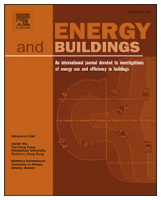
Energy & Buildings 303 (2024) 113770

Contents lists available at ScienceDirect

Energy & Buildings

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



Enhancing HVAC energy management through multi-zone occupant-centric approach: A multi-agent deep reinforcement learning solution

Xuebo Liu [a], Yingying Wu [b], Hongyu Wu [a,*]

[a] Mike Wiegers Department of Electrical and Computer Engineering, Kansas State University, Manhattan, KS 66506, USA
[b] Department of Interior Design and Fashion Studies, Kansas State University, Manhattan, KS 66506, USA

A B S T R A C T

Occupant-centric HVAC control places a premium on factors including thermal comfort and electricity cost to guarantee occupant satisfaction. Traditional approaches, reliant on static models for occupant behaviors, fall short in capturing intra-day behavioral variations, resulting in imprecise thermal comfort evaluations and suboptimal HVAC energy management, especially in multi-zone systems with diverse occupant profiles. To address this issue, this paper proposes a novel occupant-centric multi-zone HVAC control approach that intelligently schedules cooling and heating setpoints using Multi-agent Deep Reinforcement Learning (MADRL). This approach systematically takes into account stochastic occupant behavior models, such as dynamic clothing insulation adjustments, metabolic rates, and occupancy patterns. Simulation results demonstrate the efficacy of the proposed approach. Comparative case studies show that the proposed MADRL-based, occupant-centric HVAC control reduces electricity costs by 51.09% compared to rule-based approaches and 4.34% compared to single-agent DRL while maintaining multi-zonal thermal comfort for occupants.

1. Introduction

Buildings in the United States account for approximately 36% of total energy consumption, with HVAC systems being a significant contributor, particularly during hot summers [1]. Smart homes using IoT technologies and human-centered intelligent scheduling for HVAC control is a promising solution for energy-efficient and comfortable buildings [2]. Understanding and incorporating Occupant-Centric Control (OCC) is crucial for effective building energy management [3,4]. Wang et al. [5] investigated occupancy patterns in single-person offices and proposed a probabilistic model for occupancy prediction, emphasizing the complexity and time variation of occupancy intervals. Several studies have proposed algorithms and models for predicting occupancy to improve energy consumption and occupant comfort, including Reinhart [6], Page et al. [7], Klein et al. [8], and Fabi et al. [9]. Furthermore, integrating clothing behaviors and clothing decisions into HVAC control strategies can further enhance building energy efficacy [10–12]. There are a lot of literature studies on the simplified model for occupant behavior model in building energy management to reduce the electricity cost while maximizing the occupant thermal comfort in residential buildings [13,14] and commercial buildings [15–17]. However, occupancy and clothing behavior are not the only factors that have an effect on occupants' thermal comfort model. Metabolic rate with activity schedule is another important factor that should be conducted in the thermal comfort [18] considered in indoor temperature control. However, dynamic occupant behavior containing the occupancy, clothing adjustment, and metabolic rate involved in thermal comfort in HVAC control is still an unsolved problem [19].

On the other hand, model-based optimization can be time-consuming when dealing with large solution spaces, making it unsuitable for real-time decision-making. In recent years, Deep Reinforcement Learning (DRL), a model-free approach, has gained traction among engineers and researchers for tackling building energy management problems [20]. DRL-based HVAC control methods have been proposed to address challenges posed by large state-action spaces [21,22] and complex indoor environments [23,24]. However, addressing the multi-zone HVAC control problem with continuous action spaces remains a challenge despite the application of DRL techniques in previous studies. In real-world environments, multi-zone thermal control involves complex control agents and often necessitates a balance between competition and cooperation among these agents. Some studies have applied Multi-agent Deep Reinforcement Learning (MADRL) to multi-zone thermal control [25,26], while often lacking dynamic modeling. In Table 1, "Const." signifies a consistent schedule applied throughout simulations, while "Dyn." represents changing occupant activities influenced by factors such as time of day, weather, and personal choices. Although certain studies have integrated occupant presence [27–29], they do not account for metabolic rate and clothing adjustments. This gap in occupant behavior modeling

* Corresponding author.
E-mail address: hongyuwu@ksu.edu (H. Wu).

https://doi.org/10.1016/j.enbuild.2023.113770
Received 31 July 2023; Received in revised form 3 October 2023; Accepted 17 November 2023
Available online 27 November 2023

Table 1
Literature review of DRL in HVAC control.

Reference (by time)	Whole building energy simulation	Multi-area thermal zone	Multi-agent approach	Occupant behavior		
				Occupancy	Metabolic	Clothing
[21]/2017	√	√				
[30,24]/2018-19	√	√		Const.		
[31]/2020	√	√		Const.	Const.	
[32]/2021	√	√	√		Const.	Const.
[33]/2021	√	√		Const.		
[34]/2021	√	√				
[25]/2021	√	√		Const.		
[27]/2021	√	√	√	Dyn.		
[35]/2022	√	√		Const.	Const.	Const.
[28]/2022	√	√	√	Dyn.		
[26]/2022	√	√	√			
[29]/2022	√	√	√	Dyn.		
[36]/2022	√			Dyn.		
[37]/2023	√	√	√			
[38]/2023		√	√			
This work	√	√	√	Dyn.	Dyn.	Dyn.

highlights the need for novel MADRL methods capable of addressing the complexity of multi-zone environments and dynamic occupant behaviors in HVAC control.

This paper bridges this gap by proposing a multi-zone HVAC energy management scheme that aims to minimize the electricity cost and the occupants' thermal discomfort using a MADRL approach. The main contributions of this work are three-fold:

1) This paper proposes a pioneering multi-zone HVAC energy management scheme that is the first of its kind to explicitly consider the occupants' behaviors, including occupant presence, clothing conditions, and activity conditions, for minimizing the electricity cost and the occupant's thermal discomfort.
2) An MADRL approach is developed for making sequential HVAC setpoint decisions while considering the continuous action space under a whole-building simulation environment with stochastic occupant behavior. Specifically, the MADRL intelligently schedules the cooling and heating set points for the multi-zone office buildings while accounting for dynamic occupant behaviors.
3) The proposed model is trained and simulated by EnergyPlus in a practical multi-zone building with real-world datasets at daily and yearly timescales. Simulation results show the electricity cost saving of MADRL is, respectively, 4.34% and 51.09% compared to single-agent DRL and rule-based control while maintaining a high comfort level for multi-zone occupants.

For the rest of the paper, the mathematical formulation and proposed methodology are presented in Section 2. The simulation results of comparative case studies are in Section 3. Section 4 contains a discussion of this study. Section 5 presents the conclusions of the paper.

2. Problem formulation and methodology

2.1. Overview of approach

Reinforcement learning is a paradigm within machine learning wherein an autonomous agent endeavors to acquire an optimal strategy for selecting a sequence of actions within an environment to maximize its cumulative reward. The agent's decision-making process hinges on the feedback it receives in the form of a reward value following each executed action. These decisions are contingent upon the agent's interpretation of the environment, which is encapsulated by a state representation. This iterative process persists as the agent engages with the environment, with the aim of progressively improving its policy, ultimately striving to attain predefined objectives. Fig. 1 provides a visual representation of the research framework, showcasing the seamless integration of the MADRL algorithm within a complex five-zone office building environment. This environment, as shown in the middle module of Fig. 1, is accurately simulated using the Building Controls Virtual Test Bed (BCVTB) [39] in conjunction with EnergyPlus, facilitating realistic building dynamics and HVAC power consumption modeling. Within this framework, the MADRL agent (top module of Fig. 1) actively engages with the environment, effectively making decisions (heating and cooling setpoints of HVAC) pertaining to HVAC control to simultaneously optimize energy efficiency and occupant comfort. A distinctive feature highlighted in the figure is the incorporation of dynamic occupant behavior models, as shown in the bottom module of Fig. 1, represented through a stochastic process. These models dynamically account for variations in factors such as clothing insulation, occupant presence, and metabolic rates over time. Furthermore, it's essential to note the bidirectional interaction, where reward values are transmitted from EnergyPlus to the MADRL module, and optimal actions are passed from MADRL to EnergyPlus during simulation, effectively capturing occupant behavior. This integration stands at the core of the research's mission to achieve efficient multi-zone HVAC control while upholding occupant thermal comfort standards within the building.

2.2. Occupant behavior modeling

Building energy management has undergone a significant transformation with the integration of technology, sustainability considerations, and occupant comfort. Occupant-centric control (OCC) has emerged as a prominent concept, shifting the focus from a building-centric to an occupant-focused approach [40–42]. Researchers now recognize the importance of addressing occupants' needs and optimizing building systems for their comfort and energy efficiency. In the context of ensuring thermal comfort, the Predicted Mean Vote (PMV)/Predicted Percentage of Dissatisfied (PPD) index, introduced by P.O. Fanger and colleagues, is employed to measure occupants' comfort levels based on their environmental conditions [43]. This index is included in ISO 7730 (2005) and the ASHRAE standard 55 (2004) [44,45]. PMV/PPD models,

