen actions result in a difference between actual power consumed and power demand. The self-learning agent guided by the physics model optimizes space heating power consumption by interacting with environments under various conditions. The impacts

of chosen actions (power input for space heating system) on indoor thermal comfort can be presented by modeling the reduced-order RC network in state-space equation form. The penalty for indoor thermal discomfort violation is described in **Fig.3**, the thermal penalty sets lower and upper temperature bounds $[T_{lower\_bound}, T_{upper\_bound}]$ described in the yellow area to maintain a desired indoor temperature range. The reward is shaped like a bell curve and we motivate the agent to get closer to the optimal value T_m in the red dash line and then rapidly decrease when the temperature is far from the defined comfort range.

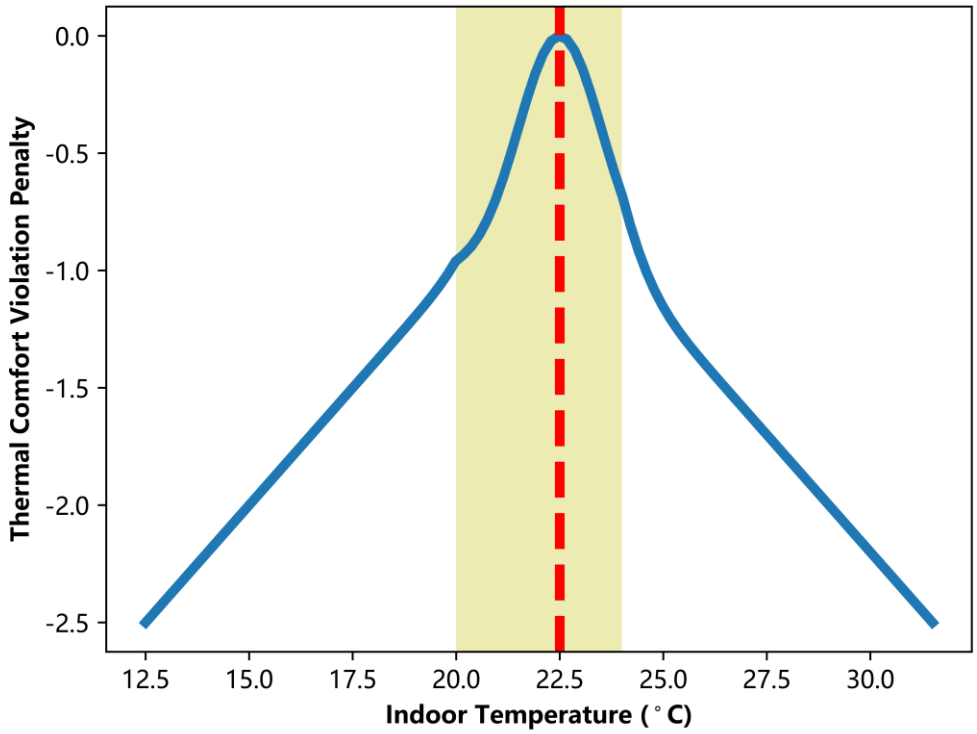

Fig. 3. Graph of the thermal comfort penalty function, parameters $\delta = 0.01$ and $\lambda = 0.2$, we set no trapezoid penalty when indoor temperature is in 20 $°C \sim 24 °C$ range.

In the controlling process of the indoor environment management, the trained current Q network chooses an action a_t based on the observed 6-dimensional state vector $S_t = [T_t^{info}, T_t^{in}, T_t^{en}, T_t^{o}, P_t^{gen}, price_t]$, including the time of day T^{info}, indoor temperature T^{in}, envelope temperature T^{en}, outdoor temperature T^{o}, PV generation P^{gen}, real-time

electric price $price_t$, the action executed by the air conditioning system, then the environment transits to the next state s_{t+1} and the agent receives the reward r_t, each transition information as (s_t, a_t, s_{t+1}, r_t) saved in the reply buffer. During the process of training, the current Q network was trained with a random mini batch of transitions sampled from the reply buffer. Squared error is calculated between a target and the predicted value of the target and current neural network in the loss function, and parameters in the current neural network will be updated by the gradient decent method, the weights ω^- of target Q network will be frozen for several time step and replaced by copying the actual Q network weights ω stabilizes the training.

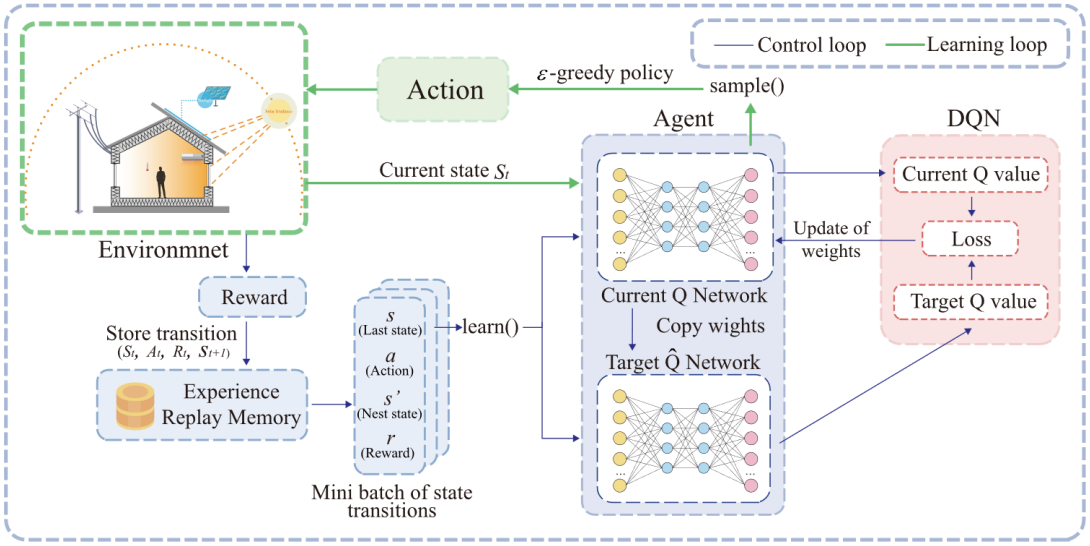

Fig. 4. The flowchart of deep RL based control approach

Result and Discussion

Results of training process

The learning process of the DRL agent is an episodic task, the convergence of the learning process of the agent can be evaluated by the episode-reward plot. Fig. 5 describes the exploration performances of different DRL algorithms as introduced in

