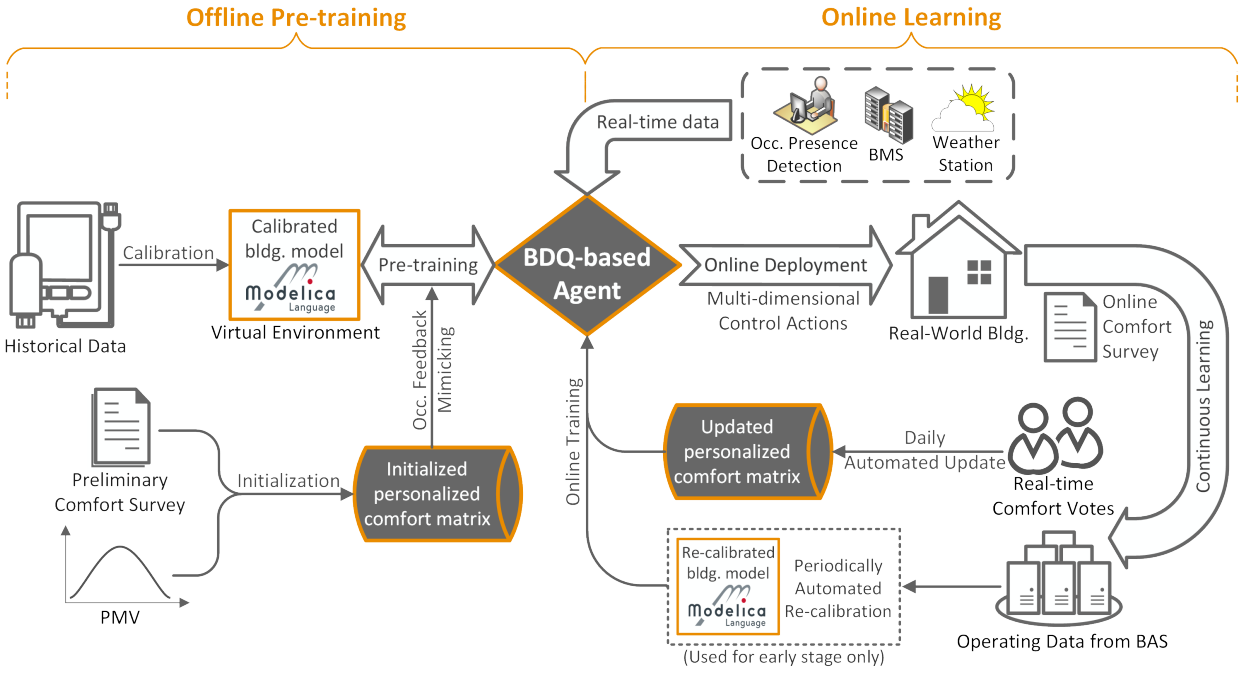
Graphical Abstract

A practical deep reinforcement learning framework for multivariate occupant-centric control in buildings

Yue Lei,Sicheng Zhan,Eikichi Ono,Yuzhen Peng,Zhiang Zhang,Takamasa Hasama,Adrian Chong



Please cite as:
Lei, Y., Zhan, S., Ono, E., Peng, Y., Zhang, Z., Hasama, T., & Chong, A. (2022). A practical deep reinforcement learning framework for multivariate occupant-centric control in buildings. Applied Energy, 324, 119742. doi: https://doi.org/10.1016/j.apenergy.2022.119742

Highlights

A practical deep reinforcement learning framework for multivariate occupant-centric control in buildings

Yue Lei,Sicheng Zhan,Eikichi Ono,Yuzhen Peng,Zhiang Zhang,Takamasa Hasama,Adrian Chong

- Framework for efficient multivariate occupant-centric building controls was proposed.
- Uses deep reinforcement learning and considers personal thermal comfort and occupancy.
- Proposed simple personal comfort model designed for practical building controls.
- Pre-trained virtually before online deployment in a real-world office.
- Reduced cooling energy by 14% and improved thermal acceptability by 11%.