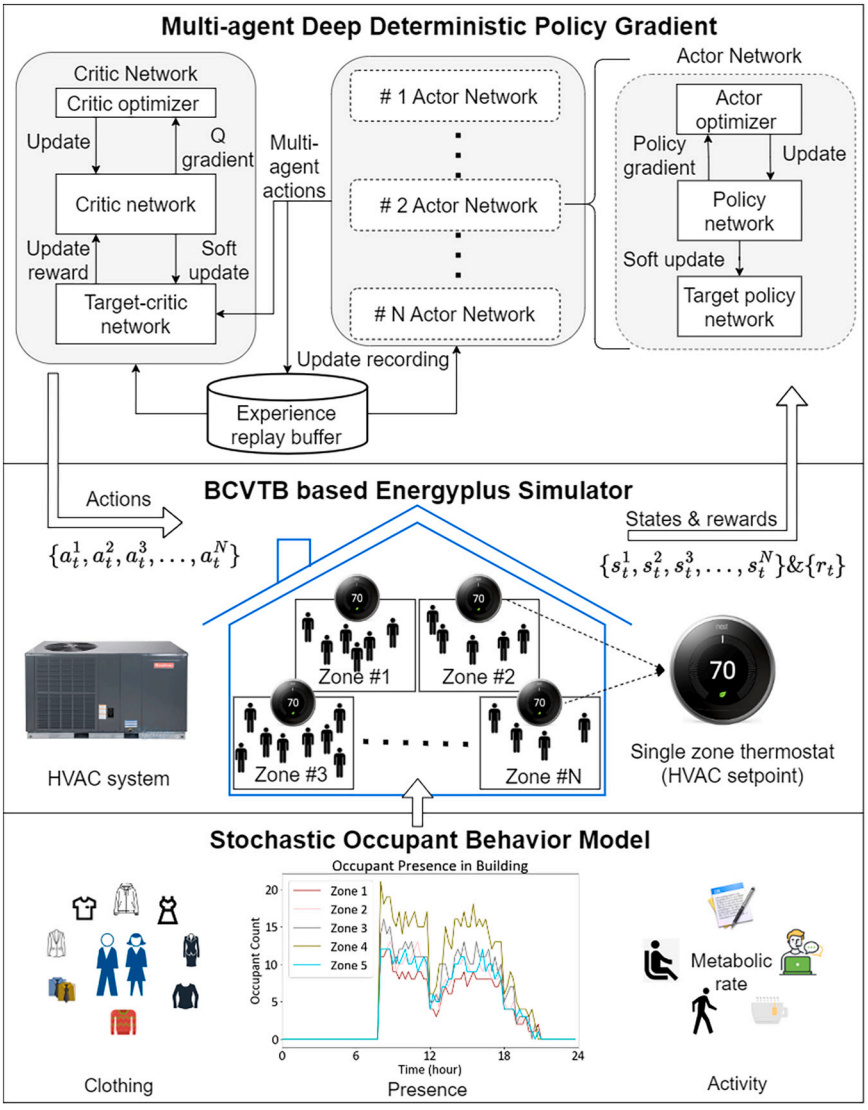


Fig. 1. Illustration of proposed MADRL in multi-zone HVAC control. Three modules are contained in this approach, 1) a MADRL algorithm: Multi-agent Deep Deterministic Policy Gradient; 2) A testbed: EnergyPlus simulator; 3) Stochastic occupant behavior model includes clothing, occupant presence, and metabolic rate. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

which take into account the indoor environment and occupant behavior factors as input, have been integrated into OCC as a part objective of the HVAC energy management scheme.

The proposed OCC allows occupants to have control over settings such as cooling and heating points in HVAC systems, enhancing their satisfaction and productivity. However, one of the critical factors in implementing OCC is understanding and integrating occupant behavior [46]. Factors such as occupant presence, activity levels, and clothing behavior are significant in optimizing thermal comfort strategies. Prior literature has often overlooked that behavior in building energy management, resulting in an incomplete understanding of occupant needs and potentially leading to inefficient control strategies. Our work contributes to addressing these challenges and developing an approach that considers and integrates these aspects into occupant-centric building energy management systems. Note that during specific time periods, such as morning opening (8 am), lunchtime (noon to 1 pm), and afternoon closing (6 pm to 7 pm), occupants are assumed to be engaged in walking activities, resulting in elevated activity levels. In the simulation, we assumed that occupants within the same zone share similar behaviors in terms of activity level and clothing insulation. This is reasonable in a multi-zone commercial building environment as the behaviors of one can easily affect others within the same zone. This assumption allowed us to group occupants within each zone and apply general patterns of behavior, which is a common approach in HVAC simulations. Initially, we established a fixed schedule and subsequently introduced a stochastic model to this schedule to simulate diverse behaviors across different zones.

2.2.1. *Occupant presence*

Occupant presence information plays a pivotal role in modern building management systems, offering invaluable insights into space utilization and opportunities for optimizing energy consumption. As illustrated in Table 1, occupant presence is a prevalent factor in HVAC control as occupant behavior. In our study, we initially implement a fixed schedule for occupant presence and subsequently introduce stochastic variables to simulate occupancy patterns across the five zones, as depicted in Fig. 2 (top). While future research avenues may explore data-driven or Markov chain models to predict time-dependent occupant behavior within simulations [47–49], our current approach serves as an initial exploration of how dynamic occupant behavior influences HVAC control strategies.

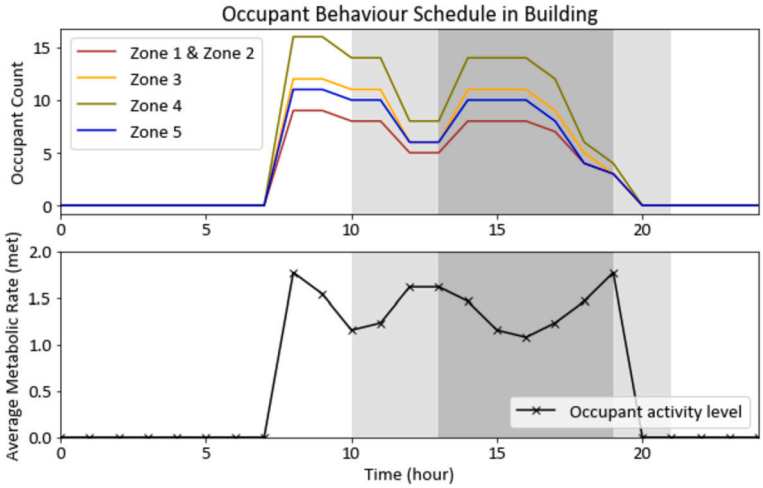


Fig. 2. One-day average occupant presence schedule (top) and activity level schedule (bottom) for each zone.

Table 2
Activity level and metabolic rate.

Office Activities	Activity Level (W/m^2)	Met
Reading, seated	55	1
Seated, quiet	60	1
Writing	60	1
Typing	65	1.1
Standing, relaxed	70	1.2
Filing, seated	70	1.2
Filing, standing	80	1.4
Walking (0.9 m/s)	115	2

2.2.2. Metabolic rate

Efficient energy management in office or residential buildings necessitates understanding occupants' metabolic rates and activity levels, which directly impact energy expenditure [18,50]. Metabolic rate, derived from oxygen consumption, indicates an individual's energy expenditure during physical activity. Activity levels, measured in watts per square meter (W/m^2), determine the heat generated by occupants. Typical office activities range from 55 W/m^2 to 115 W/m^2.

To optimize energy usage, it is essential to consider these variations in activity levels. Table 2 provides an overview of different behavioral conditions based on metabolic rates (in units: W/m^2 and Met). Tailoring energy management strategies to accommodate these diverse activity levels enables the implementation of sustainable practices in office buildings. Fig. 2 (bottom) illustrates the average activity level in the office building. The assumption is that the schedule of this behavior is average in all thermal zones and during morning opening (8 am), lunchtime (noon to 1 pm), and afternoon closing (6 pm to 7 pm), occupants are assumed to have a high activity level. By introducing random variables to the activity level, we can introduce greater diversity in this behavior, capturing more realistic occupant activity patterns.

2.2.3. Clothing behaviors

Recognizing the influence of clothing on the discomfort function (PPD in the reward function), it becomes vital to consider occupants' clothing behaviors. To achieve this, the research adopts a dynamic clothing behavior model developed by Schiavon and Lee [51], as represented in Equation (1). Within this model, o^{clo} denotes the occupant's clothing insulation, while $t_{a(out,6)}$ represents the outdoor temperature at 6 am. Schiavon and Lee's study employed multivariable linear mixed models, with the first model accounting for outdoor air temperature and the second incorporating indoor operative temperature. These models successfully explained the total variance in clothing behavior, enhancing the realism of thermal comfort assessments in office buildings. Moreover, the inclusion of stochastic elements accommodates the inherent variability in clothing behavior, making the evaluation of thermal comfort in the office building more authentic and reflective of the diverse clothing choices and individual preferences of occupants.

$$o^{clo} = \begin{cases} 1, & t_{a(out,6)} < -5^oC \\ 0.818 - 0.0364 * t_{a(out,6)}, & -5^oC \le t_{a(out,6)} < 5^oC \\ 10^{-0.1635-0.0066+t_{a(out,6)}}, & 5^oC \le t_{a(out,6)} < 26^oC \\ 0.46, & t_{a(out,6)} \ge 26^oC \end{cases} \quad (1)$$

2.3. Stochasticity in occupant behavior model

Stochastic modeling is a mathematical approach used to analyze systems involving randomness and uncertainty which employs probability theory to describe the likelihood of various outcomes, often using simulations to estimate complex systems. In addition, stochastic modeling is essential to make probabilistic predictions and assess risks in systems influenced by chance events and variability [51]. Previous studies have extensively explored appliance scheduling problems using various stochastic models, such as forecast errors in hot water usage [52], outdoor temperature [53] and renewable energy generation [54,55]. On the other hand, as shown in Table 1, occupant presence has been one of the most popular areas in recent research. However, these studies have primarily focused on only occupant presence as the main aspect of occupant behaviors, while neglecting the dynamic aspects. In addition, varying clothing behaviors and various metabolic rates are not considered. To address this research gap, here we propose a model with occupants' behaviors influenced by the time of day and weather. In addition, we introduce stochasticity [56,57] into the occupant behavior model to show the effect of variability and range of patterns. Specifically, the following parameters are defined:

$$OB(t) = \{o_t^{pres}, o_t^{clo}, o_t^{met}\}$$
$$X(t) = \{X_t^{pres}, X_t^{clo}, X_t^{met}\}$$

where t is the time index; $OB(t)$ represents the estimated values of occupant behaviors without considering any stochastic effects; $X(t)$ represents Gaussian-distributed random variables at time t for occupant presence, clothing behavior and metabolic rate. Note that $OB(t)$ values are derived from the office schedules, shown in Fig. 2 for occupant presence and metabolic rate based on different activity levels, while the clothing behavior is determined using Equation (1). Note that the

clothing insulation and activity level are measured in the average value through zone because EnergyPlus, as the simulation platform used in this study, has limitations in representing individual differences.

Based on those parameters, we define:

$$\widetilde{OB}_t = OB(t) + X(t) \tag{2}$$

where $\widetilde{OB}_t$ is the stochastic dynamic occupant behavior that provides the ability to capture the inherent time-dependent variability and uncertainty of occupant behaviors. By incorporating random variables, our model can effectively simulate the nature of occupant behavior (presence, clothing, and metabolic rate), making it the first of its kind to integrate this concept into DRL-based building energy management systems.