the method part, the episode reward curves indicate that each trained agent can learn toward convergent policy. It can be seen from Fig. 5 that the training process can be mainly divided into two processes, rewards increase rapidly during the initial stage of the training process (before 60 training episodes), then value of reward value becomes more and more stable, reaching a convergence stage at the end of the training phase. It can note that the curves of trained DQN and DDQN agents present greater fluctuations, their stability is relatively low, D3QN shows the highest cumulative reward value in the end of the training process. Overall, it can confirm that proposed DRL agents can find optimal energy management strategies for the test ZEH.

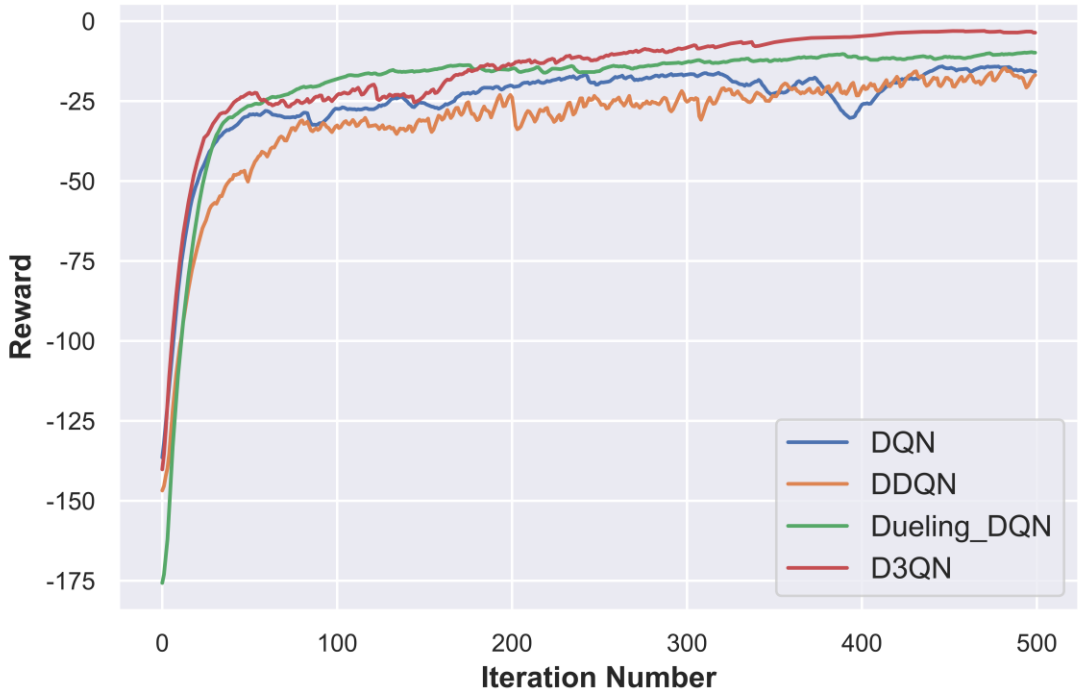

Fig. 5. Convergence performances of proposed DRL agents during training process

Comparison of testing results

In order to examine the adaptive capability of proposed DRL algorithms, we test the trained DRL agents, and compare their performances by testing data of selected two

weeks, a cold week lasts from 17th February to 23rd February in 2020 and a warm week lasts from 16th March to 23rd March in 2020. Fig. 6 describes indoor temperature plots of testing DRL agents and measured indoor temperature under baseline PI control rule during a cold week, the two black dash-dot lines define the lower and upper bounds of acceptable room temperature range, and the red dot line presents the desired indoor temperature T_m. The measured indoor temperature under the PI controller is depicted in a gray line, the purple curve is the corresponding outdoor temperature. It can be observed that the fluctuations of indoor temperature under the PI regulation are greatest. Overall, the proposed DRL agents can well control indoor thermal comfort, the maximum simulated temperature of DQN and DDQN is slightly larger than the upper thermal comfort limitation. Simulated indoor temperature profiles of Dueling-DQN and D3QN exhibit similar temporal distributions, and their average value of indoor temperatures is close to the defined T_m.

In terms of increasing local PV consumption ratio, Fig. 7 depicts a comparison of the simulated PV consumption profile by each DRL algorithm, yellow bars represent the actual PV generation, the red dash lines show the measured PV consumption, and the blue curves depict the PV consumption under each DRL agent control. As observed, the simulated temperature profiles of DQN and Double DQN agents show similar performances, their values are slightly outside the comfort range occasionally. Overall, results demonstrate that DRL agents learn to harness thermal flexibility to consume more PV production, thus providing a cost-saving opportunity while maintaining the thermal comfort of the ZEH.