A practical deep reinforcement learning framework for multivariate occupant-centric control in buildings

Yue Lei[a], Sicheng Zhan[a], Eikichi Ono[a,b], Yuzhen Peng[a], Zhiang Zhang[c], Takamasa Hasama[b] and Adrian Chong[a,*]

[a]Department of the Built Environment, National University of Singapore, 117566, Singapore
[b]Kajima Technical Research Institute Singapore, 449269, Singapore
[c]Department of Architecture and Built Environment, University of Nottingham Ningbo China, Ningbo, 315100, China

ARTICLE INFO

Keywords:
Occupant-centric control
Deep learning
Reinforcement learning
Thermal comfort
Energy efficiency

ABSTRACT

Reinforcement learning (RL) has been shown to have the potential for optimal control of heating, ventilation, and air conditioning (HVAC) systems. Although research on RL-based building control has received extensive attention in recent years, there is limited real-world implementation to evaluate its performance while keeping occupants in the loop. Additionally, many HVAC systems consist of multiple subsystems, but conventional RL algorithms face significant challenges when dealing with high-dimensional action spaces. This study proposes a practical deep reinforcement learning (DRL) based multivariate occupant-centric control framework that considers personalized thermal comfort and occupant presence. Specifically, Branching Dueling Q-network (BDQ) is leveraged as the learning agent to efficiently solve the multi-dimensional control task, and a tabular-based personal comfort modeling method is applied that is naturally integrated into human-in-the-loop operations. The BDQ agent is pre-trained in a virtual environment, followed by online deployment in a real office space for 5-dimensional action control. Based on the actual deployment and real-time comfort votes, our results showed a 14% reduction in cooling energy and an 11% improvement in total thermal acceptability.

1. Introduction

With a significant increase in building energy consumption, research on HVAC related energy conservation measures has become essential. In 2020, the residential and commercial sectors were responsible for approximately 40% of primary energy consumption and 35% carbon emissions in the U.S. [1]. Similar statistics were provided by the European Commission [2] for most EU countries. According to Fernandez et al. [3], commercial buildings have the potential to save an average of 29% in energy consumption by installing the advanced control technologies that are available today. The researchers quantified the saving potentials by investigating the improvements resulting from applying 34 different building control measures to 14 distinct building types in 16 climate zones across the U.S. In addition, while buildings are designed to provide a comfortable space for occupants, a post occupancy study revealed that over 50% of the 52,980 occupants were dissatisfied with their indoor environment [4]. Therefore, it is an ongoing challenge to utilize the advanced control and operation techniques in practical application to attain desirable building performance [5].

Leveraging increasingly affordable smart metering technologies, a growing number of studies investigated how buildings can take advantage of occupant-centric controls and the Internet of Things (IoT) to accommodate occupants' comfort and improve energy efficiency [6]. Based on existing literature on occupant-centric controls, the reported median energy savings and comfort improvement was 22% and 29.1%, respectively [7, 8]. Although many advanced building control techniques have been developed over the past decade, it remains challenging due to the high-order nonlinearity of dynamics and the complexities resulting from numerous disturbances, constraints, and uncertainties in buildings [9]. Examples of advanced building controls include model predictive control (MPC) [10, 11], linear quadratic regulator [12], and learning-based control [13, 14]. More recently, there has been increasing interest in

bdgleiy@nus.edu.sg (Y. Lei); szhan@u.nus.edu (S. Zhan); e0491209@u.nus.edu (E. Ono); yuzhen.p@nus.edu.sg (Y. Peng); zhiang.zhang@nottingham.edu.cn (Z. Zhang); t.hasama@kajima.com.sg (T. Hasama); adrian.chong@nus.edu.sg (A. Chong)
ORCID(s): 0000-0001-8103-7144 (Y. Lei); 0000-0002-9872-8555 (S. Zhan); 0000-0003-3876-1924 (E. Ono); 0000-0001-5079-0897 (Y. Peng); 0000-0002-7473-9786 (Z. Zhang); 0000-0001-5855-7782 (T. Hasama); 0000-0002-9486-4728 (A. Chong)

reinforcement learning (RL) because of its adaptability and flexibility [15]. More importantly, since RL algorithms generally do not need to analyze the building dynamics at each control step, they provide significantly more efficient run-time control than those optimization-based controllers, which opens up the possibility of applying RL-based controls to solve complex building automation problems [16].

1.1. Multivariate Building Controls

Existing studies focusing on the optimal control of HVAC systems typically consider a single control variable, such as the zone's temperature setpoint or supply airflow rate. However, modern HVAC systems may consist of more than one subsystem. Consequently, there is a need to control multiple variables simultaneously. Notable examples are radiant panels with dedicated outdoor air system (DOAS), mixed-mode ventilation systems, ceiling fans augmented air-conditioning, and personalized ventilation with ambient fan coil units. Independent univariate local controls are typically applied to operate each subsystem in these HVAC systems. However, it is necessary to ensure communication and cooperation between them to reach the full potential. Failing to consider possible subsystem interactions when developing a central control strategy could result in sub-optimal performance [17].