2.4. States, actions and rewards of reinforcement learning

Three indices are employed: $t = \{1, 2, 3, ..., T\}$ for time slots, $m = \{1, 2, 3, ..., M\}$ for building zones, and $j = \{H, C\}$ to distinguish between the HVAC dual modes of heating (H) and cooling (C). The Reinforcement Learning consists of sets of states and actions, denoted as S and A, respectively, where $s \in S$ and $a \in A$. Specifically, s_t^m represents the state of zone m at time slot t, and $a_t^{m,H}$ and $a_t^{m,C}$ indicate the heating and cooling setpoints for zone m at time slot t. In the same way, $\widetilde{OB}_t^m$ is the stochastic occupant behavior in zone m at time slot t. It is worth mentioning that $\widetilde{OB}_t^m$ is part of s_t^m. The reward function r_t is defined for the entire building at time slot t.

The state s_t includes two parts: external states and internal states. External states related to the building's outside environment encompass the outdoor dry bulb temperature, air relative humidity, wind speed, wind direction, diffuse solar radiation, and direct solar radiation. For each zone m, seven states are considered, comprising the zone air temperature, zone thermal comfort mean radiant temperature, zone air relative humidity, zone thermal comfort clothing value, thermal comfort index (PMV/PPD value), occupant counts, and zone average metabolic rate. Additionally, there are two internal states associated with electricity price and HVAC power consumption. The power consumption in our simulation represents the total electricity consumed by HVAC for the entire building, with the assumption that the entire building receives a single utility bill. As a result, power consumption is treated as a single parameter, and individual HVAC power consumption for each zone is not considered. The policy of the proposed MADDPG is a mapping that takes the current multi-zone environment observation and generates a probability distribution of actions, specifically the heating and cooling setpoints. This approach empowers the agent to make informed decisions based on the current state of the building and its zones, leading to optimized HVAC control strategies that adapt to various environmental and occupant conditions. The set tuple $\{s_t, a_t\}$ is used to represent the states, including external states and internal states, and actions of all zones in the building:

$$\{s_t, a_t\} = \{(s_t^1, ..., s_t^M), (a_t^{1,H}, a_t^{1,C}, ..., a_t^{M,H}, a_t^{M,C})\}$$

A reward function in DRL is a numerical signal that informs an agent about the desirability of its actions in a given state, with higher values indicating favorable actions and lower values representing unfavorable ones. The agent's goal is to learn a policy that maximizes the cumulative reward it receives over time. The reward function for the whole environment is defined as follows:

$$r_t(m) = -d \cdot \sum_{m=1}^{M} PPD(s_t^m) - b \cdot C_t\left(s_t, a_t\right) \tag{3}$$

In (3), the reward function consists of two components: 1) a discomfort function calculated using the PMV/PPD, and 2) an electricity cost function C_t. The weighted coefficients d_t and b_t can be determined based on historical data [51] and represent varying preferences of occupants, including cost-saving or comfort-seeking type. The electricity cost function C_t is given by:

$$C_t\left(s_t^m, a_t^m\right) = \mathrm{c}_t \sum_{m=1}^{M} p_t^m \Delta t \tag{4}$$

where, c_t is the electricity price at time slot t, Δt is the simulation time interval, and p_t^m is the electricity consumption caused by the HVAC system in zone m at time slot t. Importantly, c_t and p_t^m are part of the states that can be observed from the environment.

2.5. Multi-agent deep reinforcement learning

While classical reinforcement learning algorithms like Q-learning and policy gradient have exhibited proficiency in single-agent domains, their application presents unique challenges characterized by evolving policies, non-stationary surroundings, and the imperative for agent collaboration. In response to these multifaceted demands, the machine learning community has introduced DDPG, a prominent member of the reinforcement learning family renowned for its aptitude in handling continuous action spaces [58]. DDPG serves as a pivotal precursor to our exploration of MADDPG, an extension tailored explicitly for multi-agent domains. MADDPG advances the DDPG paradigm by empowering agents to deliberate global states and make informed decisions predicated on the actions of fellow agents, thereby enhancing coordination and overall system performance. The ensuing discourse delves deeper into the nuanced application of DDPG and the pivotal role of MADDPG in addressing the intricacies of multi-agent scenarios [59]. MADDPG is specifically crafted for multi-agent scenarios, where multiple agents interact within the same environment, potentially requiring coordination, communication, collaboration, or competition among agents. MADDPG extends DDPG to accommodate these complex multi-agent dynamics, making it suitable for modeling a wide range of cooperative or competitive interactions among autonomous agents. MADDPG offers several positive aspects in the context of multi-agent reinforcement learning:

1) Cooperative Learning: MADDPG facilitates cooperative learning by allowing agents to share information and learn from each other. By considering the joint actions and observations of all agents, MADDPG promotes coordination and collaboration among the agents, leading to better overall performance.
2) Centralized Learning, Decentralized Execution: MADDPG employs a centralized training approach, where a centralized critic network is used to estimate the Q-values based on the joint actions and observations. However, during execution, each agent acts independently based on its local observations, enabling decentralized decision-making and reducing communication requirements.
3) Handling Non-Stationarity: MADDPG is designed to handle non-stationarity in multi-agent environments, where agents' policies may change during training. By incorporating a centralized critic network that considers all agents' actions and observations, MADDPG can adapt to changing dynamics and maintain stability during training.
4) Policy Exploration and Exploitation: MADDPG combines the benefits of exploration and exploitation by utilizing the DDPG algorithm. DDPG employs an exploration policy, such as adding noise to the actions, to encourage exploration and discover new strategies. At the same time, it leverages the learned policies to exploit the most promising actions and maximize performance.

The proposed MADDPG is formulated using the Bellman equation, enabling the learning of a Q-function and a multi-agent-based policy. Similar to DDPG, a deep neural network (DNN) is utilized as the Value Function Approximation (VFA) in MADDPG. This approach integrates the actor-evaluation approach and multi-agent technique, making it suitable for handling model-free, high-dimensional, and continuous ac-

Algorithm 1 Multi-agent Deep Deterministic Policy Gradient.

Input: Environment with M zones, and occupant behavior OB
Output: M sets of actor networks
Randomly initialize evaluation network $Q(s,a|\theta^Q)$ with weights θ^Q where $a = \{a^1, a^2, ..., a^M\}$, and M actors $\mu(s|\theta^\mu) = \{\mu^1(s|\theta^{\mu,1}), \mu^2(s|\theta^{\mu,2}), ..., \mu^M(s|\theta^{\mu,M})\}$ with weights $\theta^{\mu,m}$, where $m \in \{1, 2, ..., M\}$
Initialize target-evaluation network Q' with weights $\theta^{Q'} \leftarrow \theta^Q$, and M target actors $\mu' = \{\mu'^{,1}, \mu'^{,2}, ..., \mu'^{,M}\}$ with weights $\theta^{\mu',m} \leftarrow \theta^{\mu,m}$ where $m \in \{1, 2, ..., M\}$
Initialize experience replay buffer B
for episode = 1 to E **do**
  Initialize Ornstein–Uhlenbeck process (OU) for action exploration
  Receive initial observation state s_1
  for t = 1 to T **do**
    Select action $a_t^m = \mu(s_t^m|\theta^\mu) + OU_t$ according to the policy network and exploration noise for zone m at t
    Apply stochastic-based dynamic occupant behavior based on Equation (2) to the environment
    Execute action a_t and observe reward r_t and new state s_{t+1}
    Store transition (s_t, a_t, r_t, s_{t+1}) in buffer B
    Sample a random mini-batch of M transitions (s^l, a^l, r^l, s^{l+1}) from B
    Set y^l based on Equation (6), where $\mu'(s^{l+1}|\theta^{\mu'})$ is the sets of M actors
    Update evaluation network by minimizing the loss based on Equation (7), where $a_l = \{a_l^1, a_l^2, ..., a_l^M\}$
    Update the actor policy using the sampled policy gradient for all $\mu(\cdot)$ based on Equation (8)
    Soft update the target evaluation network and all target actor networks:
      $\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$
      $\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$
  end for
end for

tion spaces in multi-zone environments. The Q-Value function, representing the value function, is given by the following expression:

$$Q^\pi(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t(s_t, a_t) + \gamma \max_{a_{t+1} \in A} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \tag{5}$$

$\pi \in \Pi$ represents the policy, which is a set of actions with a probability distribution. α denotes the learning rate, and γ is the discount factor used for future reward considerations. With Equation (5) as the foundation, the VFA of DRL can be formulated as follows:

$$y_l = r_l + \gamma Q(s_{l+1}, \mu(s_{l+1}|\theta^\mu)|\theta^Q) \tag{6}$$

where $l \in L$ denotes the index of mini-batch L, which is sampled from the experience replay buffer B. The parameters θ^μ and θ^Q correspond to the weights of the actor neural network $\mu(\cdot)$ and the evaluation network $Q(\cdot)$, respectively. The evaluation network $Q(\cdot)$ is updated through the minimization of the loss function:

$$Loss = \frac{1}{M} \sum_l (y_l - Q(s_l, a_l|\theta^Q))^2 \tag{7}$$

To facilitate the exploration of the actor-network, the Ornstein-Uhlenbeck process is utilized [60]. The actor networks are updated using the policy gradient technique, which applies the chain rule to compute the gradient of the expected return with respect to the actor parameters, represented by the approximated loss of the distribution J:

$$\nabla_{\theta^\mu} J \approx \frac{1}{M} \sum_l \nabla_a Q(s, a|\theta^Q)|_{s=s_l, a=\mu(s_l)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{s_l} \tag{8}$$

The MADDPG algorithm introduces specific modifications tailored for multi-zone HVAC control:

- During initialization, M actors and target actors are established, each corresponding to a zone within the building.
- Each actor contributes the action for its respective zone during policy execution, resulting in a set of zone-specific actions.
- The stochastic occupant behavior, represented by Equation (2), is applied to the environment (EnergyPlus) with each execution, capturing variability in occupant actions.

Table 3
HVAC parameters.

Feature	Parameters	Setting
General	High-Speed/Low-Speed Sensible Heat Ratio	0.75
	Nominal Capacity (W)	3500
Cooling	Rated Cooling COP (W/W)	3.0
	Internal Static Air Pressure (Pa)	450
Heating	Burner Efficiency	0.98
	Nominal Capacity (W)	3500
	Fan Total Efficiency	0.7
	Pressure Rise (Pa)	600
Fan	Maximum Flow Rate (m^3/s)	3.0
	Power Minimum Flow Fraction	0.25
	Motor Efficiency	0.9
	Motor In Air-stream Fraction	1.0

- The outputs of all M actors and target actors serve as inputs to the target evaluation network.
- All actor networks are updated using the sampled policy gradient, allowing the agents to learn and improve their strategies based on the environment's feedback.
- Instead of returning a single-actor network, the well-trained multiple-actor networks are returned, providing zone-specific actions for each zone in the building.

