Building and Environment 223 (2022) 109458

Contents lists available at ScienceDirect

Building and Environment

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

Energy-efficient personalized thermal comfort control in office buildings based on multi-agent deep reinforcement learning

Liang Yu [a,b,*], Zhanbo Xu [a], Tengfei Zhang [b], Xiaohong Guan [a], Dong Yue [b]

[a] Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[b] College of Automation and College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

ARTICLE INFO

Keywords:
Office buildings
HVAC systems
Personal comfort systems
Personalized thermal comfort control
Energy-efficient
Multi-agent deep reinforcement learning

different trends and peaks at different points,

ABSTRACT

In a shared office space, the percentage of occupants with satisfied thermal comfort is typically low. The main reason is that heating, ventilation, and air conditioning (HVAC) systems cannot provide individual thermal environment for each occupant within the shared office space. Although personal comfort systems (PCSs) can be adopted to implement heterogeneous thermal environments, they have limited adjustment abilities. At this time, coordinating the operations of PCSs and an HVAC system is a good choice. In this paper, the coordination control problem of PCSs and an HVAC system in a shared office space is investigated to minimize the total energy consumption while maintaining comfortable individual thermal environment for each occupant. Specifically, we first formulate an expected energy consumption minimization problem related to PCSs and an HVAC system. Due to the existence of an inexplicit building thermal dynamics model and uncertain parameters, it is challenging to solve the problem. To overcome the challenge, we reformulate the problem as a Markov game with heterogeneous agents. To promote an efficient cooperation of such agents, we propose a real-time control algorithm based on attention-based multi-agent deep reinforcement learning, which does not require an explicit building thermal dynamics model and any prior knowledge of uncertain parameters. Simulation results based on real-world traces show that the proposed algorithm can reduce energy consumption by 0.7%–4.18% and reduce average thermal comfort deviation by 64.13%–72.08% simultaneously compared with baselines.

1. Introduction

Buildings account for a large portion of total energy consumption in a country. In buildings, 40%–50% of energy consumption is attributed to heating, ventilation, and air conditioning (HVAC) systems [1,2]. Although the energy consumption of the HVAC systems is significant, the percentage of occupants with satisfied thermal comfort in office buildings is still low. According to a large-scale survey, only 38% of occupants in office buildings are satisfied with their thermal environments [2]. Since thermal comfort is closely related to the productivity and health of occupants, building energy consumption, and the energy cost of building operators, it is necessary to reduce building energy cost/consumption while maintaining the thermal comfort of occupants in office buildings [3].

There have been many studies on thermal comfort control in office buildings. Generally, such studies could be categorized into three types according to the involved systems, i.e., HVAC control [4–7][8], personal comfort systems (PCSs) control [9,10], operation of HVAC systems and PCSs [11–13]. In the first type, real-time occupancy data, feedback information, or environment data are utilized to save energy and improve thermal comfort. Although some advances have been made, the HVAC system cannot provide an individual thermal environment for each occupant in a shared office space. In the second type, PCSs (e.g., desk fans, heaters) are adopted for adjusting the local thermal environment for each occupant. As a result, such systems can potentially satisfy heterogeneous thermal comfort requirements. However, PCSs typically have limited corrective power, which is defined as the difference between two ambient temperatures at which the same thermal sensation is achieved when PCSs are or are not used [14]. Therefore, the third type considers the operation of HVAC systems and PCSs simultaneously. Experimental results showed that the proposed methods in the third type can save energy and improve individual thermal comfort environments. However, the proposed methods in existing works [11,12] assume that an explicit building thermal dynamics model and accurate forecast information about weather and occupancy

* Corresponding author at: Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China.
E-mail addresses: liang.yu@njupt.edu.cn (L. Yu), zbxu@sei.xjtu.edu.cn (Z. Xu), tfzhang@126.com (T. Zhang), xhguan@sei.xjtu.edu.cn (X. Guan), medongy@vip.163.com (D. Yue).

https://doi.org/10.1016/j.buildenv.2022.109458
Received 13 April 2022; Received in revised form 25 July 2022; Accepted 25 July 2022
Available online 4 August 2022

are available. In fact, due to many complex and random factors, it is intractable to develop an explicit building thermal dynamics model that is accurate and efficient enough for building control [3]. Moreover, the existence of forecast errors will lead to performance degradation of methods in practice.

Based on the above observation, this paper investigates the problem of optimal coordinated operation of PCSs and an HVAC system in a shared office space when an explicit building thermal dynamics model and accurate forecast information about uncertain parameters are unavailable. To be specific, after the current system states are observed in each time slot, the temperature set-point of the HVAC system and speeds of desk fans should be jointly determined. The purpose is to minimize the accumulated system energy consumption during the given time horizon while maintaining comfortable thermal environments for all occupants. Since there are temporally coupled constraints related to indoor temperature, an inexplicit building thermal dynamics model, and many uncertain parameters, it is challenging to achieve the above aim. Typical methods for multi-stage decision-making under uncertainty include stochastic programming [15], model predictive control [12], robust optimization [16], Lyapunov optimization techniques [17]. However, such methods have certain limitations, e.g., stochastic programming and robust optimization need to know prior knowledge of uncertain parameters (e.g., probability distribution, maximum and minimum values), model predictive control needs to predict future uncertain parameters, model predictive control and Lyapunov optimization techniques need to know explicit building thermal dynamics models.

With the development of Internet of things and artificial intelligence technologies, many data-driven HVAC control methods have been developed based on deep reinforcement learning (DRL) [18–21], which is a new-generation general artificial intelligence technology and has achieved great success in many fields [22], e.g., robot control, network games, and recommendation systems. Although the above DRL-based control methods can operate without an explicit building thermal dynamics model and any prior knowledge of uncertain parameters, they cannot be applied to solve the energy consumption minimization problem in this paper directly. The reason is that the size of discrete action space increases exponentially with the increase of PCS numbers. To overcome this challenge, we propose a coordinated control algorithm based on multi-agent DRL [23].

In summary, the main contributions of this paper include:

- By taking an inexplicit building thermal dynamics model, parameter uncertainty, and heterogeneous thermal comfort requirements into consideration, we formulate an expected energy consumption minimization problem related to the coordination of PCSs and an HVAC system in a shared office space. Since it is difficult to solve the problem directly, we reformulate it as a Markov game, where state, action, and reward are designed.
- We propose a real-time control algorithm to promote efficient coordination among PCSs and HVAC system based on attention-based multi-agent DRL (AMADRL), which can operate without requiring an explicit building thermal dynamics model and forecasting uncertain parameters.
- We implement a co-simulation environment based on Python and EnergyPlus for the training and testing of the proposed algorithm, which considers an inexplicit building thermal dynamics model and more practical HVAC models. Moreover, simulation results show the effectiveness of the proposed algorithm.

The rest of this paper is organized as follows. In Section 2, related works are introduced. In Section 3, we provide the system model and problem formulation. In Section 4, the AMADRL-based coordination control algorithm is proposed. In Section 5, performance evaluations are conducted. Finally, conclusions are made in Section 6.

2. Related works

Since this paper focuses on the coordinated control of an HVAC system and PCSs in a shared office space, we introduce related works from three aspects, i.e., HVAC control in office spaces, PCS management in office spaces, and the operation of HVAC systems in the presence of PCSs in office spaces.

2.1. HVAC control in office spaces

Many approaches have been proposed to control HVAC systems in office spaces, e.g., rule-based control, model-based optimization control, and model-free learning control. For example, Gupta et al. proposed an end-to-end framework to collect real-time occupant feedback (e.g., occupant preference) in multi-zone, multi-occupant shared spaces [4]. Then, such feedback information was used by an environmental learning algorithm to control the HVAC temperature set-point dynamically so that the aggregated discomfort plus energy cost could be minimized. Similarly, Lee et al. proposed a self-tuned HVAC controller for optimizing energy consumption and thermal satisfaction in an office space based on model predictive control with dynamic indoor temperature control bounds, which are calculated from occupant feedbacks [6]. In addition to feedback information, occupancy information can also be used for efficient HVAC control, e.g., Peng et al. proposed a learning-based demand-driven control strategy to adjust the HVAC temperature set-point, which can obtain predicted occupancy information based on historical data [5]. Moreover, occupancy information could be inferred by real-time environmental parameters (e.g., indoor temperature, and CO_2 concentration) and used for controlling the HVAC system efficiently, e.g., Li et al. proposed a method for saving the energy consumption of the HVAC system in irregularly occupied office spaces by utilizing real-time environmental big data [7]. When large open office spaces are considered, nonuniform thermal distribution among sub-zones is common. To ensure uniform occupant comfort in a large open office space and avoid overcooling, Shan et al. proposed a simulation method to compute the optimal temperature set-point for each sub-zone [8].

With the development of Internet of things technology and artificial intelligence technology, many methods for controlling HVAC systems in office buildings have been developed based on DRL algorithms. For example, Wei et al. proposed a Deep Q-Network (DQN) based HVAC control method for reducing energy cost while maintaining a comfortable temperature range in office buildings [18]. Similarly, Zou et al. designed a method to control air handling units in office buildings for optimizing energy consumption and maintaining high thermal comfort. The designed method was based on deep deterministic policy gradient (DDPG) and the training environment was developed based on long-short-term-memory (LSTM) networks, which can approximate real-world HVAC operations [21]. In [20], Valladares et al. proposed a Double DQN (DDQN) based method to control air conditioning units and ventilation fans for optimizing energy consumption, thermal comfort, and indoor air quality jointly. Real-world experiments showed that the proposed method can save energy by 4%–5% without sacrificing other performances. In [24], Deng et al. proposed a context-aware soft-actor–critic (SAC) based method for multi-zone HVAC control, which can encode the past experiences into context states to deal with non-stationary building environments. Simulation results showed that incorporating context in system states contributes to the reduction of energy consumption by 8.6% without sacrificing thermal comfort. Similarly, Deng et al. proposed a non-stationary DQN based method for HVAC control, which can save energy by 13% [19].

Although some advances have been made in above-mentioned works, there are still some limitations, i.e., HVAC systems cannot provide personalized thermal environment for each occupant in a shared office space and they tend to cool/heat the entire office space even the space is partially occupied.