Multivariate control problems can be solved with multi-input-multi-output (MIMO) controllers. There are extensive studies on applying the self-tuning fuzzy PID (proportional-integral-derivative) or fuzzy PI-PD based control to the MIMO HVAC systems [18]. While these controllers are effective and robust, they are often not optimal because the predictive information is not considered and the adaptive adjustment ability is limited [19]. MPC is also a viable control strategy for MIMO systems owing to its capability of handling multivariate interactions and constraints. However, developing an efficient model for online optimization is not trivial [20]. Creating a model that accurately reflects system dynamics can also be challenging due to the ill-posedness of calibrating these models [21]. More recently, multi-agent RL algorithms were applied to the building domain for multi-dimensional action controls. Examples include multi-zone building control [22, 23] and microgrid optimization [24, 25]. In particular, some studies employed clustering-based methods to enhance the learning efficiency and stability of the agents with extremely large state-action spaces [26]. However, in these multi-agent systems, each agent typically acted as a univariate controller and was responsible for one action dimension, and multiple agents worked together to complete control tasks. Although multi-agent RL is capable of multivariate building controls [27], its scalability remains challenging due to the need for many heterogeneous agents [28]. Hence, a more practical and intuitive approach to achieving scalability is implementing a single-agent multivariate control in individual rooms and extending it to multiple agent settings to control the whole building jointly.

Leveraging advances in deep learning and RL, it is computationally tractable to learn policies in high-dimensional and continuous action spaces using deep function approximators [29]. Few studies investigated single-agent multivariate RL control in the field of building HVAC controls. Sun et al. [30] implemented a Lagrangian relaxation based method to co-optimized the fresh air unit and fan coil unit. Ding et al. [31] proposed a DRL-based framework that optimizes the control of four building subsystems (HVAC, window, blind and lighting). As high-dimensional action space in zone level control become increasingly important in buildings, more investigation is needed on how to efficiently implement multivariate HVAC control with single-agent algorithms, and how to scale them to multi-zone building using fully cooperative multi-agent methods.

1.2. Implementing DRL-based HVAC Control

Wei et al. [16] first applied DRL algorithms to HVAC controls in 2017. Using simulation, they showed that deep Q-network (DQN) outperformed the baseline rule-based approach and conventional Q-learning in building energy cost and temperature violation. Since then, DRL-related research has been increasingly applied in the field of building controls [32]. However, practical implementations of DRL-based building controls remain scarce, often being limited to domestic water heating [33] and lighting [34] control applications. Actual implementations are essential because many real-world challenges are often not reflected in virtual environments [35]. For instance, DRL is notorious for requiring large amounts of training data to achieve plausible results, making the learning process very inefficient in real-world operations [36, 37]. Additionally, RL agents learn by exploring new actions, which will inevitably lead to undesirable outcomes when the algorithm performs sub-optimally [38]. Challenges encountered during the training process may also be overlooked in simulation-based studies because they usually focus on the final control performance (e.g., energy savings). Lastly, the high uncertainties in using actual thermal comfort votes [39] and the prevalence of faulty equipment and sensors in actual buildings [40, 41] are often not considered in a virtual environment.

Few studies demonstrated the improvements in control performance from using DRL in real buildings. Zhang et al. [42] pre-trained an A3C (Asynchronous Advantage Actor Critic) based agent with a calibrated building energy model

before deployment in an actual office space to avoid inefficient learning. However, this algorithm required 3.5 million interaction steps (33 years in simulation) to achieve the satisfactory performance. Chen et al. [43] adopted a novel differentiable MPC policy in place of a generic neural network in the DRL. The pre-trained policy was directly deployed in a conference room for 3 weeks and the results showed that the proposed approach saved 16.7% of cooling demand compared to the existing PID controller. Although this approach was much more sample-efficient than conventional DRL algorithms, it assumed a linear model for the system dynamics, which may not extrapolate to more complex problems. Qiu et al. [44] optimized chiller performance by employing a chilled water reset strategy with a hybrid model-free RL. Specifically, the expertise knowledge was incorporated to dynamically narrow down the action space. The proposed method was compared with expert manual control in a real central chiller plant.

In the context of using DRL for occupant-centric controls, wearable devices were often utilized to measure the physiological data. For example, Zhang et al. [45] used double deep Q-learning (Double DQN) to control the ceiling fan speed for 14 participants in an experiment room while considering their biological responses (wrist temperature, heart rate, and RR interval) and subjective thermal comfort votes in the control loop. Jung et al. [46] reduced thermal discomfort by 10.9% without increasing energy consumption using a one-dimensional convolutional neural network-based DRL algorithm that considers occupant activity from the smart wristband. While these studies showed that the thermal comfort can be maximized through the physiological information, they typically required the development of highly personalized comfort models prior to the control deployment. Consequently, occupants need to participate in dedicated comfort experiments for a few hours to collect the high granularity physiological data, which is difficult to achieve in practice.