It is essential to highlight that the MADDPG algorithm employs $2M + 2$ networks (with M being the total number of zones) compared to the four neural networks used in single-agent DDPG. This increase in the number of networks leads to longer training times for the MADDPG algorithm compared to DDPG, as will be demonstrated in Section 4.

To learn in the multi-agent environment better, we utilized Cyclical Learning Rates (CLR) [61] to enhance neural network training in MADDPG. CLR dynamically adjusts the learning rate during training, allowing it to increase and decrease within a single run. By cycling between upper and lower bounds, the network explores a wider range of learning rates, improving performance and convergence. CLR benefits neural network training by preventing instabilities and escaping saddle points. It facilitates faster traversal across the loss landscape, leading to better solutions. Therefore, implementing CLR in MADDPG optimizes the networks' generalization and optimization capabilities, while dynamic learning rate adjustment improves parameter space exploration, resulting in superior model performance.

3. Simulation result

3.1. Simulation setup

A single-floor rectangular building with five zones (containing one interior and four exterior zones) is used to simulate the practical building, which features windows on all four facades and glass doors on the south and north facades. The HVAC system incorporates a packaged variable air volume system with direct expansion cooling coils and gas heating coils, serving the five zones. The HVAC parameters are listed in Table 3. For the simulation, the one-year weather dataset of Tempa, Florida, USA, is employed, providing detailed measurements at 15-minute intervals. This dataset includes all needs for the external state, as shown in Fig. 3. Time-of-Use (ToU) electricity tariff chosen is the Pacific Gas & Electric EToU-E6, which consists of three price levels: the base, shoulder, and peak prices, represented by white, light grey, and grey colors, respectively, with unit costs of \$0.244/kWh, \$0.32/kWh, and \$0.436/kWh in Fig. 5. Note that the current validation in this study utilizes EnergyPlus as the whole-building simulator. The MADDPG algorithm is implemented in Python 3.8.10 with BCVTB serving as the interface. The computational platform used for the experiments is a PC

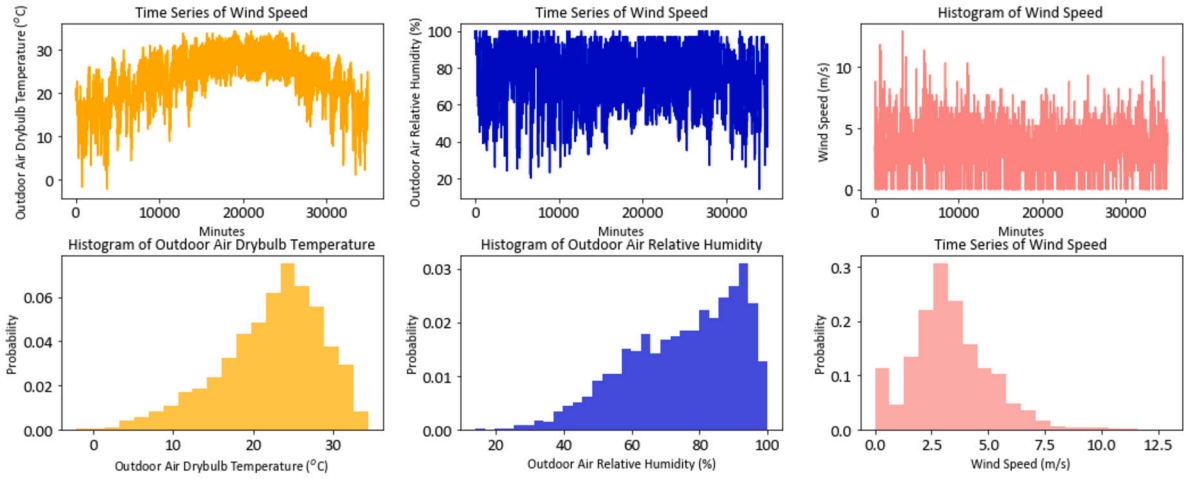


Fig. 3. One-year Tampa weather profile.

Table 4
MADDPG network information.

Actor Network	# of Neurals	Activation
Input	Input(shape = (# of state))	
Hidden 1	Dense(128, 256)	Relu
Hidden 2	Dense(256, 512)	Relu
Output	Dense(128, # of action)	Sigmoid
Critic Network	**# of Neurals**	**Activation**
State Input	Input(shape = (# of state))	
State Hidden	Dense(256, 256)	Relu
State Output	Dense(256, 256)	Relu
Action Input	Input(shape = (# of action))	
Action Output	Dense(256, 64)	Relu
Concatenate	Concatenate([State Output, Action Output])	
Hidden 1	Dense(Concatenate, 512)	Relu
Hidden 2	Dense(512,256)	Relu
Output	Dense(1)	

equipped with an Intel(R) Core(TM) i7-4790 CPU and 8 GB RAM with Windows Subsystem for Linux (WSL) Version 2.

To clearly show the efficacy of the proposed approach, we compare the proposed MADDPG with 1) a single-agent DDPG counterpart (control with occupant behavior) and 2) a rule-based control scheme (control without occupant behavior), which is discussed in [62,63]. The rule-based control highlights its reliance on outside temperature for HVAC electricity cost reduction in a whole-year hourly time step simulation. This method optimizes electricity cost by allowing slightly wider temperature ranges while ensuring occupant comfort which is deployed in realistic buildings.

3.2. Result

This section presents the simulation results for three compared methods: Rule-based method (from [62], DDPG (from [58]) and MADDPG (from Algorithm 1)), all applied to the testbed of a 5-zone building with 15 minutes time interval simulation model. The neural network information, including input, hidden, and output layers for actor and critic networks, is shown in Table 4. Table 5 displays the total training time for DDPG and MADDPG, along with the single-run execution times for the rule-based method, DDPG, and MADDPG. It is important to note that the rule-based method does not utilize any deep neural network for computation, making it significantly faster in execution time compared to the learning-based approaches. The single-run execution time refers to the time taken for a one-time decision-making process. Specifically, DDPG takes 5 hours to train and 6.25 ms for one-time decision-making, while MADDPG takes 32 hours (5.4 times longer

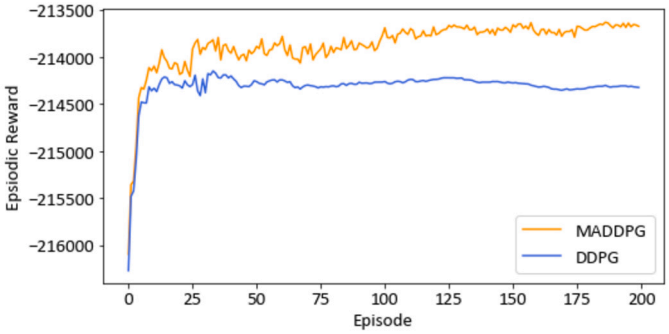


Fig. 4. Average episode reward during training of DDPG and MADDPG.

Table 5
Training and execution time comparison.

Item	Rule-based	DDPG	MADDPG
Training time	N/A	5 hr	32 hr
Single-run execution time	0.01 ms	6.25 ms	29.17 ms

than DDPG) for training and 29.17 ms (3.66 times longer than DDPG) for one-time decision-making. Additionally, the reward function plot (Fig. 4) displays the training rewards for the episodes. It is evident that MADDPG exhibits better reward returns after approximately 25 episodes of training compared to DDPG. It is important to note that MADDPG takes longer to train compared to DDPG due to the presence of more neural networks. However, this extended training time leads to higher reward returns after the training process.

In the following, we present the simulation results for three days, encompassing occupant behavior, decision results for all zones, and the whole building's electricity cost with average PMV-PPD. Fig. 5 provides a visualization of the three-day simulation of occupant behavior, illustrating fluctuations in metabolic rate, occupant presence, and clothing adjustments across all zones. The base price, shoulder price, and peak price are visually represented in white, light grey, and grey colors, respectively. It's essential to emphasize that our simulation environment is grounded in stochastic behavior modeling rather than a fixed schedule. This deliberate choice allows us to accurately capture the significant variability in occupant thermal comfort experiences over time. Furthermore, our approach offers flexibility in adjusting stochastic modeling parameters to align with real-world scenarios, leveraging insights from historical datasets [64]. It's worth noting that in our modeling, we assume that clothing adjustments play a relatively minor role in the overall variations. Detailed schedules for clothing behavior, metabolic

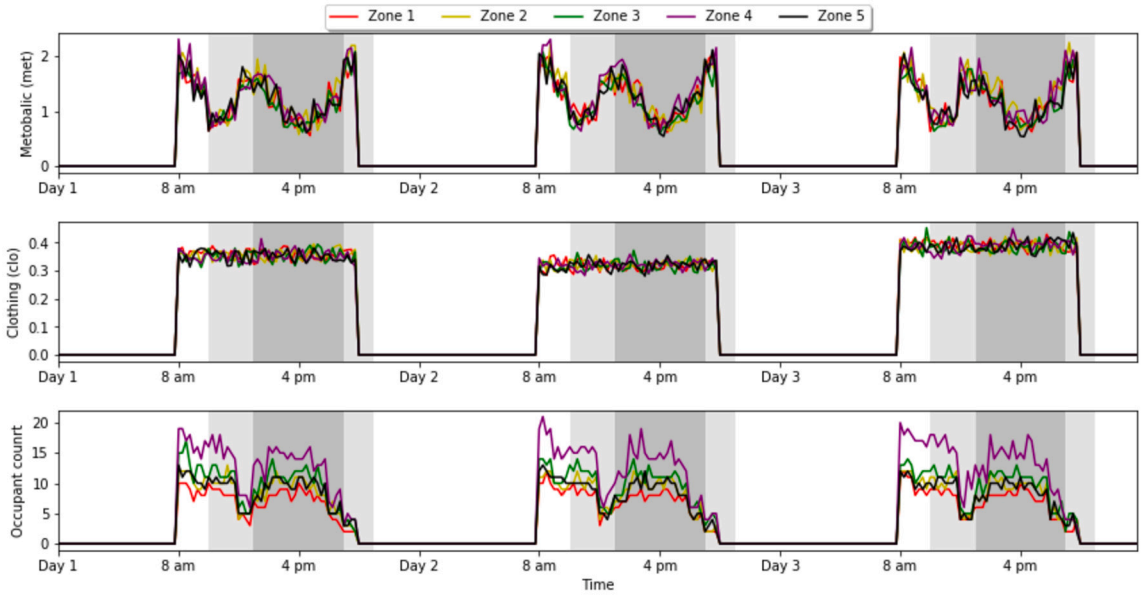


Fig. 5. Three-day behavior simulation result for all zones (base, shoulder, and peak prices are represented by white, light grey, and grey).