2.2. PCS management in office spaces

To overcome the mentioned-above drawbacks of HVAC systems, many studies have considered the use of PCSs (e.g., heaters and fans), which can adjust local thermal environments around occupants and improve occupant satisfaction. For example, Yu et al. investigated the impacts of four personalized seat heating systems in a typical office room in Shanghai during winter on energy consumption and thermal comfort [25]. Experimental results showed that overall personalized heating can provide a satisfactory thermal environment for all body parts with much less energy consumption than room air conditioners. In [9], He et al. surveyed the impacts of fan-use rates on thermal comfort, energy conservation, and human productivity. Since existing studies mainly focused on the case that PCSs are manually controlled by occupants, some control burden will be incurred. To remove the control burden for occupants, Aryal et al. proposed a framework to learn occupant preferences and control PCS devices under different levels of automation [10]. Experimental results showed that occupant satisfaction could be improved using PCSs and inquisitive automation can lead to the highest occupant satisfaction with the thermal environment. In [26], Wang et al. investigated the impact of a radiant leg warmer on the thermal comfort of office workers in winter. Experimental results showed that the radiant leg warmer can effectively improve their thermal comfort and related physiological parameters (e.g., heart rate, and blood pressure) in a cold environment. In [27], Yang et al. showed that the use of wearable cooling fans can improve the thermal comfort of office workers in warm indoor environments. More related works can be found in [28].

Although PCSs have some advantages in improving individual thermal satisfaction, they have limited corrective power. Typically, the cooling corrective power ranges from −1 to −6K [14], which means that PCSs may need to coordinate with HVAC systems, which exist widely in the current shared office spaces.

2.3. Operations of HVAC systems and PCSs in office spaces

There have been some works on the operations of HVAC systems and PCSs simultaneously in office spaces. For example, Xu et al. proposed a Lagrangian relaxation-based approach to optimize the coordination of an air conditioning system and personal fans [11]. Experimental and simulation results showed that the proposed approach can enhance building demand response capability, save energy cost, and improve individual thermal comfort environments. In [12], Kalaimani et al. proposed a PCS-aware HVAC control method based on model predictive control. Simulation results showed that the proposed method can reduce energy and improve comfort for both summer and winter when heterogeneous comfort requirements are considered. In [13], He et al. investigated the problem that how occupants use desk fans and control HVAC simultaneously. Experimental results showed that the always-on strategy of desk fans made both sensitive and insensitive subjects choose the higher temperature set-point, resulting in energy conservation. In [29], He et al. conducted an experimental study to show that the thermal needs of two occupants in a shared office space can be satisfied by an adjustable thermostat and local heating in winter. Although some positive results have been done, the above-mentioned works did not consider the optimal coordination of PCSs and an HVAC system in a shared office space under the premise that an explicit building thermal dynamics model and accurate forecast information about outdoor temperature and space occupancy are unavailable.

In summary, existing works have made great progress on improving thermal comfort in office buildings. However, no work has been developed to investigate an optimal coordination problem among an HVAC system and multiple PCSs for optimizing energy consumption and personalized thermal comfort in a shared office space without knowing explicit building thermal dynamics models and accurate forecast information about uncertain parameters. To fill the above-mentioned research gap, we propose an energy efficient personalized thermal comfort control algorithm based on attention-based multi-agent DRL, which can reduce energy consumption and improve personalized thermal comfort simultaneously compared with existing methods.

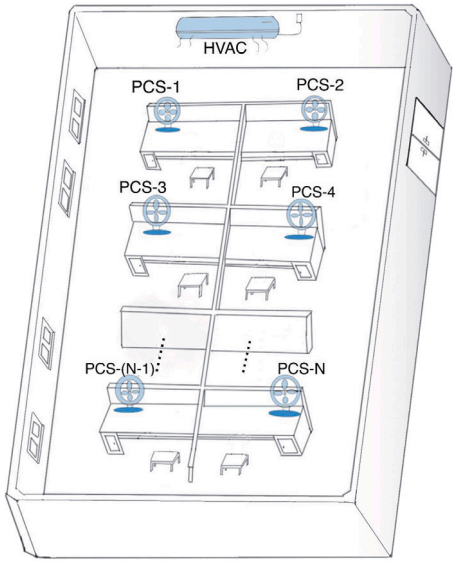


Fig. 1. Thermal comfort systems in a shared office space.

3. System model and problem formulation

We consider a shared office space as shown in Fig. 1, where an HVAC system and N PCSs can be identified. Without loss of generality, we mainly consider the cooling mode in summer and PCSs denote desktop fans. To be specific, the HVAC system changes the indoor temperature in the whole space by adjusting the temperature set-point, while each desktop fan serves one occupant by changing the local micro-environment, i.e., air speed. Since the thermal comfort of each occupant is related to both indoor temperature and air speed, it is necessary to coordinate the operations of the HVAC system and desktop fans. Specifically, in each time slot t, the temperature set-point of the HVAC system and fan speeds are jointly determined based on some factors, e.g., indoor temperature, outdoor temperature, and occupant numbers. In the following parts, we first introduce HVAC operation model, fan operation model, thermal comfort model, and energy consumption model. Then, we formulate an expected energy consumption minimization problem as well as its variant.

3.1. HVAC operation model

We consider a direct expansion HVAC system with no inverter [30], which consumes a fixed amount of power via a fixed compressor speed during the ON status. Similar to [31], its operational mode σ_t can be described as follows:

$$\sigma_t = \begin{cases} 1, & \text{if } T_{\text{in},t} > T_{\text{set},t} + b \\ 0, & \text{if } T_{\text{in},t} < T_{\text{set},t} \\ \sigma_{t-1}, & \text{otherwise} \end{cases} \tag{1}$$

In (1), $\sigma_t = 1$ and $\sigma_t = 0$ denote that the HVAC system is turned on and turned off at slot t, respectively; $T_{\text{in},t}$ and $T_{\text{set},t}$ denote indoor temperature and HVAC set-point at slot t, respectively; b denotes the temperature dead-band, which can help to avoid short cycles, i.e., the operational mode of the HVAC system changes frequently.

Let $T_{\text{set}}^{\min}$ and $T_{\text{set}}^{\max}$ be the minimum and maximum set-point under on status, respectively. Moreover, $T_{\text{set}}^{\text{off}}$ denotes that HVAC system is turned

off. Then, the setpoint $T_{\text{set},t}$ should be selected from the following discrete set:

$$T_{\text{set},t} \in \{T_{\text{set}}^{\text{off}},\ T_{\text{set}}^{\min},\ T_{\text{set}}^{\min}+1, \dots,\ T_{\text{set}}^{\max}\} \tag{2}$$

Let N_t be the total number of occupants in the office space at slot t. When there is no person in the shared office space, the HVAC system should be turned off, i.e., the following constraint is considered:

$$T_{\text{set},t} = T_{\text{set}}^{\text{off}}\ if\ N_t = 0 \tag{3}$$

3.2. PCS operation model

Each desktop fan can provide personalized thermal comfort for an occupant by adjusting its local air speed. Suppose that fan speed has M levels and $v_i^{\max}$ is the maximum fan speed of ith desktop fan. Then, fan speed should satisfy the constraint as follows:

$$v_{i,t} \in \{0, \frac{v_i^{\max}}{M}, \frac{2v_i^{\max}}{M}, \dots, v_i^{\max}\} \tag{4}$$

When occupant i is not in the space, desktop fan i will be turned off automatically, which is described as follows:

$$0 \leqslant v_{i,t} \leqslant H_{i,t} v_i^{\max} \tag{5}$$

In (5), $H_{i,t}$ is a binary variable that represents the status of occupant i. To be specific, $H_{i,t} = 1$ denotes occupant i is in the space. Otherwise, $H_{i,t} = 0$ denotes occupant i is not in the space.

3.3. Thermal comfort model

The thermal comfort of an occupant depends on many factors, e.g., indoor temperature, humidity, air speed, clothing level, and mean radiant temperature. Since we mainly focus on providing a comfortable thermal environment for occupants by jointly scheduling the HVAC system and desktop fans, how to develop an accurate personalized thermal comfort model is left for future work [32,33]. Since the proposed algorithm in Section 4 can be applicable to any thermal comfort model, we adopt a simplified PMV model $\mathcal{F}_i$ to represent thermal comfort extent of occupant i [12] as follows:

$$\beta_{i,t} = \mathcal{F}_i(T_{\text{in,t}}, v_{i,t}) \tag{6}$$

In addition, $\beta_{i,t}$ should be kept within the preferred range of occupant i. Let $\beta_i^{\min}$ and $\beta_i^{\max}$ be the lower limit and upper limit of the acceptable range of occupant i, respectively. Then, the following constraint should be satisfied:

$$\beta_i^{\min} \leqslant \beta_{i,t} \leqslant \beta_i^{\max} \tag{7}$$

Since indoor temperature has a large impact on thermal comfort, indoor temperature $T_{\text{in},t}$ should be kept within an acceptable range when the space is occupied, which can be described as follows:

$$(T_{\text{in}}^{\min} - Z^{\min}) \cdot I_{N_t>0} + Z^{\min} \leqslant T_{\text{in},t} \leqslant (T_{\text{in}}^{\max} - Z^{\max}) \cdot I_{N_t>0} + Z^{\max} \tag{8}$$

In (8), $[T_{\text{in}}^{\min}, T_{\text{in}}^{\max}]$ denotes the acceptable indoor temperature range when the space is occupied, while $[Z^{\min}, Z^{\max}]$ denotes the lower limit and upper limit of indoor temperature when the space is not occupied; $I_{N_t>0} = 1$ if $N_t > 0$ and $I_{N_t>0} = 0$ for other cases.

Let $T_{\text{out},t}$ be the outdoor temperature at slot t and ρ_t denotes unknown thermal disturbance at slot t. Then, indoor temperature dynamics can be captured by the following constraints [34]:

$$T_{\text{in},t+1} = \mathcal{G}(T_{\text{in},t}, T_{\text{out},t}, \sigma_t, \rho_t) \tag{9}$$

In (9), we assume that the expression of $\mathcal{G}$ is inexplicit, since it is difficult to develop an explicit thermal dynamics model that is accurate and efficient enough for HVAC control [35].