1.3. Demand for Occupant-centric Comfort Management in RL-based Control

Instead of controlling the indoor environment at a constant and fixed setpoint temperature, occupant-centric control intends to provide personalized comfort conditions based on occupants' feedback [47]. Studies have shown that conditioning the micro-environment of each occupant based on the matching personal comfort model often leads to co-benefit of energy savings and occupants satisfaction [48, 49]. A personal comfort model is used to predict the thermal comfort response of an individual rather than the average response of a large population [50]. Although the predicted mean vote and predicted percentage of dissatisfied (PMV-PPD) model is widely recognized, it is an aggregated model and has been shown to have poor predictive performance if applied to individuals or a small group of occupants [50, 48]. Cheung et al. [51] analyzed the accuracy of PMV-PPD model using the ASHRAE global thermal comfort database II, and found the overall accuracy was only 34%. Additionally, the reliable range of the PMV is narrow, and it is also difficult and expensive to obtain full measurements of all input variables [52, 53]. However, most HVAC control studies still use the PMV or pre-defined temperature ranges because the focus is often on achieving energy savings, and occupant satisfaction is usually not a primary consideration [54]. It has been shown that using the temperature as a reference for HVAC system operations is inadequate [55]. Thanks to the recent advancements in the occupant behavior sensing and modeling, high-accuracy occupancy presence detection technologies have become readily available [47]. Therefore, to have a tangible impact on both the energy efficiency and occupants satisfaction in buildings, more research is needed to integrate the personalized thermal comfort modeling into actionable RL-based control frameworks.

1.4. Research Gaps and Objectives

Two major gaps are identified in the existing literature.

1. Past research on the optimal HVAC control typically consider a single control variable, which cannot fully exploit the potential of HVAC systems comprising multiple subsystems. Dealing with high-dimensional action spaces is nontrivial because the computational cost for conventional DRL algorithms grows exponentially with the number of action dimensionality.
2. Efforts to incorporate personalized thermal comfort modeling into RL-based HVAC control loop have been limited, especially for real-world integration. Existing RL-based occupant-centric control studies typically rely on the personalized physiological data collected from dedicated comfort experiments prior to the control deployment, which limits their application and scalability.

Therefore, this study proposed a new DRL control framework to fill the two knowledge gaps in an experimental-based study. The objective of this study is to develop a DRL-based multivariate control framework for building HVAC systems while considering personal thermal comfort and occupant presence. To that end, we applied a novel action branching architecture to efficiently solve control problems in multi-dimensional action spaces. A simple tabular-based

personal comfort modeling method is also proposed, which fits naturally into human-in-the-loop building operations. To demonstrate the control performance, the proposed DRL agent is implemented in an actual office building with real-time occupant comfort votes.

2. Methodology

This section presents the technical details used in the proposed multivariate DRL-based control framework.

2.1. Description of the DRL Control Framework

As shown in Figure 1, there are three key components in the proposed DRL control framework, which we elaborate in the subsequent sub-sections:

- BDQ-based agent (Section 2.2).
- Personalized thermal comfort matrix (Section 2.3).
- Virtual environment for RL pre-training (Section 2.4).

Broadly speaking, the implementation of RL-based control in actual buildings consists of 2 major tasks: (1) pre-training and (2) online learning. In the first task, the DRL agent is pre-trained offline in a virtual environment to ensure a plausible control performance and mitigate sub-optimal or unsafe actions [38]. We created the virtual environment with Modelica Buildings Library [56] and used the building's historical weather and operating data to calibrate the simulation model. A one-time preliminary comfort survey and the PMV model were used to initialize the personalized comfort matrix for the agent's pre-training.

In the second task, the pre-trained agent would directly control the actual HVAC systems. During this process, the agent continuously refined its control policy by learning from the real-time feedback from the HVAC system and occupants. The daily agent online training was conducted at the end of the day (i.e., unoccupied hours). The personalized comfort matrix was updated daily based on the comfort votes, and the simulation model was also periodically re-calibrated with the latest building operating data.