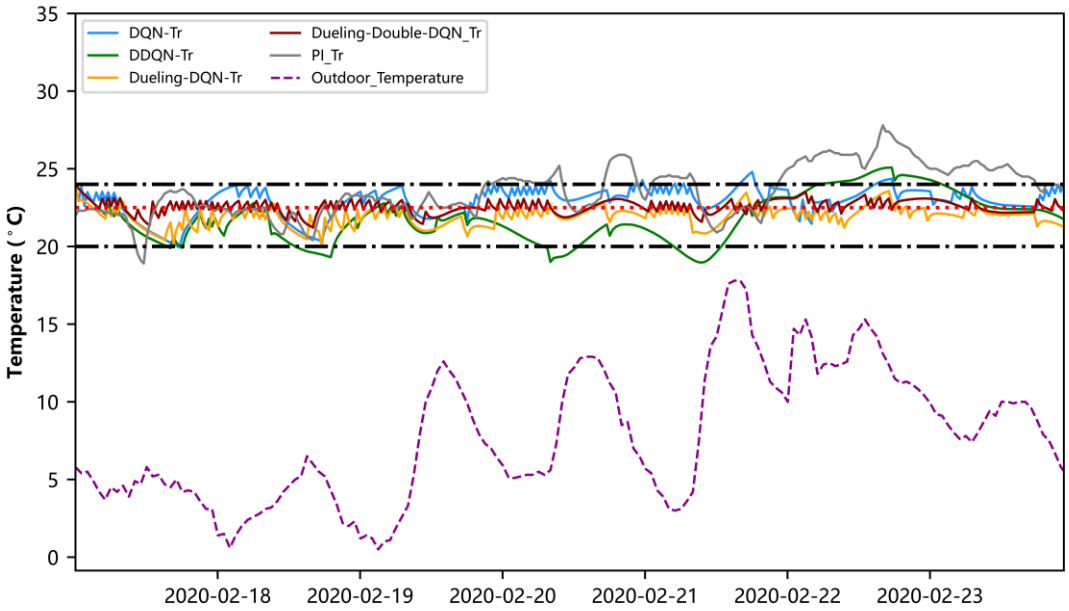

Fig. 6. Testing results of indoor temperature profiles under investigated control strategies during a cold week

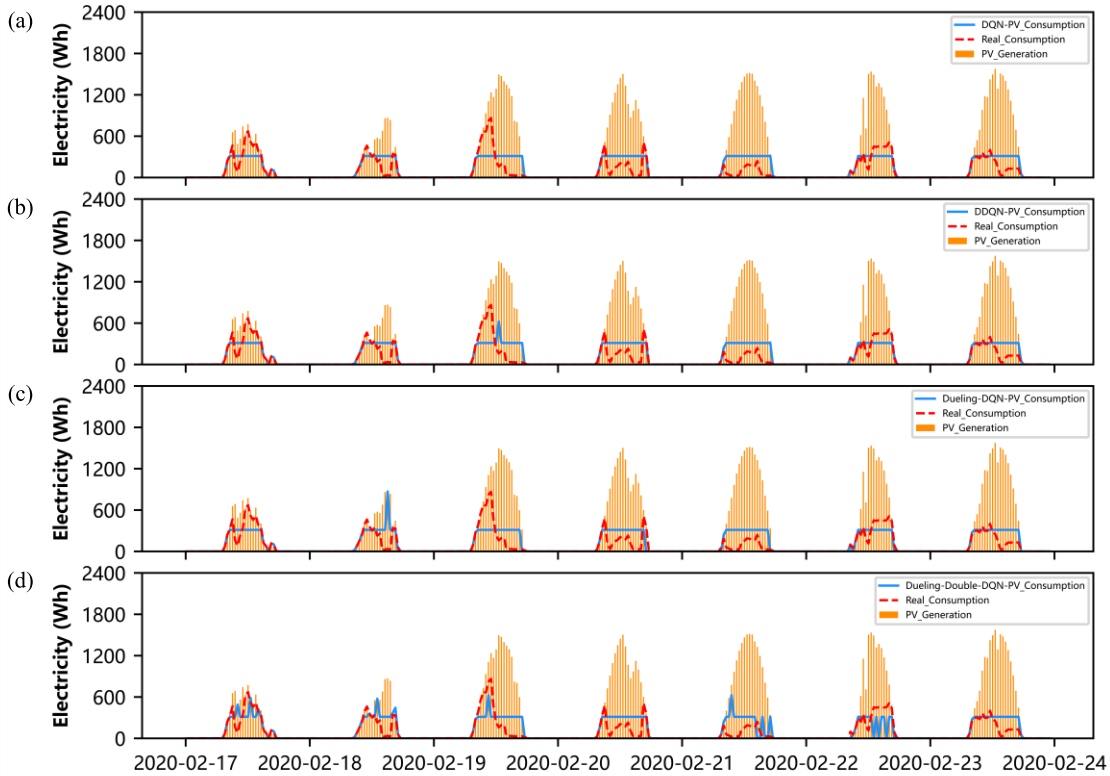

Fig. 7. Comparisons of simulated and measured PV consumption profiles during a cold week