3.4. Energy consumption model

The energy consumption consists of two parts, i.e., energy consumption related to an HVAC system and energy consumption related to PCS systems. Let ΔT be the duration of a time slot. Let P_{hvac} be the rated power of the HVAC system. Then, the energy consumption of an HVAC system can be computed as follows:

$$C_t^{\text{hvac}} = \Delta T P_{\text{hvac}} \sigma_t \tag{10}$$

Let $\omega_{\text{pcs},i}$ be a conversion parameter that transforms fan speed into power consumption, the energy consumption of the ith desktop fan at slot t can be calculated as follows:

$$C_{i,t}^{\text{pcs}} = \Delta T \omega_{\text{pcs},i} v_{i,t} \tag{11}$$

3.5. Expected energy consumption minimization problem

Based on the above models, we can formulate an expected energy consumption minimization problem under uncertainties related to the coordinated operations of PCSs and an HVAC system as follows:

$$(\mathbf{P1})\ \min_{T_{\text{set},t}, v_{i,t}} \sum_{t=0}^{L} \mathbb{E}\{C_t^{\text{hvac}} + \sum_{i=1}^{N} C_{i,t}^{\text{pcs}}\} \quad s.t.\ (1)\text{–}(9) \tag{12}$$

In (12), L is the time horizon considered in this paper and $\mathbb{E}$ denotes the expectation operator, which acts on uncertain parameters, e.g., N_t, $T_{\text{out},t}$, and ρ_t. Although the above problem can be solved by stochastic programming and model predictive control [36], some prior knowledge should be required, e.g., probability distribution or prediction values. Typically, Lyapunov optimization techniques [17] can be used to design real-time algorithms for a decision-making problem under uncertainty. However, it is not applicable when an explicit building thermal dynamics model $\mathcal{G}$ is unavailable. Although DRL-based methods in existing works [18] can operate without knowing an explicit building thermal dynamics model and prior knowledge of uncertain parameters, they cannot apply to **P1** directly since the size of discrete action space $(T_{\text{set}}^{\max} - T_{\text{set}}^{\min} + 1)M^N$ increases rapidly with the increase of PCS numbers. As a result, inefficient learning will be incurred.

3.6. Problem reformulation

To solve **P1** efficiently, we first reformulate the problem **P1** as a Markov game with $N + 1$ heterogeneous agents (i.e., the agents have different types of tasks). Then, we propose a AMADRL-based algorithm to solve the Markov game in next section. Note that a Markov game is used to describe the competition and/or cooperation relationship among multiple agents and it can be represented by a tuple $(\mathcal{S}, \mathcal{A}_1, \dots, \mathcal{A}_{N+1}, \mathcal{T}, \mathcal{R}_1, \dots, \mathcal{R}_{N+1}, \gamma)$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}_i$ denotes the action space of agent i, $\mathcal{T}$ denotes the state transition function, $\mathcal{R}_i$ denotes the reward function of agent i, γ denotes the discount factor. In time slot t, agent i takes an action $a_{i,t} \in \mathcal{A}_i$ based on its local observation $o_{i,t}$, which contains partial information of the global state s_t. After all actions are taken, the environment returns rewards and new state s_{t+1}. In this paper, agents denote learners and decision-makers (i.e., HVAC agent and PCS agents), the environment consists of many objects outside agents (e.g., occupants, HVAC, PCSs, indoor/outdoor temperature, and time). The purpose of agent i is to find an optimal policy, which can maximize the sum of discounted rewards received over the future, i.e., $\sum_{j=0}^{\infty} \gamma^j r_{i,t+j+1}(s_t, a_{1,t}, \dots, a_{N+1,t})$. Since AMADRL-based algorithm proposed in next section is model-free, the prior knowledge of $\mathcal{T}$ is not required. In the following parts, three components of Markov game associated with **P1** are defined, i.e., state, action, and reward.

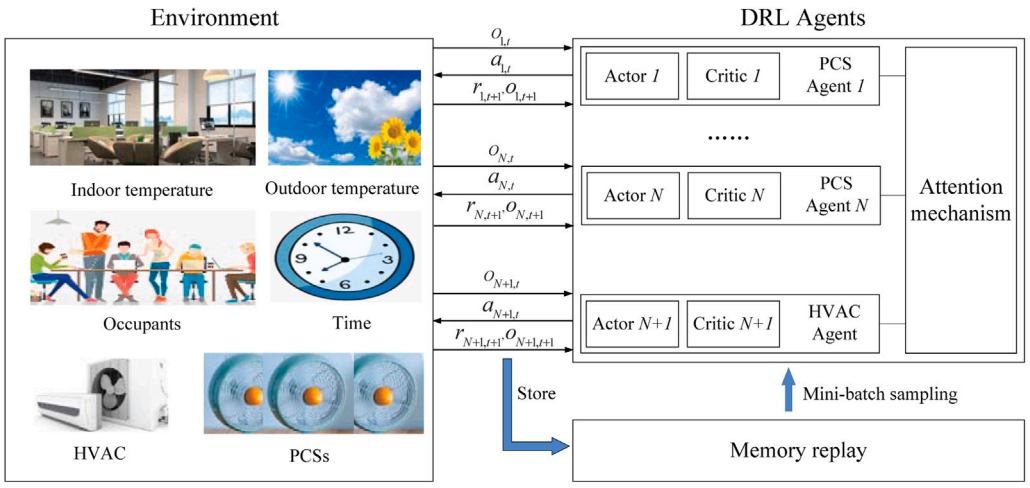


Fig. 2. The proposed control algorithm.

3.6.1. HVAC agent

For an HVAC agent, it intends to minimize HVAC energy consumption while maintaining a comfortable thermal environment for all occupants by adjusting temperature set-point. Here, a comfortable thermal environment is represented by constraints (7)–(9), while energy consumption is related to constraints (1) and (10). Therefore, the state, action, and reward of an HVAC agent are designed as follows, i.e., $o_{N+1,t} = s_{\text{hvac},t} = (t, T_{\text{in},t}, T_{\text{out},t}, H_{1,t}, \ldots, H_{N,t}, \sigma_{t-1})$, $a_{N+1,t} = a_{\text{hvac},t} = T_{\text{set},t}$, and $r_{N+1,t} = r_{\text{hvac},t} = -(\alpha C_t^{\text{hvac}} + \phi \sum_{i=1}^{N}([\beta_i^{\min} - \beta_{i,t}]^+ + [\beta_{i,t} - \beta_i^{\max}]^+)H_{i,t} + u_t)$, where α and ϕ represent the importance of energy consumption and thermal comfort deviation with respect to indoor temperature deviation, respectively; $u_t = [T_{\text{in},t+1} - ((T_{\text{in}}^{\max} - Z^{\max})I_{N_{t+1}>0} + Z^{\max})]^+ + [((T_{\text{in}}^{\min} - Z^{\min})I_{N_{t+1}>0} + Z^{\min}) - T_{\text{in},t+1}]^+$, where $I_{N_{t+1}>0}$ denotes that whether the space is occupied in next time slot or not. Since occupancy state is unlikely to change frequently within consecutive time slots, $I_{N_{t+1}>0}$ can be approximated by $I_{N_t>0}$. Although the approximation is adopted here, it does not mean that the proposed algorithm in next section needs to predict uncertain parameters. The reason is that the proposed control algorithm makes decision just based on the local observation as shown in Algorithm 2. In addition, to stabilize the learning process of HVAC agent, the size of $o_{N+1,t}$ is reduced as follows, i.e., $o_{N+1,t} = s_{\text{hvac},t} = (t, T_{\text{in},t}, T_{\text{out},t}, N_t, \sigma_{t-1})$, where $N_t = \sum_{i=1}^{N} H_{i,t}$.

3.6.2. PCS agent

For each PCS agent i, it intends to minimize fan energy consumption while providing a comfortable thermal environment for occupant i. Here, energy consumption is related to (11), while comfortable thermal environment is represented by constraints (5)–(7). Thus, the state, action, and reward of each PCS agent is designed by $o_{i,t} = s_{\text{pcs},i,t} = (t, T_{\text{in},t}, H_{i,t})$, $a_{i,t} = a_{\text{pcs},i,t} = v_{i,t}$, and $r_{i,t} = r_{\text{pcs},i,t} = -(\alpha C_{i,t}^{\text{pcs}} + \phi([\beta_i^{\min} - \beta_{i,t}]^+ + [\beta_{i,t} - \beta_i^{\max}]^+)H_{i,t})$.

4. Energy-efficient thermal comfort control algorithm design

In this section, we propose an energy-efficient thermal comfort control algorithm to solve the above-mentioned Markov game with heterogeneous agents based on AMADRL. To be specific, the framework of the proposed control algorithm is illustrated in Fig. 2, where two kinds of DRL agents interact with the environment (including outdoor temperature, occupancy state, time, HVAC system, and PCSs) and using the stored interaction information for algorithm training. To implement efficient training among PCS agents and HVAC agent, attention mechanism is adopted, which can help the current agent to know the contributions from other agents when computing its own action-value function. Since Multi-Attention-Actor–Critic (MAAC) has strong scalability than many other multi-agent DRL algorithms and supports discrete action space [37], it is adopted in this paper to train DRL agents.