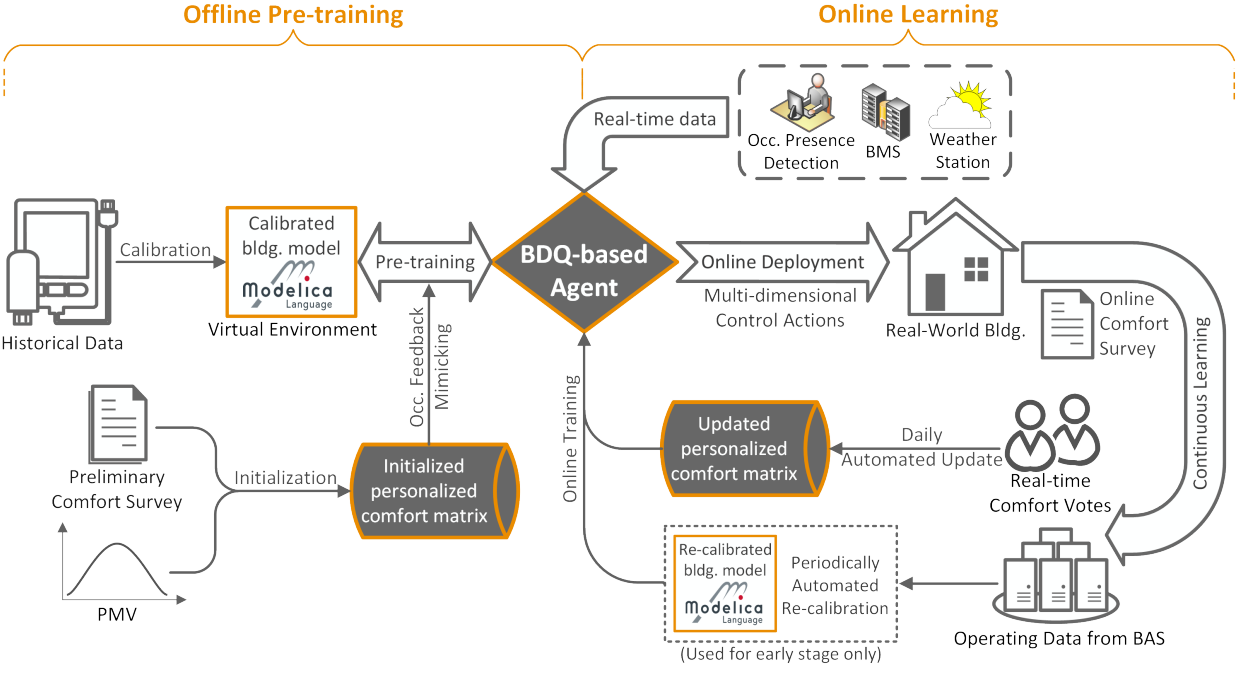


Figure 1: Block diagram of the proposed multivariate DRL control framework.

2.2. Multivariate Control with BDQ-based Agent

Due to the "curse of dimensionality", conventional DRL algorithms, such as DQN, are limited to problems with discrete and low-dimensional action spaces [28]. Yang et al. [57] showed that a univariate batch Q-learning agent required three years of training data to outperform a rule-based controller in a low exergy building. As the number of action variables increases, the training time duration will significantly increase because the number of sub-actions that need to be explicitly represented grows exponentially with the number of action dimensions [58]. Hence, this study implemented a novel action branching architecture to address the intractable training time in high-dimensional environments using conventional DRL algorithms.

2.2.1. Reinforcement Learning

Reinforcement learning is a branch of machine learning that enables an agent to interact with an environment and learn what actions to take based on the observed state of the environment. The agent learns by trial and error and is rewarded for desired behaviors and penalized for undesired behaviors [59]. In the context of HVAC controls, the RL agent refers to the controller, and the environment is the building or its virtual representation. The optimal control problem is modeled mathematically as a Markov decision process (MDP), as shown in Figure 2.

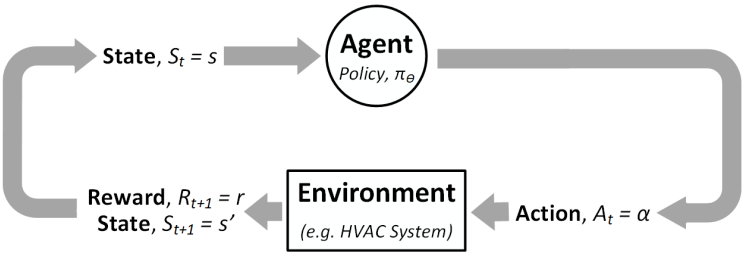


Figure 2: Standard agent-environment interaction.