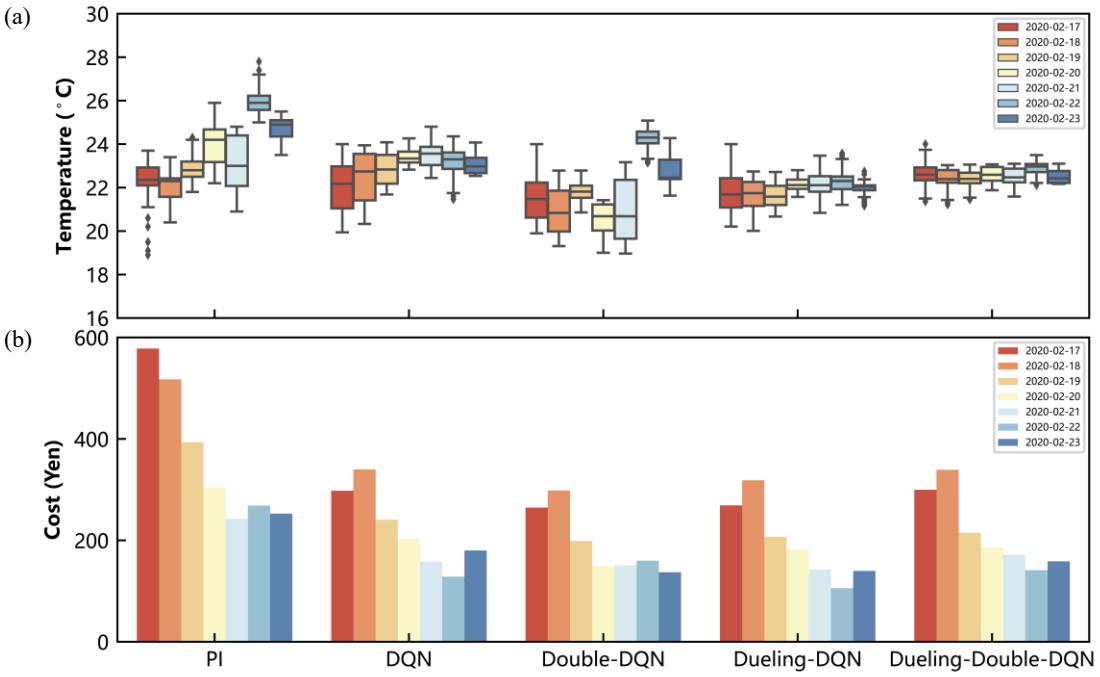

Fig. 8. Comparison of daily indoor temperature variations and electricity costs of different RL algorithms

A detailed comparisons of daily indoor temperature boxplot under different control scenarios were described in Figure.8(a), different colors indicate different days. It should be noted that variations of indoor temperature profiles are different, the temperature fluctuation under PI control is largest. The D3QN control approach shows the best performance in reducing the variations of the indoor temperature. Figure.8(b) summarized the daily electricity costs of different DRL agents, in comparison with the PI controller, each DRL agent can reduce the daily cost of space heating.

The testing temperature profiles of proposed DRL agents during a warm week are shown in Fig.9, the simulated temperatures mainly increase during daytime, more indoor temperatures of DQN and Double DQN agent control outside the thermal comfort range compared with Fig.6. The D3QN agent still presents a great stability of maintaining indoor temperature in comparison with other agent controls. The time-series profiles of the PV generation and local PV self-consumption under each agent

control are depicted in Fig.10. As demonstrated, the DQN and double DQN agents were chosen to consume more on-site PV generation during daytime by allowing greater variations of the indoor temperature.

For further comparisons and analysis, the boxplot of daily indoor temperature Fig.11(a) and energy cost Fig.11(b) of proposed DRL agents are illustrated. Compared with PI controller, presented DRL agents can effectively reduce costs in relatively cold days. Similar to the results depicted in Fig.8, the D3QN shows a great stability in maintaining the room temperature within the thermal comfort range.

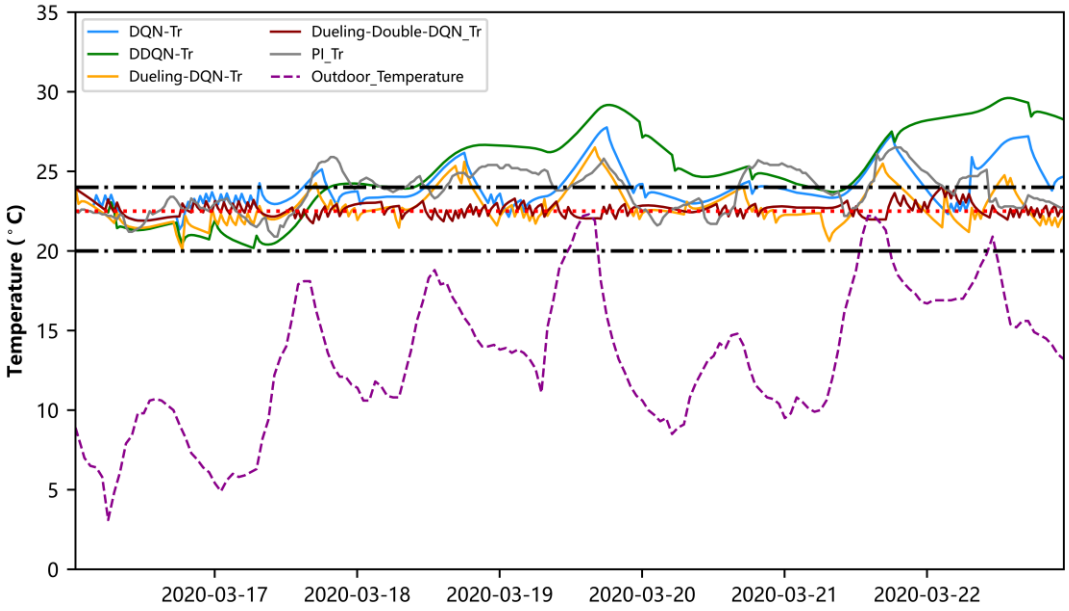

Fig. 9. Testing results of indoor temperature profiles under investigated control strategies during a warm week