In Fig. 2, each agent has an actor network and a critic network. The former takes action based on the local observation, while the latter evaluates the value of taken action at the given state. When computing the action-value function $Q_i^{\psi}(o, a)$ for agent i (where ψ is the weight parameter of critic network, $o = (o_1, o_2, \ldots, o_{N+1})$, $a = (a_1, a_2, \ldots, a_{N+1})$), the contributions from other agents should be considered and captured by attention mechanism. To be specific, attention module can generate the contributions from other agents x_i for the current agent i as follows:

$$x_i = \sum_{j \neq i} \kappa_j h(W_v e_j) \tag{13}$$

In (13), $e_j = q_j(o_j, a_j)$ and q_j be a one-layer MLP embedding function; W_v is a shared matrix that transforms e_j into a "value"; h is a non-linear activation function; W_k and W_q are shared matrixes that transform e_j into a "key" and transform e_i into a "query", respectively; κ_j is the attention weight associated with agent j and can be obtained as follows:

$$\kappa_j = \exp^{((W_k e_j)^T W_q e_i)} \Big/ \sum_{j=1}^{N} \exp^{((W_k e_j)^T W_q e_i)} \tag{14}$$

After x_i is obtained, $Q_i^{\psi}(o, a)$ can be calculated as follows:

$$Q_i^{\psi}(o, a) = f_i(z_i(o_i), x_i) \tag{15}$$

In (15), f_i is a two-layer multi-layer perceptron (MLP), z_i is a one-layer MLP embedding function. Note that parameters W_k, W_q, W_v in the attention mechanism are shared by all agents, and all critics are updated together by minimizing a joint regression loss function as follows:

$$\mathcal{L}_Q(\psi) = \sum_{i=1}^{N+1} \mathbb{E}_{(o,a,\tilde{o},r)\sim\mathcal{D}}[(Q_i^{\psi}(o, a) - y_i)^2] \tag{16}$$

In (16), $(o, a, \tilde{o}, r)$ represents a tuple in replay buffer $\mathcal{D}$, φ is the temperature parameter in soft actor–critic and it determines the balance between maximizing entropy and reward, $\bar{\theta}$ is the weight parameter of target actor network; y_i is given as follows:

$$y_i = r_i(o, a) + \gamma \mathbb{E}_{\tilde{a}\in\pi_{\bar{\theta}}(\tilde{o})}[-\varphi \log(\pi_{\bar{\theta}_i}(\tilde{a}_i|\tilde{o}_i)) + Q_i^{\bar{\psi}}(\tilde{o}, \tilde{a})] \tag{17}$$

Next, the weight parameter of actor network can be updated by policy gradient methods. To be specific, the gradient is given as follows:

$$\nabla_{\theta_i} J(\theta) = \mathbb{E}_{o\sim\mathcal{D}, a\sim\pi}[\nabla_{\theta_i} \log(\pi_{\theta_i}(a_i|o_i)) o_i(o_i, a_i)] \tag{18}$$

In (18), $\varrho_i(o_i, a_i) = -\varphi \log(\pi_{\theta_i}(a_i|o_i)) + Q_i^{\psi}(o, a) - d(o, a_{\setminus i})$, $\setminus i$ denotes the set of agents except i. Here, $Q_i^{\psi}(o, a) - d(o, a_{\setminus i})$ is called as the multi-agent advantage function, which can show that whether the current action will lead to an increase in expected return, where $d(o, a_{\setminus i}) = \sum_{\tilde{a}_i \in A_i} \pi_{\theta_i}(\tilde{a}_i|o_i) Q_i^{\psi}(o, (\tilde{a}_i, a_{\setminus i}))$.

The details of the algorithm for training DRL agents are described by Algorithm 1. In lines 1–2, input and output are defined, respectively. In lines 3–5, experience replay buffer, actor network, critic network, target actor network, and target critic network are initialized. Note that all networks are represented by deep neural networks, which is composed of an input layer, multiple hidden layers, and one output layer. In each episode k, the environment is initialized and each agent observes its local state $o_{i,1}$ as shown in line 7. In each time slot t, each agent takes an action $a_{i,t}$. Next, the joint action a_t is imposed on the environment, i.e., adjusting the HVAC temperature set-point and fan speeds. Then, each agent i observes new state $o_{i,t+1}$ and receives reward $r_{i,t+1}$ as shown in line 10. Moreover, system transitions $(o_t, a_t, o_{t+1}, r_{t+1})$ are stored in the memory replay buffer. When there are enough transitions, algorithm training is conducted as shown in lines 13–19. To be specific, B_{size} transitions are sampled randomly from the buffer. Then, the sampled data is used for updating the critic network as shown in line 18. Next, the actor network is updated by maximizing the policy gradient as shown in line 19. Note that the expectation operator in (16) and (18) is approximated by the average value. Finally, target network parameters are updated as shown in line 21.

Note that the proposed control algorithm adopts the way of centralized training and decentralized execution. In other words, when updating the critic network of one agent, the observation and action information of other PCS agents will be used. Moreover, once the process of training DRL agents is finished, just actor networks are needed for local decisions during the testing period as shown in Algorithm 2. Since just forward propagation of deep neural networks is involved in the practical testing, the proposed control algorithm has low computational complexity.

Algorithm 1 MAAC-based Algorithm for training PCS and HVAC agents

1: **Input**: Outdoor temperature and space occupation traces
2: **Output**: The weight parameter of actor network θ
3: Initialize experience replay buffer D
4: Initialize weight parameters of actor network π_i^{θ} and critic network Q_i^{ψ}
5: Initialize weight parameters of target actor network $\pi_i^{\bar{\theta}}$ and target critic network $Q_i^{\bar{\psi}}$
6: **for** k=0, 1, ⋯, $K-1$ **do**
7: Initialize environment and get initial observation state $o_{i,1}$ for agent i
8: **for** t=1, 2, ⋯, F_{train} **do**
9: Each agent i selects action $a_{i,t} \sim \pi_i^{\theta}(\cdot|o_{i,t})$ and execute it under the environment
10: Each agent i observes new state $o_{i,t+1}$ and receives reward $r_{i,t+1}$
11: Store transitions $(o_t, a_t, o_{t+1}, r_{t+1})$ in the experience replay buffer
12: **if** $\chi \geq (N+1)B_{\text{size}}$ and mod$(kF+t, T_{\text{update}})$=0 **then**
13: Sample B_{size} transitions $(o, a, \tilde{o}, r)$ from the experience replay buffer randomly
14: Calculate $Q_i^{\psi}(o_i^g, a_i^g)$ for all i and g $(1 \leq g \leq B_{\text{size}})$
15: Calculate $\tilde{a}_i^g \sim \pi_i^{\bar{\theta}}(\tilde{o}_i^g)$ for all i and g
16: Calculate $Q_i^{\bar{\psi}}(\tilde{o}_i^g, \tilde{a}_i^g)$ for all i and g
17: Update critic network by minimizing the joint regressive loss function (16)
18: Calculate $Q_i^{\psi}(o_i^g, a_i^g)$ for all g and i
19: Update policies using (18)
20: Update weight parameters of target actor network and target critic network:
21: $\bar{\psi} \leftarrow \xi\psi + (1-\xi)\bar{\psi}$, $\bar{\theta} \leftarrow \xi\theta + (1-\xi)\bar{\theta}$
22: **end if**
23: **end for**
24: **end for**

Algorithm 2 Energy-efficient Thermal Comfort Control Algorithm

1: **Input**: Load the weight parameters of actor network π_i^{θ} for each PCS agent and HVAC agent
2: **Output**: Fan speeds of PCSs and temperature set-point at time slot t
3: Initialize environment and get initial observation state $o_{i,1}$ for agent i
4: **for** t=1, 2, ⋯, F_{test} **do**
5: Each agent i $(1 \leq i \leq N+1)$ selects action $a_{i,t} \sim \pi_i^{\theta}(\cdot|o_{i,t})$ in parallel
6: All agents execute actions under the environment
7: Each agent i observes new state $o_{i,t+1}$
8: **end for**
9: Performance metrics (e.g., total energy consumption) are collected

5. Performance evaluation

In this section, we first introduce the simulation setup. Next, we define performance metrics. Then, three baselines for performance comparisons are described. Finally, we provide simulation results and discussions.

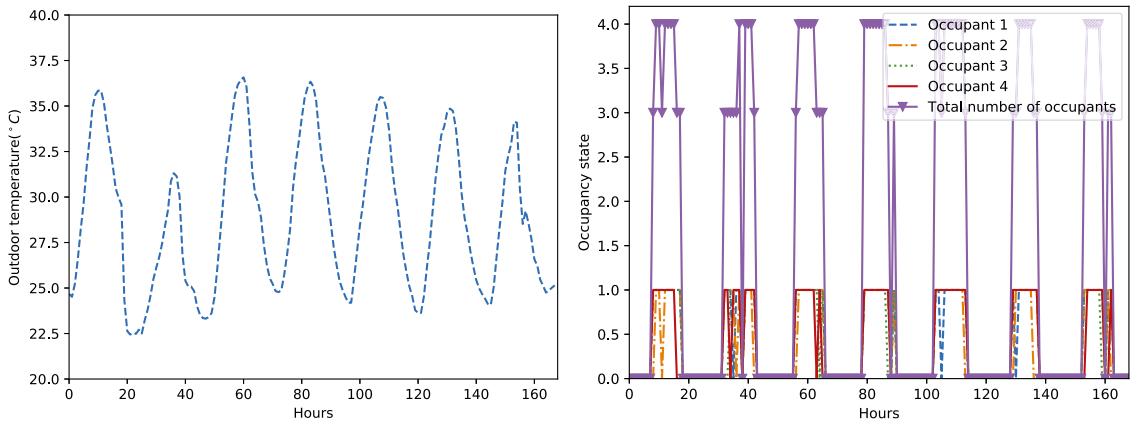
5.1. Simulation setup

In simulations, real-world traces related to outdoor temperature and occupancy are used. Since the cooling mode of an HVAC system in the summer and desk fans are considered, we use the hourly outdoor temperature data from June 1 to August 31, 2018, in Pecan Street database,[1] which is the largest real-world open energy database on the planet and includes the data related to the Mueller neighborhood in Austin, Texas, USA. Moreover, occupancy data from [38] are used in simulations. Since the duration of a time slot is 15 min, outdoor temperature and occupancy state are assumed to be constant within an hour for simplicity. Moreover, we consider four occupants in a shared office space and occupancy information obtained from [38] is used for randomly generating occupancy state of each occupant. Note that the traces of hourly outdoor temperature and occupancy state are shown in Fig. 3. The program code of this paper is implemented using Python 3.7 and executed by a desktop computer with Intel Core(TM) i9-9900 CPU and 64 GB RAM. To simulate the building thermal dynamics, we adopt the following model $\mathcal{G}$ similar to many existing works [39–41], i.e., $T_{\text{in},t+1} = \mathcal{G}(T_{\text{in},t}, T_{\text{out},t}, \sigma_t, \rho_t) = \varepsilon_{\text{hvac}} T_{\text{in},t} + (1 - \varepsilon_{\text{hvac}})(T_{\text{out},t} - P_{\text{hvac}} \sigma_t \eta_{\text{hvac}}/\zeta) + \rho_t$, where $\varepsilon_{\text{hvac}} = 0.7$, $\eta_{\text{hvac}} = 2.5$, and $\zeta = 0.14$. Moreover, we evaluate the impact of random thermal disturbance ρ_t in simulations to illustrate the robustness of the proposed algorithm against model uncertainty. Note that the above model structure is not used for optimization energy consumption similar to model-based methods (e.g., model predictive control, and Lyapunov optimization techniques), but used to obtain environment data for model-free learning. The main simulation parameters are listed as follows, i.e., $K = 6000$, $F_{\text{train}} = 96$, $F_{\text{test}} = 2976$, $T_{\text{set}}^{\min} = 22$ °C, $T_{\text{set}}^{\max} = 28$ °C, $T_{\text{set}}^{\text{off}}$=0, $M = 2$, $v_i^{\max} = 1$ m/s, $Z^{\min} = 0$ °C, $Z^{\max} = 30$ °C, $\Delta T = 15$ min, $P_{\text{hvac}} = 2$ kW, $\omega_{\text{pcs}} = 0.03$ kW[12]. Similar to [12], we assume that $\beta_{i,t} = F_i(T_{\text{in},t}, v_{i,t}) = \varsigma_{1,i} T_{\text{in},t} + 0.76 v_{i,t}^2 - 2.14 v_{i,t} - \varsigma_{2,i}$, where $\varsigma_{1,i} \in \{0.37, 0.35, 0.33, 0.36\}$ and $\varsigma_{2,i} \in \{9.22, 8.5, 9.5, 9.1\}$. The heterogeneous thermal comfort requirements of occupants are configured as follows, i.e., $-0.7 \leq \beta_{1,t} \leq 0$[12], $-0.5 \leq \beta_{2,t} \leq 0.1$, $-1 \leq \beta_{3,t} \leq 0.2$, and $-0.3 \leq \beta_{4,t} \leq 0.3$. $\alpha = 10$, $\phi = 2$. In each actor network, three layers are considered and the number of neurons in the hidden layer is 128, the buffer size is 240000, $B_{\text{size}} = 120$, $\gamma = 0.995$, $\varphi = 0.1$, $\xi = 0.001$, the learning rate of actor network and critic network is 0.0001 and 0.001, respectively.