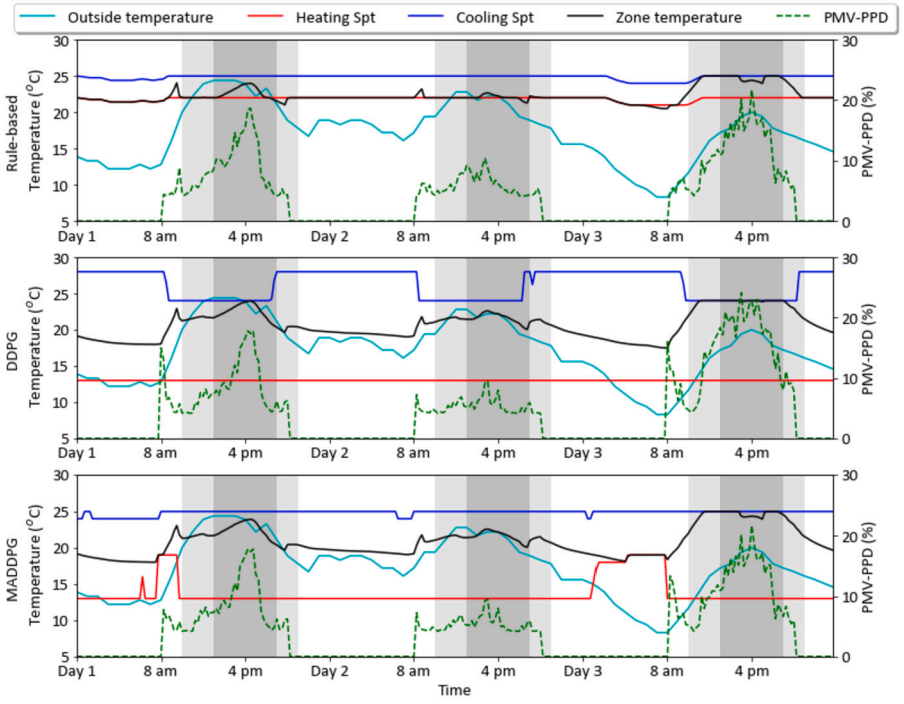
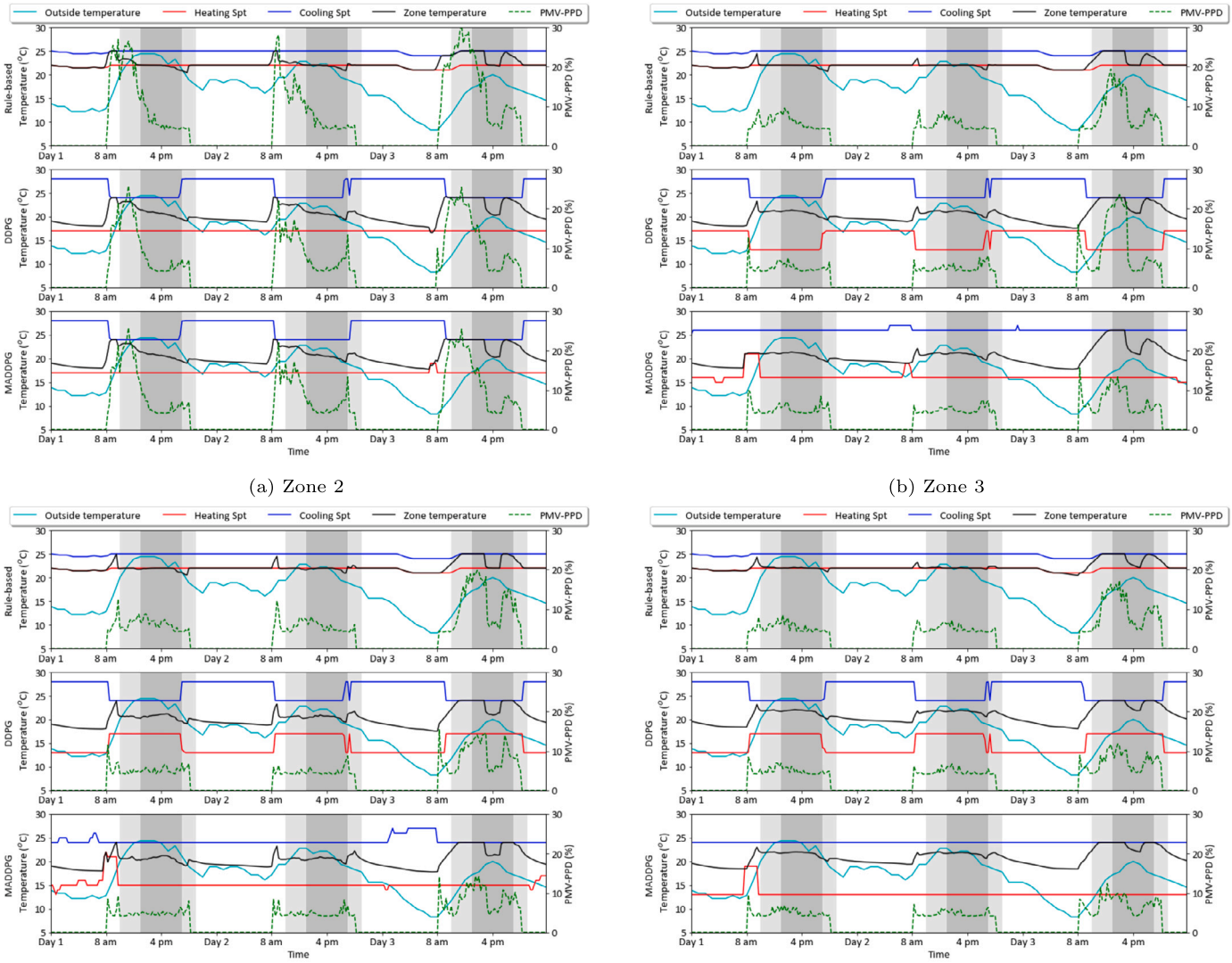


Fig. 6. Results of Zone 1 over a three-day period, with the Rule-based method (top), DDPG (middle), and MADDPG (bottom). The red line represents heating setpoints, the blue line indicates cooling setpoints, the black line represents the zone's average temperature, and the green dashed line represents the PMV-PPD value, while the light blue line depicts the outside temperature.

rate, and occupant presence are elaborated upon in Section 2.2 for a more comprehensive understanding.

Fig. 6 compares decision results, including heating and cooling setpoints of HVAC, for the three methods, along with room and outside temperatures. The weather dataset contains a wealth of information, with the outside temperature being the major factor affecting the room temperature. In the rule-based simulations, the control range (between heating and cooling setpoints) is narrow, as it solely considers the outside temperature for the control strategy. In addition, the heating and cooling set points for our rule-based method are derived from [62,63], in which high cooling and heating set points are observed in the early morning. Therefore, the rule-based method is typically less adaptive than DDPG and MADDPG which considers occupant behavior due to limited control range. As a result, the room temperature is controlled within this range without considering occupant behavior. In contrast, both DDPG and MADDPG show a wider range of temperature control as they take into account the occupant's behavior. Specifically, DDPG cools the zone once on Day 1 around 4 pm and requires no control on Day 2, while on Day 3, it cools the zone from 9 am to 6 pm. On the other hand, MADDPG adapts to the outside environment on Day 1 and Day 2. On Day 3, it maintains climate control from 9 am to 6 pm but undergoes a two-hour shift from the grey area (high electricity price) to the white area (based price). Furthermore, there are two pre-heating phase on Day 1 around 8 am and Day 3 from about 5 am to 8 am to raise the



(a) Zone 2

(b) Zone 3

(c) Zone 4

(d) Zone 5

Fig. 7. Results of Zone 2, 3, 4, and 5 over a three-day period. The red line represents heating setpoints, the blue line indicates cooling setpoints, the black line represents the zone's average temperature, and the green dashed line represents the PMV-PPD value, while the light blue line depicts the outside temperature.

zone temperature due to its base price period, leading to lower electricity costs while ensuring similar thermal comfort levels. The decisions made by MADDPG demonstrate a more adaptive and cost-effective approach, taking into account both the outside environment and occupant behavior, resulting in better energy efficiency and comfort management. Fig. 7 presents a visual representation of the three-day simulation results for Zones 2 to 5, offering insights into the strategies employed by the three methods. It's worth noting that the single-agent method demonstrates consistent actions across all zones, while MADDPG exhibits adaptive actions tailored to each zone's unique requirements. Importantly, MADDPG effectively avoids peak ToU hours, as evidenced by its morning actions around 7 am or 8 am, leading to cost savings. Additionally, MADDPG maintains a high level of thermal comfort, as reflected in the lower PMV-PPD values compared to the single-agent method.

Fig. 8 illustrates the evaluation metrics for the three methods, including HVAC whole building cost and average PMV-PPD in the five zones. As expected, the rule-based method incurs the highest electricity cost due to its narrow control climate range, which limits zone temperatures and leads to increased electricity consumption. This highlights the crucial importance of considering occupant behavior in the simulation. Regarding thermal comfort, the PMV-PPD values for the first two days are similar across all methods. However, on the last day, both the rule-based and MADDPG methods control the temperature at 25 ^{o}C, while DDPG maintains it at 24 ^{o}C. This slight difference in temperature control results in varied performance in thermal comfort, as evident in the PMV-PPD results. Notably, the pre-heating on the last day contributes to higher electricity costs in MADDPG but provides more comfortable thermal control compared to DDPG. The findings underscore the significance of incorporating occupant behavior into the simulation. By accounting for occupant preferences and adjusting temperature settings accordingly, MADDPG optimizes both electricity costs and thermal comfort, offering a more adaptive and efficient HVAC control strategy for the building. This approach provides a more personalized and responsive solution that takes into account occupant comfort while achieving energy savings.