When executing a control task, a closed-loop process is performed as follows: At each time step, an agent observes a state $S_t = s$ representing the current situation of the environment. Following its control policy π, the agent takes action $A_t = \alpha$ at time t, i.e., $\pi(s) = \alpha$ for a deterministic policy. In response to the action, after one time step, i.e., at time $t + 1$, the agent finds itself at a new state $S_{t+1} = s'$ and receives a reward $R_{t+1} = r$ from the environment. Note that in DRL, the policy is commonly approximated by a deep neural network (DNN). The aim of the agent is to find an optimal policy that maximizes the discounted accumulated rewards $\sum_t^\infty \gamma^t R_t$ by exploring different control policies. $\gamma \in [0, 1)$ is the discount factor.

2.2.2. Branching Dueling Q-network

This study leveraged the Branching Dueling Q-Network (BDQ), a branching variant of the Dueling Double Deep Q-Learning Network [58], to tackle multivariate control problems.

- **Double DQN** In the standard DQN, the target network parameters are updated to reflect desired changes in Q-Values. However, it has been revealed that the target network is most likely to overestimate Q-Values since the same estimator is used to select and evaluate actions, leading to sub-optimal and unstable training [60]. The idea of the double DQN is to reduce overestimations by using two independent Q-Values estimators for action selection (on the online network) and action evaluation (on the target network). This algorithm not only yields more accurate value estimations but also finds better policies.

- **Dueling DQN** The key idea behind the dueling architecture relies on the representation of the Q-Value of a state-action pair $Q(s, a)$ as two streams: the state-value $V(s)$ and the advantages for each action $A(s, a)$. As shown in Equation 1, the estimates of the value function are computed as the value of a given state and the value of the advantage of taking a particular action in this state compared to all other possible actions in this state [61]. However, Equation 1 cannot be directly used because it is unidentifiable. A common practice to address this issue it to force the advantage estimator to be zero at the best action for that state. Consequently, the maximum predicted advantage can be subtracted from the function as shown in Equation 2. Many studies have suggested that the maximum operator in Equation 2 can be replaced with an averaging operator for better performance

[58, 62]. Implementing the dueling architecture often leads to significant improvements in learning efficiency [61].

$$Q(s,a) = V(s) + A(s,a) \tag{1}$$

$$Q(s,a) = V(s) + \left(A(s,a) - \max_{a' \in \mathcal{A}} A(s,a') \right) \tag{2}$$

Figure 3 illustrated the action branching network architecture for the proposed BDQ agent. The shared decision module consists of multilayer neural networks, which extract features from the input state. The latent representations are then decomposed into the state value and state-dependent action advantages for each subsequent independent branch. The latter is distributed on the several action branches, and each branch is responsible for controlling one dimension of the action. The Q-values for each action dimension is calculated by combining the state value and action advantages through a special aggregation layer. Finally, the argmax function is used to concatenate the joint-action tuples from sub-action branches. This network architecture achieves a linear increase in the number of network outputs by allowing independence for each action dimension, which is particularly efficient for multivariate tasks.

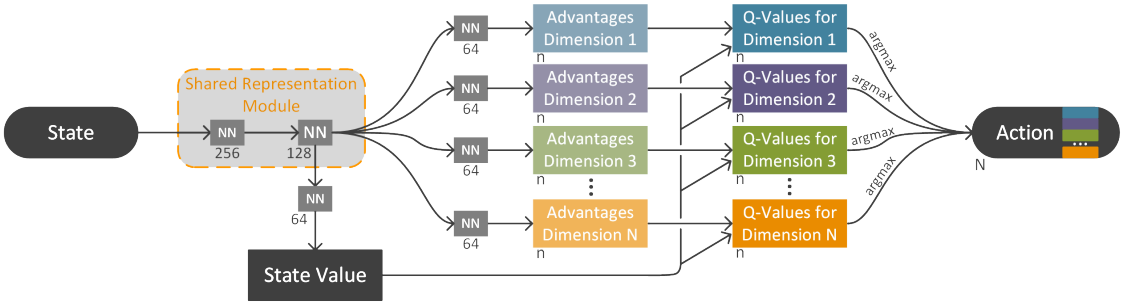


Figure 3: Visualization of the action branching network architecture for the proposed BDQ agent. The gray rectangles denote the weights of the fully connected neural network layers. The advantage dimension and Q-value dimension is also indicated.