5.2. Performance metrics

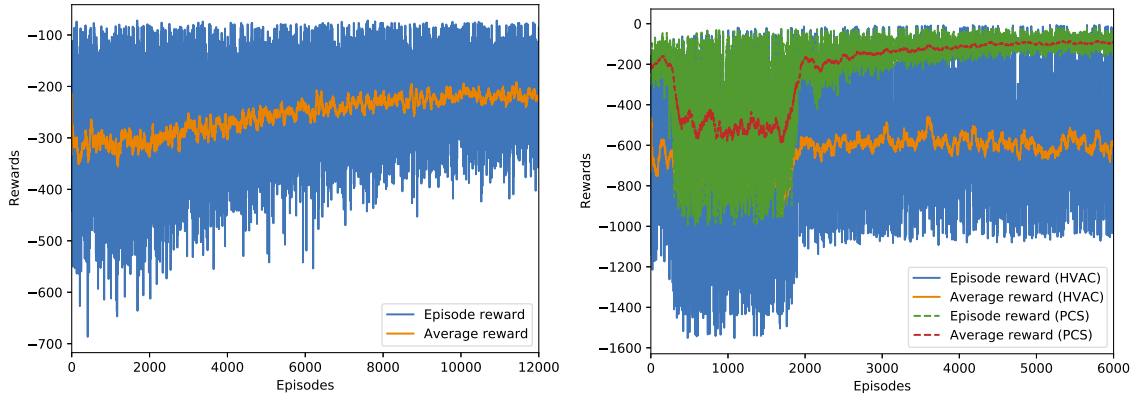
To describe testing performances of different schemes, we adopt total energy consumption (TEC, in kWh) and average thermal comfort deviation (ACD) as performance metrics, which are defined as follows: $\text{TEC} = \sum_{t=0}^{F_{\text{test}}} (C_t^{\text{hvac}} + \sum_{i=1}^{N} C_{i,t}^{\text{pcs}})$, $\text{ACD} = \frac{1}{F_{\text{test}} N} \sum_{t=0}^{F_{\text{test}}} (\sum_{i=1}^{N} ([\beta_i^{\min} - \beta_{i,t}]^+ + [\beta_{i,t} - \beta_i^{\max}]^+) H_{i,t})$.

[1] https://www.pecanstreet.org/



(a) Outdoor temperature (b) Occupancy state

Fig. 3. Real-world traces used in simulations.



(a) B2 (b) The proposed algorithm

Fig. 4. Convergence curves under B2 and the proposed algorithm.

5.3. Benchmarks

To evaluate whether the way of adjusting HVAC temperature set-point and fan speeds of PCSs dynamically and cooperatively can lead to better performance or not, three baselines are adopted for performance comparisons. To be specific, baseline1 (B1) adopts a fixed HVAC temperature set-point strategy and does not consider the use of PCSs. Since different fixed set-points would result in different tradeoffs between TEC and ACD, we mainly consider the fixed HVAC temperature set-point that can lead to the lowest ACD. Baseline2 (B2) adopts Double Deep Q-network (DDQN) based DRL algorithm [21] to adjust the HVAC temperature set-point dynamically and does not consider the use of PCSs[2]. Moreover, the state, action, and reward under B2 are the same as those of the HVAC agent under the proposed algorithm. Baseline3 (B3) adopts a fixed HVAC temperature set-point strategy that is the same as B1. Moreover, B3 considers the use of PCSs. For each PCS, it operates at the maximum/minimum fan speed when the corresponding occupant is present/not present in the shared office space.

5.4. Algorithmic convergence performance

Since B2 and the proposed algorithm are DRL-based schemes, it is necessary to check their convergence performances, which are shown in Fig. 4. It can be seen that the curve of the episode reward (i.e., the total reward received within an episode) under B2 fluctuates frequently, which is caused by action exploration and uncertain system parameters. To show the tendency of episode reward curve under B2 more clearly, we provide the average rewards over the past 50 episodes. It can be observed that the average reward curves generally increase and become more and more stable. The same tendency can be found in the reward curves associated with HVAC and PCS (note that the total reward curve of all PCSs is provided for simplicity), which indicates that the convergence of the proposed algorithm.

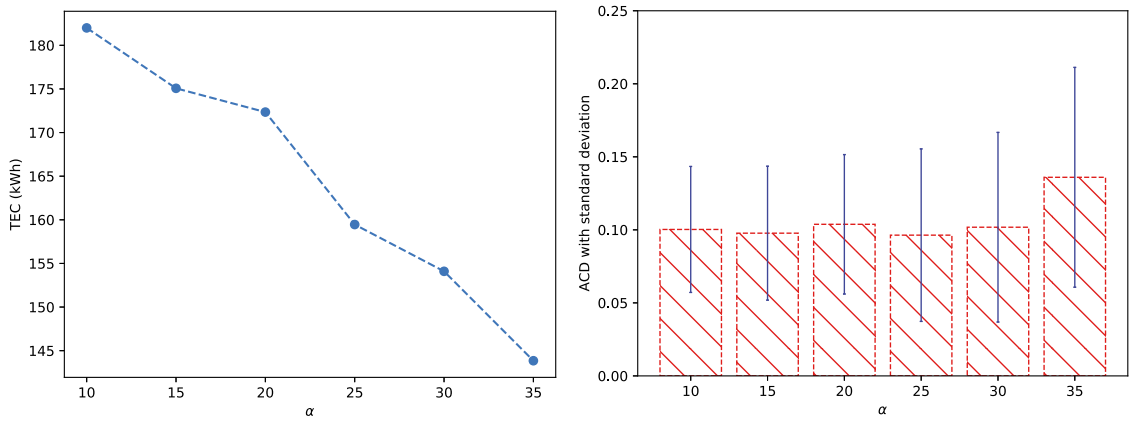
5.5. Algorithmic flexibility

By adjusting the value of α, the proposed algorithm can achieve different tradeoff performances. As shown in Fig. 5, the total energy consumption becomes lower and ACD generally becomes higher with the increase of α. The reason is that α represents the importance of energy consumption relative to indoor temperature deviation. Moreover, the increase of α means that the importance of thermal comfort deviation ϕ is indirectly reduced, resulting in higher ACD.

5.6. Algorithmic effectiveness

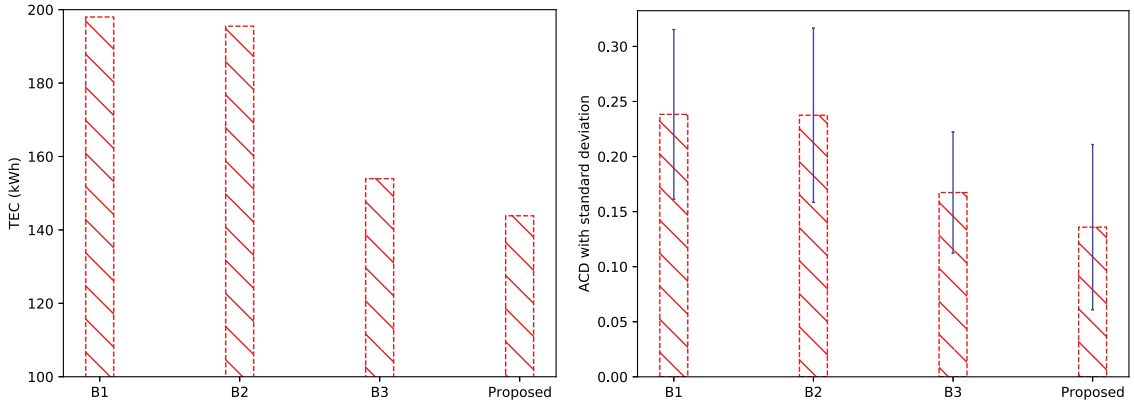
Performance comparisons of all schemes are shown in Fig. 6, where the schemes of B1 and B3 with the lowest ACD are selected. In Fig. 6, it can be seen that the proposed algorithm can achieve the best performance among all schemes. To be specific, compared with B1, B2, and B3, the proposed algorithm with $\alpha = 35$ can reduce total energy

[2] Note that the size of action space will be 648 (i.e., 3*3*3*3*8) if all control decisions are made by a single DDQN-based DRL agent. Since DDQN is inefficient when action space size is large, the coordination among an HVAC system and PCSs is not considered under B2.



(a) TEC (b) ACD with standard deviation

Fig. 5. Tradeoff performance of the proposed algorithm.



(a) TEC (b) ACD with standard deviation

Fig. 6. Performance comparisons of all schemes.