An interesting finding is that the single-agent DDPG demonstrates a nearly identical control action for all zones, as depicted in Fig. 9. The control pattern remains consistent over the time horizon, with a relaxation of the control range during nights and a narrowing of the range during daytime. However, a crucial limitation of this approach is that it uses similar temperature setpoints for all zones, therefore covering each other plot line in the top (rule-based) and middle (DDPG) in Fig. 9, which does not reflect the actual real-world conditions in the

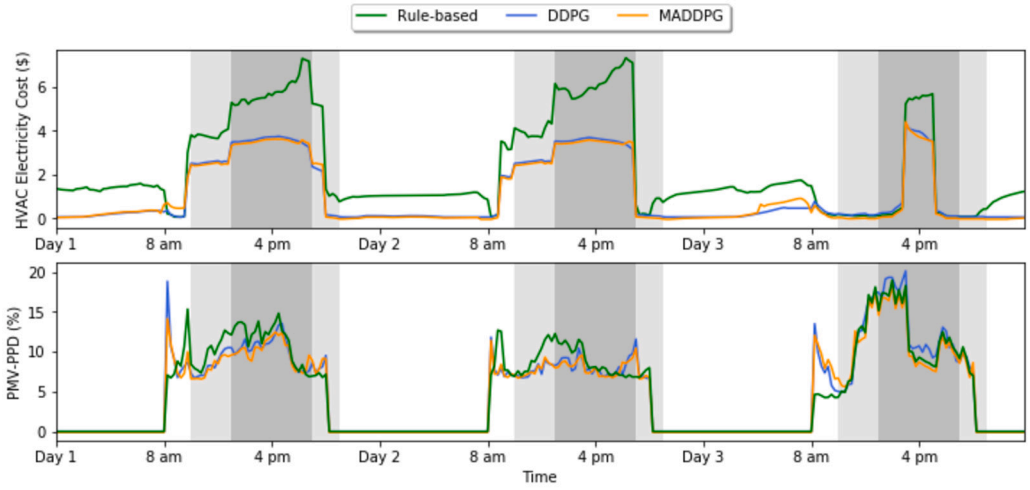


Fig. 8. Three days simulation of electricity cost and thermal comfort result (PMV-PPD) for the whole building.

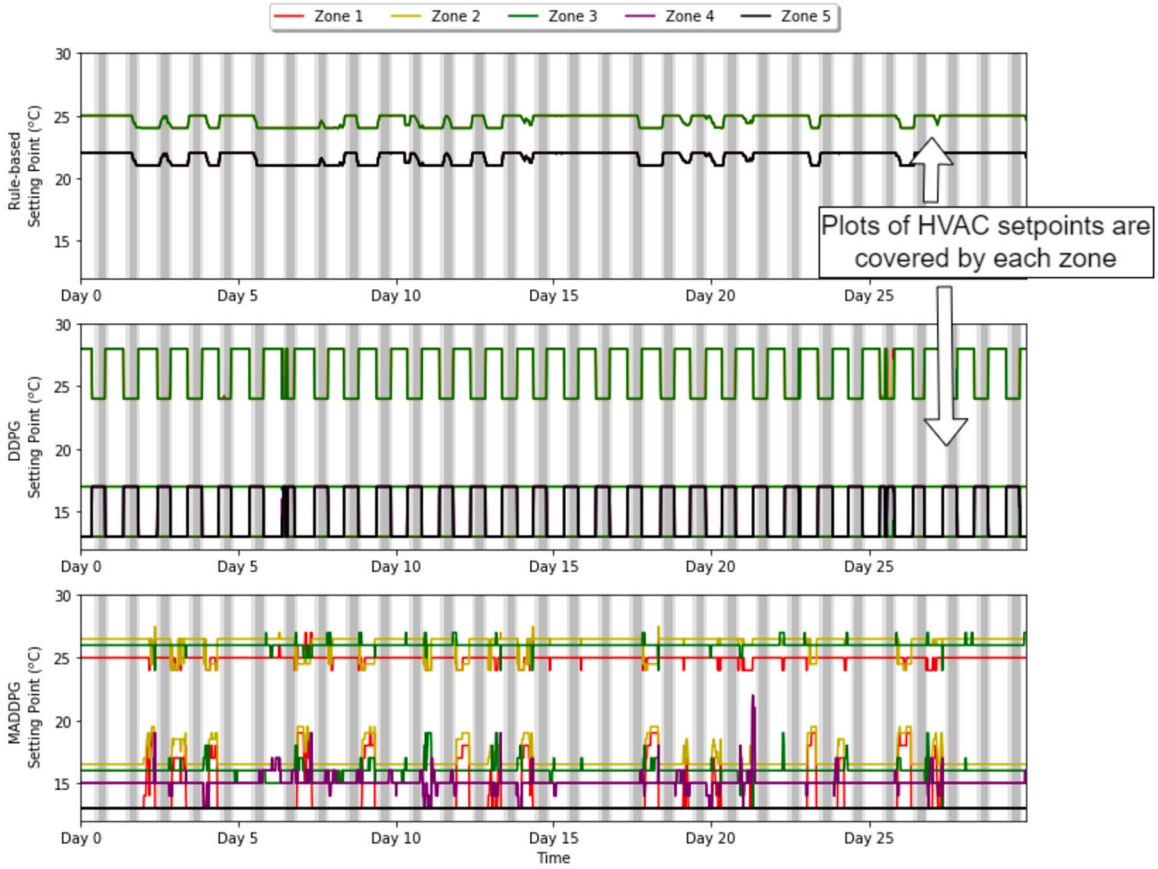


Fig. 9. One-month control patterns of HVAC for all zones.

simulations. Fig. 10 displays the decision distribution over one month of simulation. It becomes evident that MADDPG provides more options in terms of temperature settings, while DDPG and rule-based control exhibit a more limited range of options in the simulation results. This phenomenon indicates that MADDPG employs a more detailed and zone-specific control strategy, while DDPG applies the similar control approach to all five zones. Consequently, DDPG fails to capture the distinct differences between zones within the building, resulting in different simulation outcomes. This observation highlights the advantage of using MADDPG, which allows for more personalized and adaptive control strategies for each zone, leading to enhanced performance and efficiency in the overall HVAC management within the building.

To comprehensively represent the simulation results at the long-term level, one-year worth of weather data was employed in this study from the Tampa dataset from January 1 to December 31. Table 6 presents a one-year comparison of the three methods, showcasing the average daily electricity cost ($/day) and average PMV-PPD across all five zones during the year-long simulation. The rule-based control approach exhibits the highest electricity cost but maintains the best thermal comfort levels. However, this advantage in thermal comfort comes

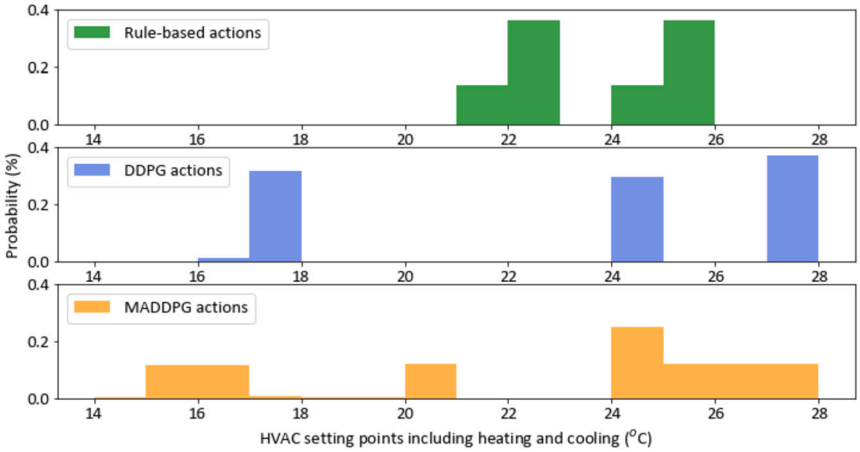


Fig. 10. One-month HVAC decisions histogram comparison of whole building.

Table 6
One-year comparison between different methods.

Item	Rule-based	DDPG	MADDPG
Average Daily Electricity Cost ($ / Day)	216.77	110.83	106.02
Avg. PMV-PPD (%)	10.36	11.43	10.39

at the cost of overlooking occupant behavior, as the rule-based control solely relies on outside temperature, leading to non-optimal energy usage. In contrast, MADDPG stands out with exceptional performance, achieving the lowest electricity cost, which is 51.09% lower than the rule-based control and 4.36% lower than DDPG. Importantly, MADDPG not only excels in cost-efficiency but also enhances thermal comfort levels compared to DDPG. Its capacity to strike a harmonious balance between cost-effectiveness and occupant comfort positions it as a highly promising and superior choice for HVAC control in buildings.

MADDPG achieves this through a multi-agent control strategy, which takes into account the unique thermal transfer characteristics and occupant behaviors in different zones. The control strategy, as in DDPG, would not capture these differences and result in suboptimal multi-zone thermal control, as shown in Fig. 10. MADDPG's incorporation of multi-agent techniques, training the agents to cooperate and coordinate within the environment, leads to better results in this study. The approach effectively optimizes both electricity cost and thermal comfort, demonstrating its superiority over the rule-based and single-agent DDPG methods. The consideration of occupant behavior and zone-specific thermal dynamics in MADDPG results in more adaptive and efficient HVAC control, offering a more desirable outcome for building occupants and energy management.

4. Discussion and future study

The findings from the figures and table in this study highlight the effectiveness of the MADDPG approach in multi-zone HVAC control. MADDPG operates by making pre-emptive decisions to avoid peak ToC price periods while ensuring consistent thermal comfort. Its capability to customize HVAC setpoints individually for different zones, as opposed to applying uniform actions across all zones, significantly enhances its ability to reduce electricity costs compared to the rule-based approach. Furthermore, MADDPG exhibits superior thermal comfort performance when contrasted with the single-agent DDPG method. In addition, over the course of one year, MADDPG consistently outperformed the single-agent DDPG methods in terms of average daily electricity cost, while maintaining superior thermal comfort levels compared to DDPG, striking a balance between cost reduction and occupant satisfaction. This improved performance is attributed to MADDPG's ability to adaptively design HVAC setpoints based on zone-specific characteristics and occupant behaviors, effectively utilizing multi-agent techniques to optimize both electricity cost and thermal comfort.