Formally, for an action dimension $d \in \{1, ..., N\}$, we discretize the it into n feasible sub-actions. A key concept of the dueling network architecture is to explicitly decouple the state value and the action advantages in deep Q-networks to enhance the learning efficiency and robustness [61]. Consequently, the individual branch's Q-value at state s when decision $a_d \in \mathcal{A}_d$ is taken can be expressed by:

$$Q_d(s,a_d) = V(s) + \left(A_d(s,a_d) - \frac{1}{n} \sum_{a'_d \in \mathcal{A}_d} A_d(s,a'_d) \right) \tag{3}$$

Where $V(s)$ denotes the common state value computed by the shared decision module and $A_d(s,a_d)$ denotes the corresponding state-dependent action advantage. The temporal-difference (TD) target for BDQ agent can be calculated as:

$$y = r + \gamma \frac{1}{N} \sum_{d} Q_d^- \left(s', \underset{a'_d \in \mathcal{A}_d}{argmax} Q_d(s', a'_d) \right) \tag{4}$$

Where r is the immediate reward after taking sub-action a_d; γ is the discount factor; and Q_d^- is the d^{th} branch of the target network Q^-, which will be periodically updated with the latest weights from the online network Q. Let $\mathcal{D}$

denote the replay buffer and a denotes the joint-action tuple $(a_1, a_2, ..., a_N)$, the weights of the BDQ agent are iteratively adjusted by minimizing the loss function:

$$L = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\frac{1}{N}\sum_{d}\left(y_d - Q_d(s,a_d)\right)^2\right] \tag{5}$$

2.2.3. BDQ-based Agent Training

The pseudocode for detailed training process of the proposed BDQ-based agent is outlined in algorithm 1. We considered each day as an episode.

Algorithm 1: The training process of the BDQ-based agent for hybrid HVAC system

```
Randomly initialize the weights θ of the online network Q and the target network Q⁻;
for i = 0, ..., # Episodes do
    Observe the initial state s_t;
    for t = 0, ..., # Control_Steps do
        Compute the greedy action with BDQ agent;
        Sample a_t from a Gaussian distribution with its mean at the greedy action and standard deviation
          based on a linear schedule;
        Execute a_t if the action is within the action limits, otherwise re-sample a_t;
        Observe the new state s_{t+1};
        Compute reward r_t by Equation 8;
        Store the transitions (s_t, a_t, r_t, s_{t+1}) to the experience replay buffer D;
        Calculate priority sampling weight and store in D;
        if t mod Training_Steps = 0 then
            Sample a mini-batch of transitions from D;
            Estimate Q-value by Equation 3;
            Compute TD target by Equation 4;
            Perform gradient descent step on the loss by Equation 5 to update the parameters θ in the online
              network Q;
            Update priorities of sampled transitions;
        end
        if t mod Network_Update_Steps = 0 then
            Update the target network Q⁻ = Q
        end
    end
end
```

2.3. Personalized Thermal Comfort Matrix

Although many studies have been conducted on personal comfort models, they often lack a vision for real-world integration in building controls [50]. This study proposes a personalized thermal comfort matrix that aligns with standard thermal comfort surveys and integrates individuals' comfort needs into the DRL control loop. Specifically, the comfort matrices (one for each occupant) are initialized based on the PMV model and a one-time preliminary comfort survey. The PMV model is used to develop the neutral-preference comfort matrices, then modified based on the survey responses. After initialization, the comfort matrices are used to pre-train the agent. Unlike many other data-driven personal comfort models requiring extensive training data, the proposed method relies only on a one-time survey before enabling online control. During the online learning phase, the initialized comfort matrices will be continuously updated through the real-time comfort votes. Gradually, each comfort matrix becomes customized for the corresponding occupant.

Assuming that the personalized comfort level is mainly determined by two controllable indoor environment variables (e.g., room temperature and air velocity), we will discretize them into m and n intervals, respectively. Let C^k

Table 1
Summary of data sources and their usage at different agent training phases

Data	Offline pre-training	Online learning	
		Early stage	Late stage
Thermal comfort feedback	Initialized personalized comfort matrix	Updated personalized comfort matrix	
HVAC energy feedback	Calibrated virtual environment	Periodically re-calibrated virtual environment by GP-BO	Real-time BMS data

denotes the m-by-n comfort matrix for k^{th} occupant, where each entry $(c_{ij}^{k}) \in \mathbb{R}^{m \times n}$ represents the thermal discomfort penalty value at i^{th} and j^{th} environment condition. In particular, the thermal discomfort penalty must align with the standard thermal comfort surveys, such as the ASHRAE 7-point thermal sensation scale or the 4-point thermal acceptability scale [63]. An example of the thermal discomfort penalty values can be defined in the following format to map the personalized comfort level to the RL reward calculation: {Clearly acceptable (0), Just acceptable (-1), Just unacceptable (-2), Clearly unacceptable (-3)}. Note that instead of directly learning from occupants' feedback, the proposed agent always interact with the comfort matrices for two reasons. First, since real-time comfort votes at a high temporal granularity are difficult to acquire, the control time step is usually much smaller than the comfort votes interval. Consequently, a personal comfort model is needed to compute the reward for the agent at every control time step. Second, the real-time comfort votes may include inconsistent responses due to the recall bias [39], which will confuse the learning agent. Hence, the comfort matrices also act as a "buffer" to consolidate inconsistent comfort votes, thereby stabilizing the online learning. The detailed process for the initialization and online update of the proposed personalized comfort matrix is illustrated in Section 3.3. This study employed a two-dimensional matrix because it was assumed that the thermal comfort is mainly related to the two environmental variables in the experiment setting, which can be easily extended to higher dimensions if more variables are considered.