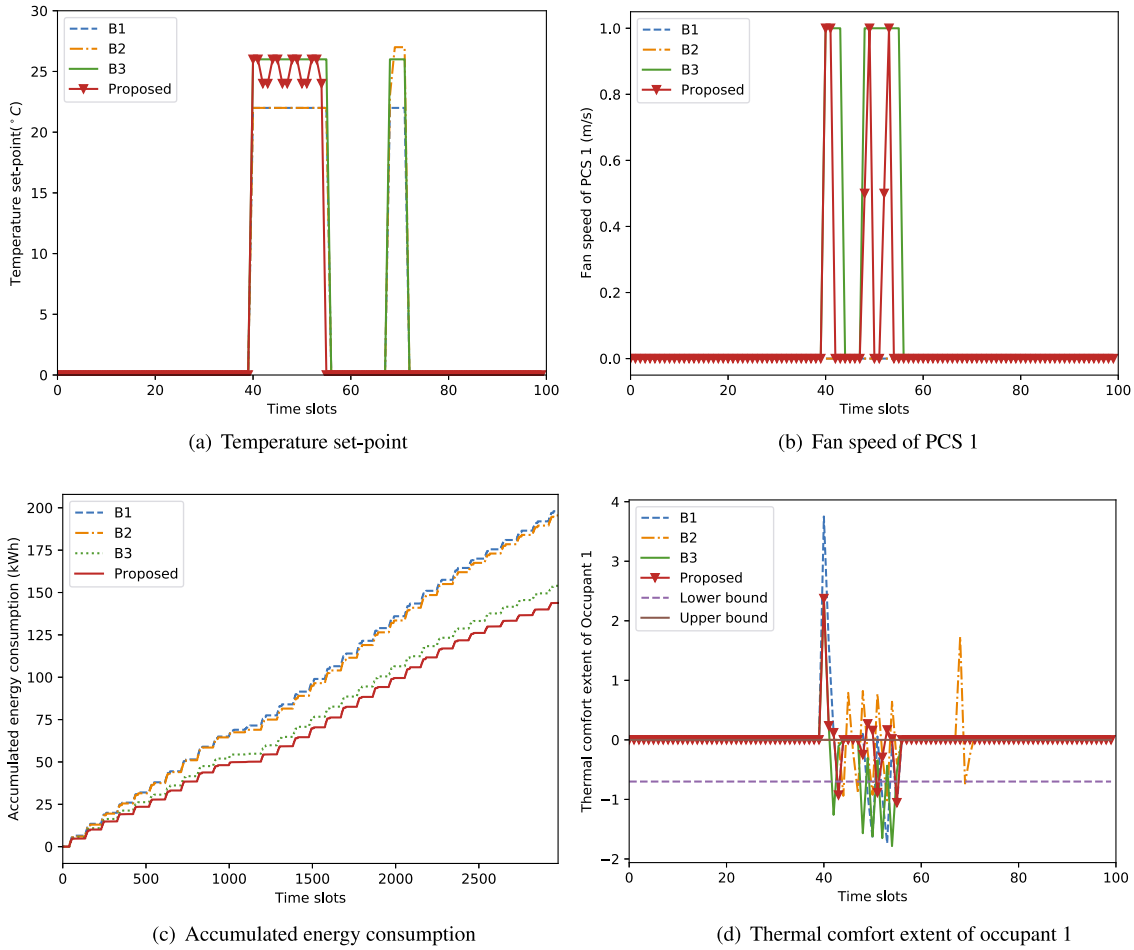
consumption by 27.35%, 26.42%, and 6.56%, respectively. Moreover, compared with B1, B2, and B3, the proposed algorithm with $\alpha = 35$ can reduce average thermal comfort deviation by 42.97%, 42.77%, and 18.77%, respectively. To explain the reason why the proposed algorithm achieves the best performance, more detail results are presented in Fig. 7, where just the fan speed of PCS 1 and thermal comfort extent of occupant 1 are provided for ease of presentation. In Figs. 7(a) and (b), it can be seen that the proposed algorithm sometimes turns off the HVAC system and reduces fan speeds for saving energy consumption, i.e., $T_{\text{set},t} = 0$ and $v_{1,t} = 0$. In contrast, B3 adopts the fixed temperature set-point and the fixed fan speed during office hours, resulting in higher energy consumption and lower thermal comfort extent as shown in Figs. 7(c) and (d). Moreover, the proposed algorithm can achieve the best thermal comfort during office hours as shown in Fig. 7(d) since the amplitude and frequency of violating the thermal comfort range of occupant 1 (i.e., [-0.7,0]) are the smallest among all schemes.

5.7. Algorithmic robustness

To show the performance of the proposed algorithm under inaccurate building thermal dynamics models, we evaluate the impact of thermal disturbances on the proposed algorithm. Here, thermal disturbance ρ_t is assumed to follow a uniform distribution with parameters $[\vartheta_l,\ \vartheta_u]^o F$ and $\vartheta_u = -\vartheta_l \in \{1.8, 3.6, 5.4\}^\circ F$, i.e., the maximum indoor temperature disturbance is 1, 2, and 3 °C, respectively. In Fig. 8, it can be seen that the proposed algorithm can generally achieve the best performance among all schemes. Therefore, the proposed algorithm has strong robustness against uncertainty in building thermal dynamics models.

5.8. Algorithmic versatility

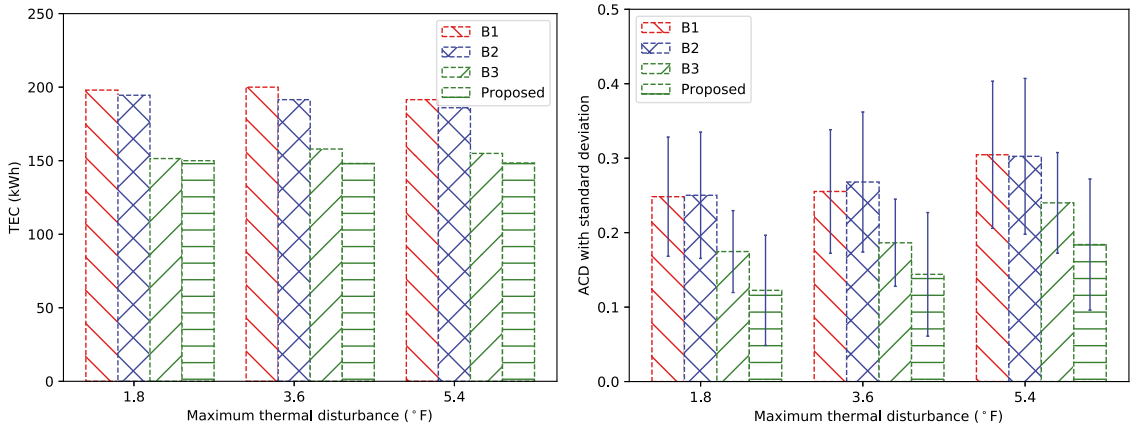
To show that the proposed algorithm does not depend on HVAC models mentioned in Section 3 (e.g., (1), (2), (9), and (10)), we consider a different building environment modeled by EnergyPlus v9.5. To be specific, we select a small office model "*ASHRAE9012016_OfficeSmall_Denver*" from the example files of EnergyPlus, which consists of 5 zones as shown in Fig. 9(a). Moreover, the indoor temperature of each zone can be adjusted by a direct expansion HVAC system, which can operate under cooling and heating modes. In this paper, we mainly focus on controlling the indoor temperature of "Core_ZN ZN" by dynamically setting the heating setpoint and the cooling setpoint. During the working hours, the heating setpoint is 22°C and the cooling setpoint is dynamically selected by the proposed algorithm from the range [22,28]. During the non-working hours, the heating setpoint is 0°C, and the cooling setpoint is dynamically selected by the proposed algorithm from the range [22,30]. For simplicity, we assume that "Core_ZN ZN" is occupied during the working hours and is unoccupied during the non-working hours. To support the training of DRL agents, a co-simulation framework is implemented based on Python and EnergyPlus, which can be observed in Fig. 9(b). Note that the communication between Python and EnergyPlus is implemented based on the Energy Management System (EMS) feature in EnergyPlus and a Python interface program, which can observe environment data



(a) Temperature set-point (b) Fan speed of PCS 1

(c) Accumulated energy consumption (d) Thermal comfort extent of occupant 1

Fig. 7. Performance comparisons of all schemes.

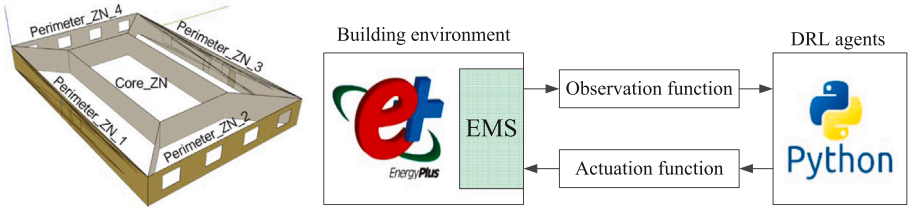


(a) TEC (b) ACD with standard deviation

Fig. 8. The impact of thermal disturbances on the proposed algorithm.

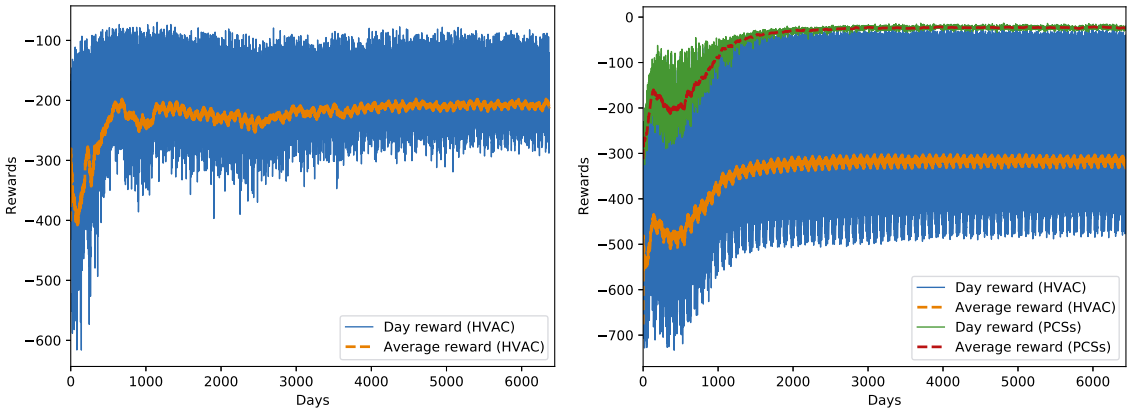
and send action command information at the beginning of each time slot. In simulations, we use the real-world weather data in Tampa, Florida, USA, which has the largest cooling demand from June 1 to September 30 among five cities (i.e., San Francisco, Golden, Tampa, Chicago, and Washington) with available weather data in EnergyPlus v9.5. To be specific, the data during June 1 and August 31 are used for training, while the data in September are used for testing. In the training process, a training episode consists of 92 days (i.e., the duration from June 1 to August 31) and each hour is divided into six-time slots. The convergence curves of B2 and the proposed algorithm can be found in Fig. 10, where 70 training episodes (note that the weather data in 92 days are used repeatedly in all training episodes) are needed for algorithmic convergence.