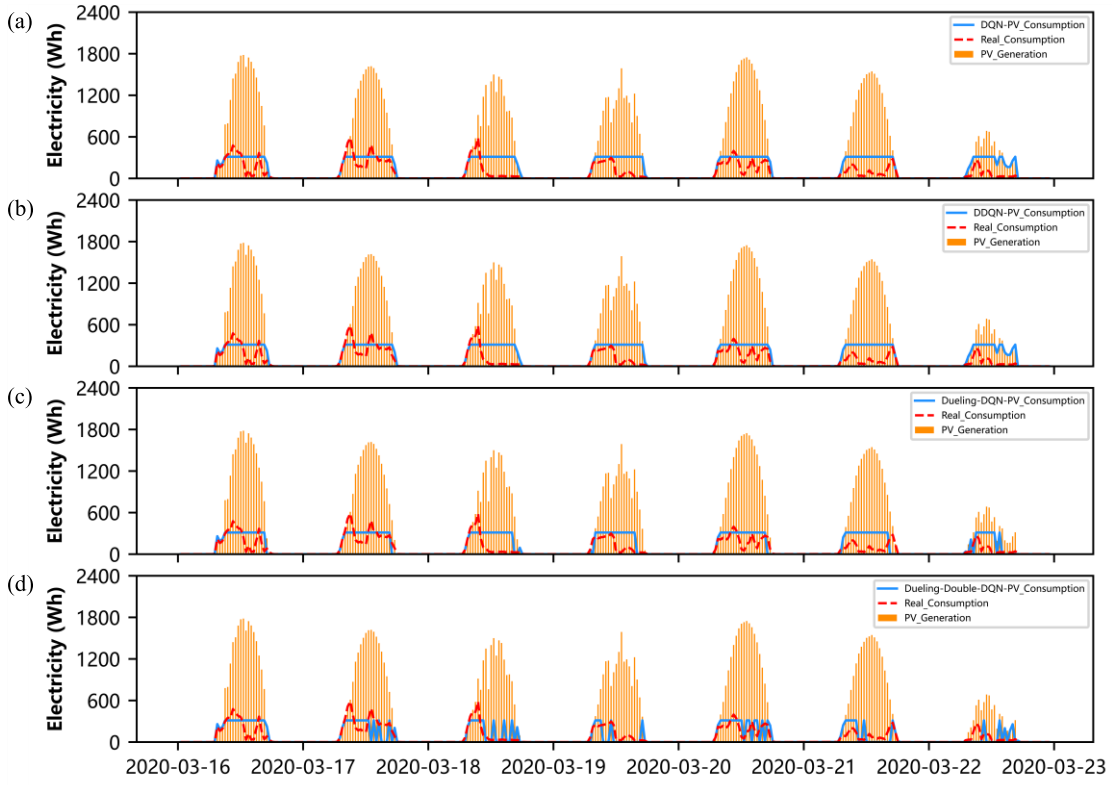

Fig. 10. Comparison of simulated and measured PV consumption profiles during a warm week

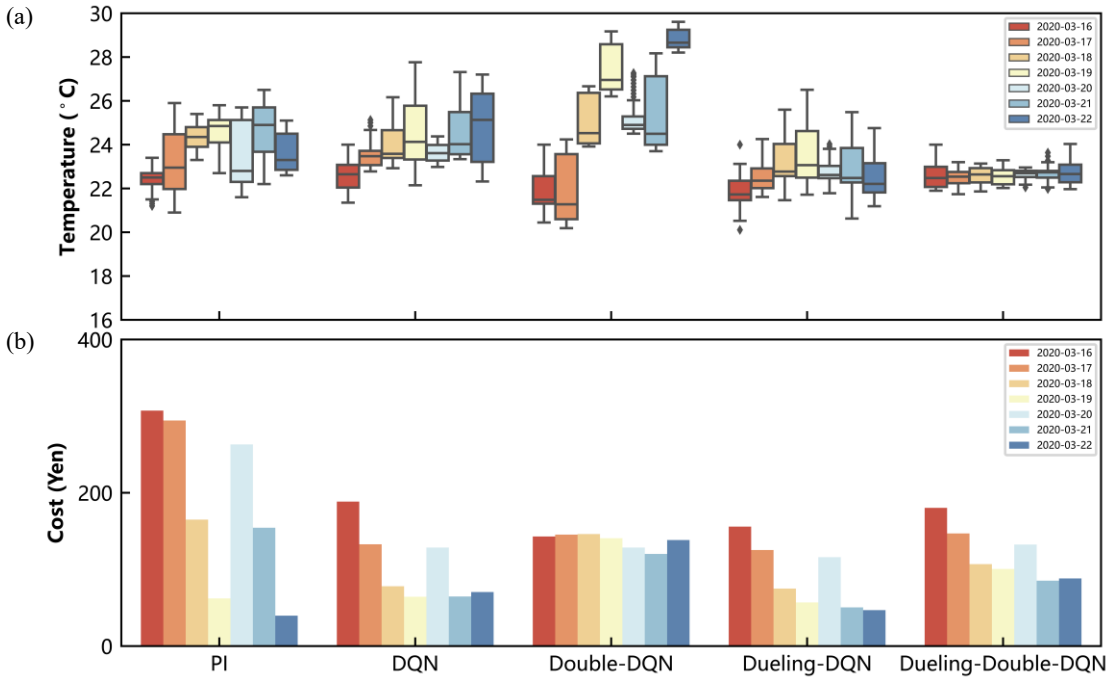

Fig. 11. Comparison of daily indoor temperature variations and electricity costs of different control algorithms

Table 4. Comparisons of electricity cost and local PV consumption ratio between a baseline and proposed DRL agents

Periods	Variables	DQN	Dueling-DQN	Double-DQN	D3QN
Cold week	Electricity cost	-39.46%	-46.62%	-46.85%	-40.92%
	On-site PV consumption	+25.07%	+22.58%	+25.89%	+22.80%
Warm week	Electricity cost	-43.43%	-51.28%	-25.15%	-34.61%
	On-site PV consumption	+73.47%	+49.35%	+73.45%	0.0%