Regarding the issues of similar actions for all zones in DDPG, we realize that during the training phase, all agents participate in centralized learning, sharing a global value function that captures interactions and dependencies between agents and the environment. However, this collective approach does not distinguish individual agents, treating those agents as a unified entity. Consequently, the shared knowledge acquired through single-agent DDPG tends to result in similar actions among agents, particularly when they face similar environmental conditions such as a testbed of multi-zone HVAC building. However, MADDPG is tailored for multi-agent environments, which facilitates communication and coordination among multiple agents, enabling them to learn distinct policies and make varied decisions based on observations and other agents' actions.

This study serves as a proof of concept for applying MADRL to multi-zone thermal control while considering dynamic occupant behavior. One of the critical avenues for future research is the validation of this approach in real-world environments. The validation process entails several essential steps to ensure the algorithm's practicality and effectiveness in realistic scenarios. Firstly, it involves establishing the necessary hardware and software infrastructure, gathering real-world data, including occupant behavior and weather variations, and constructing a simulation model for virtual testing using the EnergyPlus simulator. Following this, the MADDPG algorithm undergoes offline training before its implementation within a realistic building system. Continuous system operation over an extended period facilitates ongoing performance monitoring of MADDPG, including model calibration for building models and parameter updates based on user feedback. Finally, tracking electricity consumption and occupant comfort, followed by comprehensive data analysis, evaluates MADDPG's performance compared to baseline approaches in realistic buildings.

5. Conclusions

This study introduces a multi-zone HVAC energy management approach using MADDPG to minimize electricity costs and enhance occupants' thermal comfort. The MADDPG model proves its effectiveness in multi-zone HVAC control by accounting for dynamic occupant behavior and zone-specific thermal dynamics. Simulation results highlight its capacity to optimize both electricity cost and thermal comfort, surpassing rule-based and single-agent DDPG methods. The incorporation of multi-agent techniques empowers personalized and adaptive control strategies for each zone, leading to improved HVAC management performance and efficiency. MADDPG demonstrates its adaptability by making informed cooling and heating setpoint decisions based on ex-

ternal conditions and occupant behavior. Moreover, over a year-long simulation, MADDPG consistently achieves the lowest electricity cost while striking a remarkable balance between energy savings and occupant comfort. The inclusion of stochastic modeling, as exemplified in this study, opens new avenues for more realistic and sophisticated building energy management systems. Future studies include considering an energy management system that focuses on multi-zone HVAC control with the effect of bioclimatic and passive strategy to achieve higher reward feedback.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgement

This work is supported in part by the U.S. National Science Foundation under Grant No. 1856084 for the FEWtures project.

References

[1] A. Pratt, D. Krishnamurthy, M. Ruth, H. Wu, M. Lunacek, P. Vaynshenk, Transactive home energy management systems: the impact of their proliferation on the electric grid, IEEE Electrif. Mag. 4 (4) (2016) 8–14, https://doi.org/10.1109/MELE.2016.2614188.

[2] X. Zhang, M. Pipattanasomporn, T. Chen, S. Rahman, An iot-based thermal model learning framework for smart buildings, IEEE Int. Things J. 7 (1) (2020) 518–527, https://doi.org/10.1109/JIOT.2019.2951106.

[3] J.Y. Park, M.M. Ouf, B. Gunay, Y. Peng, W. O'Brien, M.B. Kjærgaard, Z. Nagy, A critical review of field implementations of occupant-centric building controls, Build. Environ. 165 (2019) 106351, https://doi.org/10.1016/j.buildenv.2019.106351, https://www.sciencedirect.com/science/article/pii/S036013231930561X.

[4] Z. Nagy, B. Gunay, C. Miller, J. Hahn, M. Ouf, S. Lee, B.W. Hobson, T. Abuimara, K. Bandurski, M. André, C.-L. Lorenz, S. Crosby, B. Dong, Z. Jiang, Y. Peng, M. Favero, J.Y. Park, K. Nweye, P. Nojedehi, H. Stopps, L. Sarran, C. Brackley, K. Bassett, K. Govertsen, N. Koczorek, O. Abele, E. Casavant, M. Kane, Z. O'Neill, T. Yang, J. Day, B. Huchuk, R.T. Hellwig, M. Vellei, Ten questions concerning occupant-centric control and operations, Build. Environ. (2023) 110518, https://doi.org/10.1016/j.buildenv.2023.110518, https://www.sciencedirect.com/science/article/pii/S0360132323005450.

[5] D. Wang, C.C. Federspiel, F. Rubinstein, Modeling occupancy in single person offices, Energy Build. 37 (2) (2005) 121–126, https://doi.org/10.1016/j.enbuild.2004.06.015, https://www.sciencedirect.com/science/article/pii/S0378778804001951.

[6] C.F. Reinhart, Lightswitch-2002: a model for manual and automated control of electric lighting and blinds, Sol. Energy 77 (2004) 15–28, https://doi.org/10.1016/J.SOLENER.2004.04.003.

[7] J. Page, D. Robinson, N. Morel, J.L. Scartezzini, A generalised stochastic model for the simulation of occupant presence, Energy Build. 40 (2008) 83–98, https://doi.org/10.1016/J.ENBUILD.2007.01.018.

[8] L. Klein, J.Y. Kwak, G. Kavulya, F. Jazizadeh, B. Becerik-Gerber, P. Varakantham, M. Tambe, Coordinating occupant behavior for building energy and comfort management using multi-agent systems, Autom. Constr. 22 (2012) 525–536, https://doi.org/10.1016/J.AUTCON.2011.11.012.

[9] V. Fabi, R.K. Andersen, S. Corgnati, Verification of stochastic behavioral models of occupants' interactions with windows in residential buildings, Build. Environ. 94 (2015) 371–383, https://doi.org/10.1016/J.BUILDENV.2015.08.016.

[10] X. Liu, Y. Wu, H. Wu, PV-EV integrated home energy management considering residential occupant behaviors, Sustainability (Switzerland) 13 (24) (2021), https://doi.org/10.3390/su132413826.

[11] X. Liu, Y. Wu, H. Zhang, H. Wu, Hourly occupant clothing decisions in residential HVAC energy management, J. Build. Eng. 40 (2021) 102708, https://doi.org/10.1016/j.jobe.2021.102708.

[12] X. Liu, Y. Wu, H. Zhang, B. Liu, L. Edmonds, H. Wu, Home energy management with clothing integrated thermal comfort and ev soc concern, in: 2022 IEEE Power & Energy Society General Meeting (PESGM), Denver, CO, USA, 2022, pp. 01–05.

[13] D. Liu, Y. Xu, Q. Wei, X. Liu, Residential energy scheduling for variable weather solar energy based on adaptive dynamic programming, IEEE/CAA J. Autom. Sin. 5 (1) (2018) 36–46, https://doi.org/10.1109/JAS.2017.7510739.

[14] H. Wu, A. Pratt, P. Munankarmi, M. Lunacek, S.P. Balamurugan, X. Liu, P. Spitsen, Impact of model predictive control-enabled home energy management on large-scale distribution systems with photovoltaics, Adv. Appl. Energy 6 (2022) 100094, https://doi.org/10.1016/J.ADAPEN.2022.100094.

[15] F. Luo, Z.Y. Dong, K. Meng, J. Wen, H. Wang, J. Zhao, An operational planning framework for large-scale thermostatically controlled load dispatch, IEEE Trans. Ind. Inform. 13 (1) (2017) 217–227, https://doi.org/10.1109/TII.2016.2515086.

[16] F. Luo, G. Ranzi, C. Wan, Z. Xu, Z.Y. Dong, A multistage home energy management system with residential photovoltaic penetration, IEEE Trans. Ind. Inform. 15 (1) (2019) 116–126, https://doi.org/10.1109/TII.2018.2871159.

[17] Y. Liu, D. Zhang, H.B. Gooi, Optimization strategy based on deep reinforcement learning for home energy management, CSEE J. Power Energy Syst. 6 (3) (2020) 572–582, https://doi.org/10.17775/CSEEJPES.2019.02890.

[18] G. Barone, A. Buonomano, C. Forzano, G. Giuzio, A. Palombo, G. Russo, A new thermal comfort model based on physiological parameters for the smart design and control of energy-efficient hvac systems, Renew. Sustain. Energy Rev. 173 (2023) 113015, https://doi.org/10.1016/j.rser.2022.113015, https://www.sciencedirect.com/science/article/pii/S1364032122008966.

[19] X. Zhao, Y. Yin, Z. He, Z. Deng, State-of-the-art, challenges and new perspectives of thermal comfort demand law for on-demand intelligent control of heating, ventilation, and air conditioning systems, Energy Build. 295 (2023) 113325, https://doi.org/10.1016/j.enbuild.2023.113325, https://www.sciencedirect.com/science/article/pii/S0378778823005558.

[20] Y. Fu, S. Xu, Q. Zhu, Z. O'Neill, V. Adetola, How good are learning-based control v.s. model-based control for load shifting? Investigations on a single zone building energy system, Energy 273 (2023) 127073, https://doi.org/10.1016/j.energy.2023.127073, https://www.sciencedirect.com/science/article/pii/S036054422300467X.

[21] T. Wei, Y. Wang, Q. Zhu, Deep reinforcement learning for building HVAC control, in: Proceedings of the 54th Annual Design Automation Conference 2017, Austin, TX, USA, 2017, pp. 1–6.

[22] Y. Ye, D. Qiu, H. Wang, Y. Tang, G. Strbac, Real-time autonomous residential demand response management based on twin delayed deep deterministic policy gradient learning, Energies 14 (3) (2021) 531, https://doi.org/10.3390/EN14030531.

[23] Y. Zhang, X. Bai, F.P. Mills, J.C. Pezzey, Rethinking the role of occupant behavior in building energy performance: a review, Energy Build. 172 (2018) 279–294, https://doi.org/10.1016/J.ENBUILD.2018.05.017.

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

[25] A. Kathirgamanathan, E. Mangina, D. Finn, Development of a soft actor critic deep reinforcement learning approach for harnessing energy flexibility in a large office building, Energy AI 5 (2021), https://doi.org/10.1016/j.egyai.2021.100101.

[26] R. Homod, H. Togun, A.K. Hussein, F.N. Al-Mousawi, Z. Yaseen, W. Al-Kouz, H. Abd, O. Alawi, M. Goodarzi, O. Hussein, Dynamics analysis of a novel hybrid deep clustering for unsupervised learning by reinforcement of multi-agent to energy saving in intelligent buildings, Appl. Energy 313 (2022), https://doi.org/10.1016/j.apenergy.2022.118863.