The performance comparisons under all schemes are shown in Table 1, where $\overline{\text{B1}}$ is equivalent to B1 when the cooling setpoint can lead to the lowest ACD. Moreover, $\overline{\text{B3}}$ is equivalent to B3 when the cooling setpoint can lead to the lowest ACD. It can be seen that the proposed



(a) Building structure (b) Co-simulation framework

Fig. 9. Building structure and co-simulation framework.



(a) B2 (b) Proposed

Fig. 10. Algorithmic convergence performances under a different building environment.

Table 1
Algorithmic performance comparisons.

Schemes	Setpoint	ACD					TEC
		user 1	user 2	user 3	user 4	all users	
$\overline{B1}$	22	0.1993	0.0075	0.0069	0.0159	0.0574	1996.2511
	23	0.1044	0.0081	0.0076	0.0890	**0.0523**	**1983.1294**
	24	0.0715	0.0619	0.0538	0.1931	0.0951	1968.3148
	25	0.0596	0.1419	0.1286	0.3055	0.1589	1952.2923
	26	0.0950	0.2324	0.2148	0.4211	0.2408	1934.8766
	27	0.1539	0.3282	0.3067	0.5385	0.3318	1916.3161
	28	0.2261	0.4266	0.4016	0.6565	0.4277	1896.7100
B2	Dynamic	0.0395	0.0016	0.0015	0.1200	**0.0407**	**1978.2749**
$\overline{B3}$	22	0.9841	0.6858	0.3557	0.5131	0.6347	2045.629
	23	0.7920	0.4948	0.1698	0.2975	0.4385	2019.4117
	24	0.5816	0.2856	0.0003	0.0611	0.2321	1988.0769
	25	0.3707	0.0758	0.0012	0.0057	0.1133	1953.6440
	26	0.1599	0.0030	0.0024	0.0870	**0.0631**	**1914.6618**
	27	0.0006	0.0099	0.0037	0.3280	0.0856	1872.9917
	28	0.0015	0.2197	0.1734	0.5666	0.2403	1828.6019
Proposed	Dynamic	0.0053	0.0003	0	0.0528	**0.0146**	**1900.1156**

algorithm can achieve the lowest ACD among all schemes, which means that the proposed algorithm can provide the best personalized thermal comfort for occupants in the shared office space. Furthermore, B2 can achieve a lower ACD than $\overline{B1}$ and $\overline{B3}$, which indicates the advantage of the DRL-based algorithm. In addition, the proposed algorithm can achieve lower TEC than $\overline{B1}$ when cooling setpoints are lower than 28 and $\overline{B3}$ when cooling setpoints are lower than 27. Compared with B1, B2, and B3, the proposed algorithm can reduce ACD by 72.08%, 64.13%, and 76.86%, respectively. Meanwhile, the proposed algorithm can reduce corresponding TEC by 4.18%, 3.95%, and 0.7%, respectively.

To explain the reason why the proposed algorithm can achieve the best performance than B1, B2, and B3, we provide more specific information in Fig. 11. It can be seen that B3 intends to choose a higher cooling setpoint than B1 during working hours with the help of PCSs, resulting in lower energy consumption. Compared with B1, B2 can achieve lower ACD and TEC by dynamically choosing cooling setpoints, which indicates the advantage of DRL technology. Compared with B3, the proposed algorithm can utilize the advantages of PCSs and DRL technology, i.e., creating more opportunities of turning off the HVAC system during working hours as shown in Fig. 11(d) by dynamically selecting reasonable cooling setpoints. As a result, the proposed algorithm achieves better performance than B3.

5.9. Discussions

In the above subsections, we have shown the convergence, flexibility, effectiveness, robustness, and versatility of the proposed algorithm. In this subsection, we would like to compare the findings obtained in this study with those obtained in existing works. Moreover, we

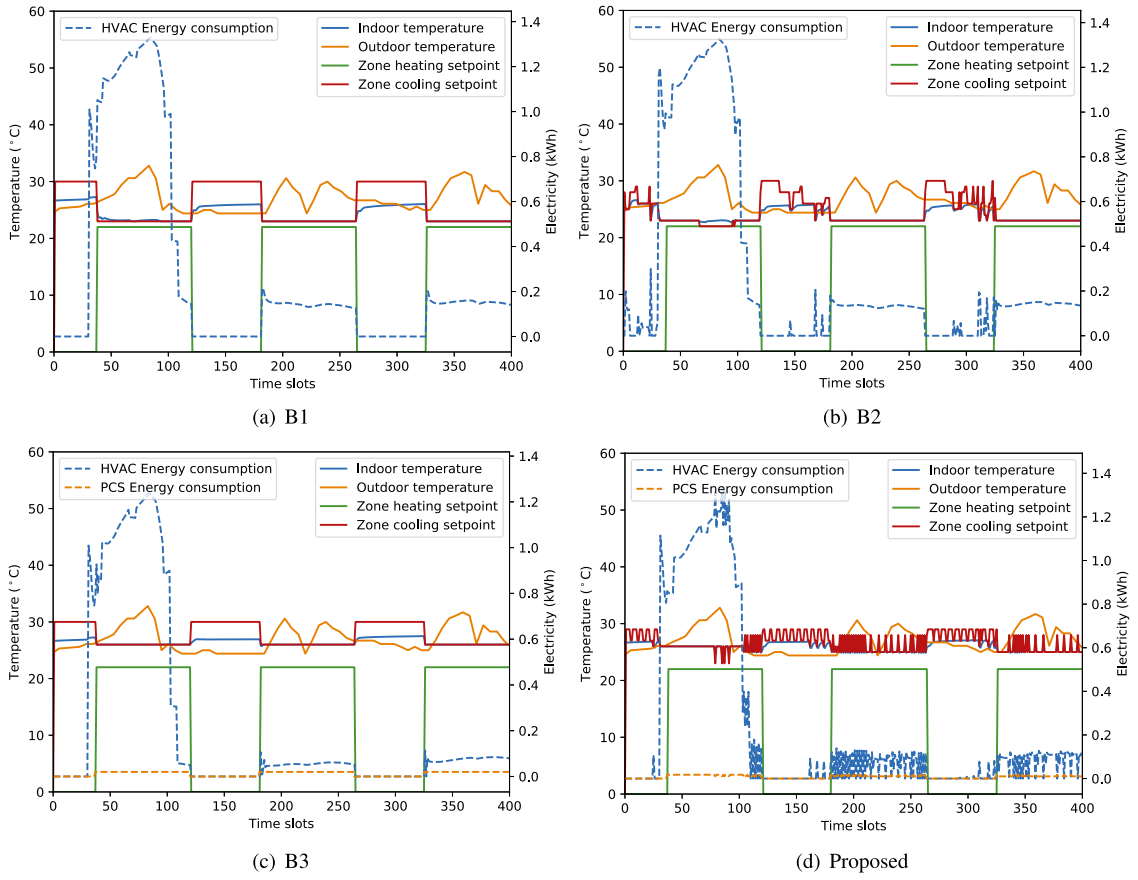


Fig. 11. Performance details among all schemes.

intend to point out the potential limitations of this study and discuss its practical applications and implications.

Similar to existing DRL-based HVAC control methods [20], this study also shows that dynamically adjustment of HVAC temperature set-points by DRL techniques contributes to saving energy and improving thermal comfort simultaneously. However, different from existing DRL-based works [18,20], the proposed algorithm can support efficient coordination among an HVAC system and PCSs in a shared office space without knowing an explicit building thermal dynamics model and any prior knowledge of uncertain parameters. To be specific, the proposed algorithm can reduce TEC by 0.7%–4.18% and ACD by 64.13%–76.86% simultaneously compared with existing methods without coordination.

Although the proposed algorithm has some strengths, it also has some potential limitations similar to existing DRL-based works [3,18,20]. The first limitation is that the actual performance of the proposed algorithm may be affected by training environment inaccuracy, which is common sense in the field of DRL. Typically, two training environment construction methods for building energy management could be adopted for supplementing the proposed algorithm. One method is to set up a long-short term memory (LSTM) based environment model using historical data [21]. Another method is to use measure data for calibrating a whole building energy model (BEM) that is represented by a BEM engine [42], e.g., EnergyPlus. Although the above methods are efficient, they cannot ensure that the obtained training environments are perfect (i.e., no inaccuracy). Fortunately, the proposed algorithm has strong robustness to imperfect training environments as shown in Section 5.7. The second limitation is that the actual control of an HVAC temperature set-point and fan speeds of PCSs may be interrupted by the direct control of occupants. Even so, a similar study has indicated that DRL-based thermal comfort control algorithms are still efficient when direct control of occupants is involved [20]. To reduce the occurrence frequency of interruption events in practice, two possible methods could be adopted, i.e., developing accurate personal comfort models for all occupants beforehand or considering fan speed preferences in the process of algorithmic design, which will be our future work.

In summary, the proposed algorithm can be applied to control an HVAC temperature set-point and fan speeds of PCSs efficiently in a shared office space/building. Although this study considers a direct expansion HVAC system, similar algorithms can be designed for other types of HVAC systems, e.g., variable air volume, and variable refrigerant flow. Therefore, the proposed algorithm has wide applications in various office buildings.