Table 5 Comparisons of indoor temperature between a baseline and proposed DRL agents

Periods	Variable	PI	DQN	Dueling-DQN	Double-DQN	D3QN
Cold week	Mean-T_m (℃)	+1.09	+0.43	+0.56	+0.68	+0.06
	Max-T_m (℃)	+5.30	+2.30	+1.50	+2.58	+1.50
	Min-T_m(℃)	-3.60	-2.56	-2.49	-3.53	-1.28
Warm week	Mean-T_m (℃)	+1.25	+1.48	+0.28	+2.61	+0.11
	Max-T_m (℃)	+4.00	+5.26	+4.01	+7.11	+1.53
	Min-T_m(℃)	-1.60	-1.15	-2.39	-2.31	-0.76

Table 4 summarizes the changes in weekly energy costs and on-site PV consumption ratios under proposed DRL agents, measurement results were used as a benchmark (under PI controller). Overall, testing results demonstrate that designed DRL agents reach good performances in cost saving, this could be a result of increasing local PV consumption ratios and optimal adjustment of indoor temperature regarding the dynamic price information. As observed, the cost saving ratio of Dueling-DQN is over 46%, and percentages of increased local PV consumption were over 22% in the cold week. We note that the indoor temperature maintained by the D3QN agent is more stable and closer to the desired temperature, while the local PV consumption ratios remain unchanged in the warm week.

Table 5 compares the simulated thermal comfort violations of proposed DRL agents

using the difference between the average indoor temperature and user-defined T_m, the deviation between maximum indoor temperature and T_m, and the difference between minimum indoor temperature and T_m. As reported, D3QN agent performs better than other agents in maintaining the stability of thermal comfort range As described in Fig.9, the temperature values under DQN and Double DQN agents control fluctuate more for increasing local PV consumption ratios. To conclude, there is a trade-off between the local PV consumption ratio and indoor temperature variations.

Model-based control for optimizing building energy systems such as MPC has been widely investigated in previous literature. However, the results of the MPC-based controller highly depend on the accuracy of the built model, the computational load is high for long-term optimization. Deep RL algorithms provide a great opportunity to improve building energy system management performance. Although deep RL agents can effectively handle the uncertainty of energy systems and weather conditions, the RL-based approach requires a large amount of training data and lacks interpretability. In this study, combining the physics-based RC network model, the proposed RL framework learns dynamic control policies through interaction with the environment. The simulation results validate the effectiveness of the proposed hybrid RL framework in controlling the house space heating system. Besides, the convergence speed is fast, the training process takes about 3 hours on average. Overall, we see good potential for the hybrid deep RL approach in field of building space heating system management and beyond.

Our results have the potential for improvement. A careful design of the reward functions

is key to finding a good control policy. Even using expert knowledge, defining the reward functions appropriately is essential to learn effective control policies. The selection of invalid parameters can lead to unstable training or ineffective control performance. Therefore, further efforts are needed to test different reward functions (that optimize the same objective). This will help provide a more standardized definition of reward function well adapted for DRL frameworks.

Conclusions and future work

The intermittency of PV generation, the uncertainties of energy consumption and weather conditions increase the complexity of developing a dynamic building energy management strategy. This work presents a hybrid model-based reinforcement learning framework to develop a real-time energy management strategy for a ZEH. The proposed control framework synthesizes a model-based and data-driven approach to improve learning efficiency and model interpretability, which does not require long-term data compared with the model-free RL approach. We formulated the Markov decision process, and then employed RL agents to optimize room heating cost, indoor thermal comfort and local PV self-consumption under different conditions. The main conclusions are drawn as follows:

The proposed RL agents were trained and tested using less monitored data. The results verified the effectiveness of proposed RL agents in reducing space heating costs with significant training efficiency, a long-term training data is not required.

Regarding building thermal flexibility, the proposed RL agents can learn the dynamic

of the system and achieve the optimized objective, optimally adjust real-time heating energy consumption and local PV self-consumption, while maintaining indoor thermal comfort with many uncertainties.

Detailed performances of trained DQN controllers including DQN, DDQN, Dueling-DQN and D3QN in terms of cost saving, indoor thermal comfort and on-site PV utilization are illustrated. The cost-saving benefits of proposed agents are obvious, the changes in on-site PV consumption ratio highly depend on the outdoor conditions.

Regarding indoor thermal comfort, the D3QN learning agent achieved better control policies under the test weeks, value of regulated indoor temperature is close to the expected indoor temperature. Compared with measurement results, the D3QN-based controller achieves 40.9% cost savings and 22.8% rise in consumed PV generation in the test cold week.

A multi-agent deep reinforcement learning control framework can be proposed to transform single agent control into multi-agent control in future work by utilizing more controllable electrical equipment in the integrated building energy systems. The proposed control framework can be applied to the complex energy system with high shares of renewable energy, providing new ideas for decarbonizing building energy systems and optimizing building operational performances.

Acknowledgments

This study was supported by Shandong Natural Science Foundation ‘Research on Flexible District Integrated Energy System under High Penetration Level of Renewable Energy’, grant number ZR2021QE084 and the ‘Development of Smart Building

Management Technologies Towards Carbon Neutrality', grant number XJ20220028.