[27] L. Yu, Y. Sun, Z. Xu, C. Shen, D. Yue, T. Jiang, X. Guan, Multi-agent deep reinforcement learning for hvac control in commercial buildings, IEEE Trans. Smart Grid 12 (1) (2021) 407–419, https://doi.org/10.1109/TSG.2020.3011739.

[28] A. Naug, M. Quinones-Grueiro, G. Biswas, Deep reinforcement learning control for non-stationary building energy management, Energy Build. 277 (2022), https://doi.org/10.1016/j.enbuild.2022.112584.

[29] D. Bayer, M. Pruckner, Enhancing the Performance of Multi-Agent Reinforcement Learning for Controlling HVAC Systems, 2022, pp. 187–194.

[30] Z. Zhang, K.P. Lam, Practical implementation and evaluation of deep reinforcement learning control for a radiant heating system, in: Proceedings of the 5th Conference on Systems for Built Environments, Shenzen, China, 2018, pp. 148–157.

[31] D. Azuatalam, W.L. Lee, F. de Nijs, A. Liebman, Reinforcement learning for whole-building hvac control and demand response, Energy AI 2 (2020), https://doi.org/10.1016/J.EGYAI.2020.100020.

[32] J. Li, W. Zhang, G. Gao, Y. Wen, G. Jin, G. Christopoulos, Toward intelligent multi-zone thermal control with multiagent deep reinforcement learning, IEEE Int. Things J. 8 (14) (2021) 11150–11162, https://doi.org/10.1109/JIOT.2021.3051400.

[33] D. Coraci, S. Brandi, M. Piscitelli, A. Capozzoli, Online implementation of a soft actor-critic agent to enhance indoor temperature control and energy efficiency in buildings, Energies 14 (2021), https://doi.org/10.3390/en14040997.

[34] D. Deltetto, D. Coraci, G. Pinto, M. Piscitelli, A. Capozzoli, Exploring the potentialities of deep reinforcement learning for incentive-based demand response in a cluster of small commercial buildings, Energies 14 (2021), https://doi.org/10.3390/en14102933.

[35] Z. Li, Z. Sun, Q. Meng, Y. Wang, Y. Li, Reinforcement learning of room temperature set-point of thermal storage air-conditioning system with demand response, Energy Build. 259 (2022), https://doi.org/10.1016/j.enbuild.2022.111903.

[36] M. Esrafilian-Najafabadi, F. Haghighat, Towards self-learning control of hvac systems with the consideration of dynamic occupancy patterns: application of model-free deep reinforcement learning, Build. Environ. 226 (2022) 109747, https://doi.org/10.1016/j.buildenv.2022.109747, https://www.sciencedirect.com/science/article/pii/S0360132322009775.

[37] X. Lu, Y. Fu, Z. O'Neill, Benchmarking high performance hvac rule-based controls with advanced intelligent controllers: a case study in a multi-zone system in modelica, Energy Build. 284 (2023) 112854, https://doi.org/10.1016/J.ENBUILD.2023.112854.

[38] S. Qiu, Z. Li, Z. Pang, Z. Li, Y. Tao, Multi-agent optimal control for central chiller plants using reinforcement learning and game theory, Systems 11 (2023), https://doi.org/10.3390/systems11030136.

[39] M. Wetter, Co-simulation of building energy and control systems with the building controls virtual test bed, J. Build. Perform. Simul. 4 (3) (2010) 185–203, https://doi.org/10.1080/19401493.2010.518631.

[40] X. Wang, B. Dong, Physics-informed hierarchical data-driven predictive control for building hvac systems to achieve energy and health nexus, Energy Build. 291 (2023) 113088, https://doi.org/10.1016/j.enbuild.2023.113088, https://www.sciencedirect.com/science/article/pii/S0378778823003183.

[41] H. Li, H. Johra, F. de Andrade Pereira, T. Hong, J. Le Dréau, A. Maturo, M. Wei, Y. Liu, A. Saberi-Derakhtenjani, Z. Nagy, A. Marszal-Pomianowska, D. Finn, S. Miyata, K. Kaspar, K. Nweye, Z. O'Neill, F. Pallonetto, B. Dong, Data-driven key performance indicators and datasets for building energy flexibility: a review and perspectives, Appl. Energy 343 (2023) 121217, https://doi.org/10.1016/j.apenergy.2023.121217, https://www.sciencedirect.com/science/article/pii/S0306261923005810.

[42] M. Kong, B. Dong, R. Zhang, Z. O'Neill, Hvac energy savings, thermal comfort and air quality for occupant-centric control through a side-by-side experimental study, Appl. Energy 306 (2022) 117987, https://doi.org/10.1016/j.apenergy.2021.117987, https://www.sciencedirect.com/science/article/pii/S0306261921012903.

[43] P.O. Fanger, Thermal Comfort. Analysis and Applications in Environmental Engineering, Danish Technical Press, Copenhagen, 1970, p. 1970.

[44] UNI EN ISO 7730, Ergonomics of the thermal environment — analytical determination and interpretation of thermal comfort using calculation of the PMV and PPD indices and local thermal comfort criteria, 2005.

[45] ASHRAE Standard 55/2004, Thermal environmental conditions for human occupancy, 2004.

[46] H.B. Gunay, W. O'Brien, I. Beausoleil-Morrison, A critical review of observation studies, modeling, and simulation of adaptive occupant behaviors in offices, Build. Environ. 70 (2013) 31–47, https://doi.org/10.1016/J.BUILDENV.2013.07.020.

[47] D. Wang, C.C. Federspiel, F. Rubinstein, Modeling occupancy in single person offices, Energy Build. 37 (2005) 121–126, https://doi.org/10.1016/J.ENBUILD.2004.06.015.

[48] H.B. Gunay, W. O'Brien, I. Beausoleil-Morrison, Implementation and comparison of existing occupant behaviour models in energyplus, J. Build. Perform. Simul. 9 (2015) 567–588, https://doi.org/10.1080/19401493.2015.1102969, https://www.tandfonline.com/doi/abs/10.1080/19401493.2015.1102969.

[49] S. Carlucci, M.D. Simone, S.K. Firth, M.B. Kjærgaard, R. Markovic, M.S. Rahaman, M.K. Annaqeeb, S. Biandrate, A. Das, J.W. Dziedzic, G. Fajilla, M. Favero, M. Ferrando, J. Hahn, M. Han, Y. Peng, F. Salim, A. Schlüter, C. van Treeck, Modeling occupant behavior in buildings, Build. Environ. 174 (2020) 106768, https://doi.org/10.1016/J.BUILDENV.2020.106768.

[50] H. Choi, B. Jeong, J. Lee, H. Na, K. Kang, T. Kim, Deep-vision-based metabolic rate and clothing insulation estimation for occupant-centric control, Build. Environ. 221 (2022) 109345, https://doi.org/10.1016/j.buildenv.2022.109345, https://www.sciencedirect.com/science/article/pii/S0360132322005789.

[51] A. Pratt, B. Banerjee, T. Nemarundwe, Proof-of-concept home energy management system autonomously controlling space heating, in: 2013 IEEE Power Energy Society General Meeting, Vancouver, BC, Canada, 2013, pp. 1–5.

[52] X. Liu, H. Wu, L. Wang, M.N. Faqiry, Stochastic home energy management system via approximate dynamic programming, IET Energy Syst. Integr. 2 (4) (2020) 382–392, https://doi.org/10.1049/iet-esi.2020.0060.

[53] H. Wu, A. Pratt, S. Chakraborty, Stochastic optimal scheduling of residential appliances with renewable energy sources, in: IEEE Power Energy Society General Meeting, 2015, pp. 1–5.

[54] M. Shafie-Khah, P. Siano, A stochastic home energy management system considering satisfaction cost and response fatigue, IEEE Trans. Ind. Inform. 14 (2) (2018) 629–638, https://doi.org/10.1109/TII.2017.2728803.

[55] A. Heidari, F. Maréchal, D. Khovalyg, Reinforcement learning for proactive operation of residential energy systems by learning stochastic occupant behavior and fluctuating solar energy: balancing comfort, hygiene and energy use, Appl. Energy 318 (2022) 119206, https://doi.org/10.1016/J.APENERGY.2022.119206.

[56] H. Wu, M. Shahidehpour, Stochastic scuc solution with variable wind energy using constrained ordinal optimization, IEEE Trans. Sustain. Energy 5 (2) (2014) 379–388, https://doi.org/10.1109/TSTE.2013.2289853.

[57] H. Wu, M. Shahidehpour, A. Alabdulwahab, A. Abusorrah, Thermal generation flexibility with ramping costs and hourly demand response in stochastic security-constrained scheduling of variable energy sources, IEEE Trans. Power Syst. 30 (6) (2015) 2955–2964, https://doi.org/10.1109/TPWRS.2014.2369473.

[58] T.P. Lillicrap, J.J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning, in: 4th International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2015.

[59] R. Lowe, Y. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, I. Mordatch, Multi-agent actor-critic for mixed cooperative-competitive environments, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017, https://proceedings.neurips.cc/paper/2017/file/68a9750337a418a86fe06c1991a1d64c-Paper.pdf.

[60] D.T. Gillespie, Exact numerical simulation of the Ornstein-Uhlenbeck process and its integral, Phys. Rev. E 54 (2) (1996) 2084–2091, https://doi.org/10.1103/PhysRevE.54.2084.

[61] L.N. Smith, Cyclical learning rates for training neural networks, arXiv:1506.01186, 2017.

[62] D. Gyalistras, M. Gwerder, et al., Use of Weather and Occupancy Forecasts for Optimal Building Climate Control (Opticontrol): Two Years Progress Report, Terrestrial Systems Ecology ETH Zurich, Switzerland and Building Technologies Division, vol. 158, Siemens Switzerland Ltd., Zug, Switzerland, 2010.

[63] D. Gyalistras, M. Gwerder, F. Oldewurtel, C.N. Jones, M. Morari, B. Lehmann, K. Wirth, V. Stauch, Analysis of energy savings potentials for integrated room automation, in: Clima - RHEVA World Congress, 2010, https://infoscience.epfl.ch/record/169732.

[64] D. Yan, W. O'Brien, T. Hong, X. Feng, H.B. Gunay, F. Tahmasebi, A. Mahdavi, Occupant behavior modeling for building performance simulation: current state and future challenges, Energy Build. 107 (2015) 264–278, https://doi.org/10.1016/J.ENBUILD.2015.08.032.