6. Conclusion

In this paper, we investigated an optimal coordination control problem among an HVAC system and PCSs in a shared office space. Due to the solving challenges caused by inexplicit building thermal dynamics models, uncertain parameters, temporally coupled constraints, and large action space, we reformulated the optimization problem as a cooperative Markov game. To solve the Markov game efficiently, we proposed an energy-efficient thermal comfort control algorithm based on AMADRL, which combines multi-agent DRL and attention mechanism. Note that the proposed algorithm did not require an explicit building thermal dynamics model and any prior knowledge of uncertain parameters. Extensive simulation results based on real-world traces showed the convergence, flexibility, effectiveness, robustness, and versatility of the proposed algorithm. In terms of algorithmic effectiveness, the proposed algorithm can reduce energy consumption by 0.7%–4.18% and reduce average thermal comfort deviation by 64.13%–72.08% simultaneously compared with a DRL-based method and two rule-based methods, which indicates the necessity of adjusting an HVAC temperature set-point and fan speeds of PCSs dynamically and cooperatively. In future work, we intend to evaluate the extent

of performance improvement that brought by the proposed algorithm when other types of HVAC systems (e.g., variable air volume, and variable refrigerant flow) are considered.

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.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFA0702200, in part by the National Natural Science Foundation of China under Grant 62192751, Grant 61972214, Grant 62073173, Grant 61833011, in part by China Postdoctoral Science Foundation under Grant 2020M673406, in part by Qinlan Project of Jiangsu Province (2022), and in part by 1311 Talent Project of Nanjing University of Posts and Telecommunications.

References

[1] Y. Yang, G. Hu, C.J. Spanos, Stochastic optimal control of HVAC system for energy-efficient buildings, IEEE Trans. Control Syst. Technol. 30 (1) (2021) 376–383.
[2] D. Li, C.C. Menassa, V.R. Kamat, Non-intrusive interpretation of human thermal comfort through analysis of facial infrared thermography, Energy Build. 176 (2018) 246–261.
[3] L. Yu, S. Qin, M. Zhang, C. Shen, T. Jiang, X. Guan, A review of deep reinforcement learning for smart building energy management, IEEE Internet Things J. 8 (15) (2021) 12046–12063.
[4] S.K. Gupta, S. Atkinson, I. OBoyle, J. Drogo, K. Kar, S. Mishra, J.T. Wen, BEES: Real-time occupant feedback and environmental learning framework for collaborative thermal management in multi-zone, multi-occupant buildings, Energy Build. 125 (2016) 142–152.
[5] Y. Peng, A. Rysanek, Z. Nagy, A. Schlüter, Occupancy learning-based demand-driven cooling control for office spaces, Build. Environ. 122 (2017) 145–160.
[6] S. Lee, J. Joe, P. Karava, I. Bilionis, A. Tzempelikos, Implementation of a self-tuned HVAC controller to satisfy occupant thermal preferences and optimize energy use, Energy Build. 194 (2019) 301–316.
[7] W. Li, C. Koo, T. Hong, J. Oh, S.H. Cha, S. Wang, A novel operation approach for the energy efficiency improvement of the HVAC system in office spaces through real-time big data analytics, Renew. Sustain. Energy Rev. 127 (2020) 109885.
[8] X. Shan, N. Luo, K. Sun, T. Hong, Y.-K. Lee, W.-Z. Lu, Coupling CFD and building energy modelling to optimize the operation of a large open office space for occupant comfort, Sustainable Cities Soc. 60 (2020) 102257.
[9] Y. He, W. Chen, Z. Wang, H. Zhang, Review of fan-use rates in field studies and their effects on thermal comfort, energy conservation, and human productivity, Energy Build. 194 (2019) 140–162.
[10] A. Aryal, B. Becerik-Gerber, G.M. Lucas, S.C. Roll, Intelligent agents to improve thermal satisfaction by controlling personal comfort systems under different levels of automation, IEEE Internet Things J. 8 (8) (2020) 7089–7100.
[11] Z. Xu, S. Liu, G. Hu, C.J. Spanos, Optimal coordination of air conditioning system and personal fans for building energy efficiency improvement, Energy Build. 141 (2017) 308–320.
[12] R. Kalaimani, M. Jain, S. Keshav, C. Rosenberg, On the interaction between personal comfort systems and centralized HVAC systems in office buildings, Adv. Build. Energy Res. 14 (1) (2020) 129–157.
[13] Y. He, N. Li, N. Li, J. Li, J. Yan, C. Tan, Control behaviors and thermal comfort in a shared room with desk fans and adjustable thermostat, Build. Environ. 136 (2018) 213–226.
[14] H. Zhang, E. Arens, Y. Zhai, A review of the corrective power of personal comfort systems in non-neutral ambient environments, Build. Environ. 91 (2015) 15–41.
[15] D.T. Nguyen, H.T. Nguyen, L.B. Le, Coordinated dispatch of renewable energy sources and HVAC load using stochastic programming, in: 2014 IEEE InteRnational Conference on Smart Grid Communications, SmartGridComm, IEEE, 2014, pp. 139–144.
[16] H. Zhang, D. Yue, C. Dou, K. Li, X. Xie, Event-triggered multiagent optimization for two-layered model of hybrid energy system with price bidding-based demand response, IEEE Trans. Cybern. 51 (4) (2019) 2068–2079.
[17] L. Yu, T. Jiang, Y. Zou, Online energy management for a sustainable smart home with an HVAC load and random occupancy, IEEE Trans. Smart Grid 10 (2) (2019) 1646–1659.
[18] T. Wei, Y. Wang, Q. Zhu, Deep reinforcement learning for building HVAC control, in: Proceedings of the 54th Annual Design Automation Conference 2017, 2017, pp. 1–6.
[19] X. Deng, Y. Zhang, H. Qi, Towards optimal HVAC control in non-stationary building environments combining active change detection and deep reinforcement learning, Build. Environ. (2022) 108680.
[20] W. Valladares, M. Galindo, J. Gutiérrez, W.-C. Wu, K.-K. Liao, J.-C. Liao, K.-C. Lu, C.-C. Wang, Energy optimization associated with thermal comfort and indoor air control via a deep reinforcement learning algorithm, Build. Environ. 155 (2019) 105–117.
[21] Z. Zou, X. Yu, S. Ergan, Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network, Build. Environ. 168 (2020) 106535.
[22] V. François-Lavet, P. Henderson, R. Islam, M.G. Bellemare, J. Pineau, et al., An introduction to deep reinforcement learning, Found Trends® Mach Learn 11 (3–4) (2018) 219–354.
[23] T.T. Nguyen, N.D. Nguyen, S. Nahavandi, Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications, IEEE Trans. Cybern. 50 (9) (2020) 3826–3839.
[24] X. Deng, Y. Zhang, Y. Zhang, H. Qi, Towards smart multi-zone HVAC control by combining context-aware system and deep reinforcement learning, IEEE Internet Things J. (2022) 1, http://dx.doi.org/10.1109/JIOT.2022.3175728.
[25] G. Yu, Z. Gu, Z. Yan, H. Chen, Investigation and comparison on thermal comfort and energy consumption of four personalized seat heating systems based on heated floor panels, Indoor Built Environ 30 (8) (2021) 1252–1267.
[26] H. Wang, J. Wang, W. Li, S. Liang, Experimental study on a radiant leg warmer to improve thermal comfort of office workers in winter, Build. Environ. 207 (2022) 108461.
[27] B. Yang, T.-H. Lei, P. Yang, K. Liu, F. Wang, On the use of wearable face and neck cooling fans to improve occupant thermal comfort in warm indoor environments, Energies 14 (23) (2021) 8077.
[28] W. Song, Z. Zhang, Z. Chen, F. Wang, B. Yang, Thermal comfort and energy performance of personal comfort systems (PCS): A systematic review and meta-analysis, Energy Build. 256 (2022) 111747.
[29] Y. He, N. Li, J. Lu, N. Li, Q. Deng, C. Tan, J. Yan, Meeting thermal needs of occupants in shared space with an adjustable thermostat and local heating in winter: An experimental study, Energy Build. 236 (2021) 110776.
[30] B. Rajasekhar, W. Tushar, C. Lork, Y. Zhou, C. Yuen, N.M. Pindoriya, K.L. Wood, A survey of computational intelligence techniques for air-conditioners energy management, IEEE Trans. Emerg. Top. Comput. Intell. 4 (4) (2020) 555–570, http://dx.doi.org/10.1109/TETCI.2020.2991728.
[31] Y. Du, H. Zandi, O. Kotevska, K. Kurte, J. Munk, K. Amasyali, E. Mckee, F. Li, Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning, Appl. Energy 281 (2021) 116117.
[32] A. Aryal, B. Becerik-Gerber, Thermal comfort modeling when personalized comfort systems are in use: Comparison of sensing and learning methods, Build. Environ. 185 (2020) 107316.
[33] S. Lee, P. Karava, A. Tzempelikos, I. Bilionis, A smart and less intrusive feedback request algorithm towards human-centered HVAC operation, Build. Environ. 184 (2020) 107190.
[34] L. Yu, W. Xie, D. Xie, Y. Zou, D. Zhang, Z. Sun, L. Zhang, Y. Zhang, T. Jiang, Deep reinforcement learning for smart home energy management, IEEE Internet Things J. 7 (4) (2019) 2751–2762.
[35] 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) (2020) 407–419.
[36] G. Mantovani, L. Ferrarini, Temperature control of a commercial building with model predictive control techniques, IEEE Trans. Ind. Electron. 62 (4) (2014) 2651–2660.
[37] S. Iqbal, F. Sha, Actor-attention-critic for multi-agent reinforcement learning, in: International Conference on Machine Learning, PMLR, 2019, pp. 2961–2970.
[38] K.S. Liu, E.V. Pinto, S. Munir, J. Francis, C. Shelton, M. Berges, S. Lin, Cod: a dataset of commercial building occupancy traces, in: Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments, 2017, pp. 1–2.
[39] H. Huang, L. Chen, E. Hu, A new model predictive control scheme for energy and cost savings in commercial buildings: An airport terminal building case study, Build. Environ. 89 (2015) 203–216.
[40] A.A. Thatte, L. Xie, Towards a unified operational value index of energy storage in smart grid environment, IEEE Trans. Smart Grid 3 (3) (2012) 1418–1426.
[41] P. Constantopoulos, F.C. Schweppe, R.C. Larson, ESTIA: A real-time consumer control scheme for space conditioning usage under spot electricity pricing, Comput. Oper. Res. 18 (8) (1991) 751–765.
[42] 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.