References

[1] Wei Y-M, Chen K, Kang J-N, Chen W, Wang X-Y, Zhang X. Policy and management of carbon peaking and carbon neutrality: A literature review. Engineering. 2022.
[2] Tan X-C, Wang Y, Gu B-H, Kong L-S, Zeng A. Research on the national climate governance system toward carbon neutrality—A critical literature review. Fundamental Research. 2022;2:384-91.
[3] Clarke J, Searle J. Active Building demonstrators for a low-carbon future. Nature Energy. 2021;6:1087-9.
[4] Svetozarevic B, Begle M, Jayathissa P, Caranovic S, Shepherd RF, Nagy Z, et al. Dynamic photovoltaic building envelopes for adaptive energy and comfort management. Nature Energy. 2019;4:671-82.
[5] Zhou N, Khanna N, Feng W, Ke J, Levine M. Scenarios of energy efficiency and CO2 emissions reduction potential in the buildings sector in China to year 2050. Nature Energy. 2018;3:978-84.
[6] Ohta H. Japan's Policy on Net Carbon Neutrality by 2050. East Asian Policy. 2021;13:19-32.
[7] Li Y, Gao W, Zhang X, Ruan Y, Ushifusa Y, Hiroatsu F. Techno-economic performance analysis of zero energy house applications with home energy management system in Japan. Energy and Buildings. 2020;214.
[8] Zhang X, Gao W, Li Y, Wang Z, Ushifusa Y, Ruan Y. Operational performance and load flexibility analysis of Japanese zero energy house. International Journal of Environmental Research and Public Health. 2021;18:6782.
[9] Kuwahara R, Kim H, Sato H. Evaluation of Zero-Energy Building and Use of Renewable Energy in Renovated Buildings: A Case Study in Japan. Buildings. 2022;12:561.
[10] Li Y, Zhang X, Gao W, Qiao J. Lessons Learnt From the Residential Zero Carbon District Demonstration Project, Governance Practice, Customer Response, and Zero-energy House Operation in Japan. Frontiers in Energy Research. 2022;10.
[11] Xue X, Wang S, Sun Y, Xiao F. An interactive building power demand management strategy for facilitating smart grid optimization. Applied Energy. 2014;116:297-310.
[12] Khorasany M, Shokri Gazafroudi A, Razzaghi R, Morstyn T, Shafie-khah M. A framework for participation of prosumers in peer-to-peer energy trading and flexibility markets. Applied Energy. 2022;314.
[13] Nguyen H-T, Safder U, Loy-Benitez J, Yoo C. Optimal demand side management scheduling-based bidirectional regulation of energy distribution network for multi-residential demand response with self-produced renewable energy. Applied Energy. 2022;322.
[14] Farrokhifar M, Bahmani H, Faridpak B, Safari A, Pozo D, Aiello M. Model predictive control for demand side management in buildings: A survey. Sustainable Cities and Society. 2021;75.
[15] Shi H, Chen Q. Building energy management decision-making in the real world: A comparative study of HVAC cooling strategies. Journal of Building Engineering. 2021;33.
[16] Fu Y, O'Neill Z, Adetola V. A flexible and generic functional mock-up unit based threat injection framework for grid-interactive efficient buildings: A case study in Modelica. Energy and

Buildings. 2021;250.
[17] Ding Y, Lyu Y, Lu S, Wang R. Load shifting potential assessment of building thermal storage performance for building design. Energy. 2022;243.
[18] Zhang X, Gao W, Li Y, Wang Z, Ushifusa Y, Ruan Y. Operational Performance and Load Flexibility Analysis of Japanese Zero Energy House. Int J Environ Res Public Health. 2021;18.
[19] Jensen SØ, Marszal-Pomianowska A, Lollini R, Pasut W, Knotzer A, Engelmann P, et al. IEA EBC Annex 67 Energy Flexible Buildings. Energy and Buildings. 2017;155:25-34.
[20] Oliveira Panão MJN, Mateus NM, Carrilho da Graça G. Measured and modeled performance of internal mass as a thermal energy battery for energy flexible residential buildings. Applied Energy. 2019;239:252-67.
[21] Sánchez Ramos J, Pavón Moreno M, Guerrero Delgado M, Álvarez Domínguez S, F. Cabeza L. Potential of energy flexible buildings: Evaluation of DSM strategies using building thermal mass. Energy and Buildings. 2019;203.
[22] Kim D, Braun JE. MPC solution for optimal load shifting for buildings with ON/OFF staged packaged units: Experimental demonstration, and lessons learned. Energy and Buildings. 2022;266.
[23] Niu J, Tian Z, Lu Y, Zhao H. Flexible dispatch of a building energy system using building thermal storage and battery energy storage. Applied Energy. 2019;243:274-87.
[24] Saavedra A, Negrete-Pincetic M, Rodríguez R, Salgado M, Lorca Á. Flexible load management using flexibility bands. Applied Energy. 2022;317.
[25] Wei Z, Calautit J. Investigation of the effect of the envelope on building thermal storage performance under model predictive control by dynamic pricing. Smart Energy. 2022;6.
[26] Korkas CD, Terzopoulos M, Tsaknakis C, Kosmatopoulos EB. Nearly optimal demand side management for energy, thermal, EV and storage loads: An Approximate Dynamic Programming approach for smarter buildings. Energy and Buildings. 2022;255.
[27] Afroz Z, Shafiullah GM, Urmee T, Shoeb MA, Higgins G. Predictive modelling and optimization of HVAC systems using neural network and particle swarm optimization algorithm. Building and Environment. 2022;209.
[28] Esrafilian-Najafabadi M, Haghighat F. Impact of occupancy prediction models on building HVAC control system performance: Application of machine learning techniques. Energy and Buildings. 2022;257.
[29] Afroz Z, Shafiullah GM, Urmee T, Higgins G. Modeling techniques used in building HVAC control systems: A review. Renewable and Sustainable Energy Reviews. 2018;83:64-84.
[30] Hu C, Xu R, Meng X. A systemic review to improve the intermittent operation efficiency of air-conditioning and heating system. Journal of Building Engineering. 2022;60:105136.
[31] Ulpiani G, Borgognoni M, Romagnoli A, Di Perna C. Comparing the performance of on/off, PID and fuzzy controllers applied to the heating system of an energy-efficient building. Energy and Buildings. 2016;116:1-17.
[32] Killian M, Kozek M. Ten questions concerning model predictive control for energy efficient buildings. Building and Environment. 2016;105:403-12.
[33] Yao Y, Shekhar DK. State of the art review on model predictive control (MPC) in Heating Ventilation and Air-conditioning (HVAC) field. Building and Environment. 2021;200.
[34] Lee D, Ooka R, Ikeda S, Choi W, Kwak Y. Model predictive control of building energy systems with thermal energy storage in response to occupancy variations and time-variant electricity prices. Energy and Buildings. 2020;225.

[35] Dmitrewski A, Molina-Solana M, Arcucci R. CntrlDA: A building energy management control system with real-time adjustments. Application to indoor temperature. Building and Environment. 2022;215.
[36] Li Y, O'Neill Z, Zhang L, Chen J, Im P, DeGraw J. Grey-box modeling and application for building energy simulations - A critical review. Renewable and Sustainable Energy Reviews. 2021;146.
[37] Zhang S, Hu X, Xie S, Song Z, Hu L, Hou C. Adaptively coordinated optimization of battery aging and energy management in plug-in hybrid electric buses. Applied Energy. 2019;256.
[38] Sutton RS, Barto AG. Reinforcement learning: An introduction: MIT press; 2018.
[39] Perera ATD, Kamalaruban P. Applications of reinforcement learning in energy systems. Renewable and Sustainable Energy Reviews. 2021;137.
[40] Wang Z, Hong T. Reinforcement learning for building controls: The opportunities and challenges. Applied Energy. 2020;269.
[41] Touzani S, Prakash AK, Wang Z, Agarwal S, Pritoni M, Kiran M, et al. Controlling distributed energy resources via deep reinforcement learning for load flexibility and energy efficiency. Applied Energy. 2021;304.
[42] Zheng C, Li W, Li W, Xu K, Peng L, Cha SW. A Deep Reinforcement Learning-Based Energy Management Strategy for Fuel Cell Hybrid Buses. International Journal of Precision Engineering and Manufacturing-Green Technology. 2021;9:885-97.
[43] An Y, Niu Z, Chen C. Smart control of window and air cleaner for mitigating indoor PM2.5 with reduced energy consumption based on deep reinforcement learning. Building and Environment. 2022;224.
[44] Han M, May R, Zhang X, Wang X, Pan S, Yan D, et al. A review of reinforcement learning methodologies for controlling occupant comfort in buildings. Sustainable Cities and Society. 2019;51.
[45] Pinosky A, Abraham I, Broad A, Argall B, Murphey TD. Hybrid control for combining model-based and model-free reinforcement learning. The International Journal of Robotics Research. 2022.
[46] Swazinna P, Udluft S, Hein D, Runkler T. Comparing model-free and model-based algorithms for offline reinforcement learning. arXiv preprint arXiv:220105433. 2022.
[47] Plaat A. Model-Based Reinforcement Learning. In: Plaat A., editor. Deep Reinforcement Learning. Singapore: Springer Nature Singapore; 2022. p. 135-67.
[48] Lee H, Kim K, Kim N, Cha SW. Energy efficient speed planning of electric vehicles for car-following scenario using model-based reinforcement learning. Applied Energy. 2022;313.
[49] Wang D, Shen Y, Wan J, Sha Q, Li G, Chen G, et al. Sliding mode heading control for AUV based on continuous hybrid model-free and model-based reinforcement learning. Applied Ocean Research. 2022;118.
[50] Totaro S, Boukas I, Jonsson A, Cornélusse B. Lifelong control of off-grid microgrid with model-based reinforcement learning. Energy. 2021;232.
[51] Zhang W, Wang J, Xu Z, Shen Y, Gao G. A generalized energy management framework for hybrid construction vehicles via model-based reinforcement learning. Energy. 2022;260.
[52] Yu L, Xu Z, Zhang T, Guan X, Yue D. Energy-efficient personalized thermal comfort control in office buildings based on multi-agent deep reinforcement learning. Building and Environment. 2022;223.

[53] Valladares W, Galindo M, Gutiérrez J, Wu W-C, Liao K-K, Liao J-C, et al. Energy optimization associated with thermal comfort and indoor air control via a deep reinforcement learning algorithm. Building and Environment. 2019;155:105-17.
[54] Watkins CJ, Dayan PJMl. Q-learning. 1992;8:279-92.
[55] Tangkaratt V, Mori S, Zhao T, Morimoto J, Sugiyama MJNn. Model-based policy gradients with parameter-based exploration by least-squares conditional density estimation. 2014;57:128-40.
[56] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing atari with deep reinforcement learning. 2013.
[57] 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.
[58] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning. Proceedings of the AAAI conference on artificial intelligence2016.
[59] Wang Z, Schaul T, Hessel M, Hasselt H, Lanctot M, Freitas N. Dueling network architectures for deep reinforcement learning. International conference on machine learning: PMLR; 2016. p. 1995-2003.
[60] Lopez-Martinez D, Eschenfeldt P, Ostvar S, Ingram M, Hur C, Picard R. Deep reinforcement learning for optimal critical care pain management with morphine using dueling double-deep q networks. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC): IEEE; 2019. p. 3960-3.
[61] Moriyama T, De Magistris G, Tatsubori M, Pham T-H, Munawar A, Tachibana R. Reinforcement Learning Testbed for Power-Consumption Optimization. Singapore: Springer Singapore; 2018. p. 45-59.
[62] Hu M, Xiao F, Jørgensen JB, Wang S. Frequency control of air conditioners in response to real-time dynamic electricity prices in smart grids. Applied Energy. 2019;242:92-106.