2.4. Virtual Environment for Agent Training

RL agents learn to make better decisions by trial and error, inevitably leading to unsafe actions. Virtual environments play an essential role in RL research in mitigating undesirable outcomes during the exploration and accelerating the training process. In this study, the offline pre-training and the early stage of online training are done by having the agent interact with a Modelica-based virtual environment, as shown in Figure 1. The BDQ agent training relies on the prioritized experience replay technique. It stores trajectories of experience in a replay buffer and lets off-policy agents reuse past experiences and prioritize important transitions. At each training step, a mini-batch of experiences will be sampled from the replay buffer to update the weights of the deep neural network. This technique has been shown to significantly improve the sample efficiency and stability of the network during training [64]. We selected Modelica over other traditional building simulation tools, such as EnergyPlus and TRNSYS, because it inherently models the building control dynamics without relying on additional software or scripts. To allow co-simulation between the DRL agent and the virtual environment, an OpenAI gym [65] interface wrapper for the pre-compiled Modelica model is developed based on PyFMI [66].

The data sources data sources and their usage at different agent training phases are summarized in Table 1. During offline pre-training, the virtual environment provides synthetic building operating data based on historical weather files, and the initialized comfort matrix mimics the comfort feedback of the actual occupants. The deep neural network training and hyperparameters tuning are performed to maximize the reward in this interactive simulation environment. Finally, the pre-trained agent with fine-tuned hyperparameters is prepared to control the actual HVAC system.

By contrast, the online agent training is divided into two stages. At the early stage, the agent still interacts with the virtual environment. This is because, in principle, the size of the replay buffer should be large enough to contain a wide variety of experiences; otherwise, the network tends to overfit the most recent trajectories and forget what it has learned from previous experiences. Consequently, the replay buffer needs to collect many experiences before starting training. Since only limited data is collected at the early stage of online deployment, the agent is trained daily in the virtual environment with the updated personalized comfort matrices. However, at the late stage, once the reply buffer is filled with the data from the actual HVAC system, the agent will directly learn from the BMS data onwards. Zhang and Sutton [67] compared the performance of different replay buffer sizes in two notable control tasks and concluded

that 10^3 to 10^4 is the optimal size. Hence, the transition from online agent training with the virtual environment data to the actual environment data takes about 5 to 10 weeks.

Throughout the early stage of online training, automated model calibration is conducted periodically to address the discrepancy between the virtual and actual environment. Gaussian Process-based Bayesian Optimization (GP-BO) is applied for calibration. Over BO iterations, the calibrated parameter sets and corresponding objective values are collected. A Gaussian Process is used to model the parameter $\mapsto$ objective relationship, and parameters that possess a higher probability of minimizing the objective function are accordingly selected for the next iteration until the threshold is met. Several studies have revealed that the Bayesian calibration provides better accuracy with reduced parameter uncertainty [68]. More details about GP-BO can be found in [69].

3. Case Study

3.1. Building and System

The proposed DRL control framework was implemented in an actual office building. Located on the National University of Singapore (NUS) campus, the SDE4 building is a net-zero energy and WELL platinum certified building. It employs a novel hybrid cooling system that consists of DOAS units and ceiling fans. This system can deliver better indoor air quality since the DOAS provides 100% conditioned fresh air. With the help of ceiling fans, significant energy savings can be achieved because the building is operated at an elevated room temperature and air movement, using the extended thermal comfort boundaries suggested by ASHRAE Standard 55 [70]. According to a previous study [71], the optimal operating condition for this hybrid system is 27°C room temperature with fan speed mode 3, which leads to energy savings while achieving both occupant thermal and IAQ (indoor air quality) satisfaction. Compared with 24°C temperature setpoint in typical office buildings, they estimated that the annual HVAC energy consumption was reduced by 26% under this operating condition. However, the occupant presence was not considered in their study, which could lead to a sub-optimal thermal environment in the actual building. Considering the the personalized thermal comfort and occupant presence, the proposed DRL-based controller can dynamically choose the optimal temperature setpoint and speed for each ceiling fan to improve the control performance.

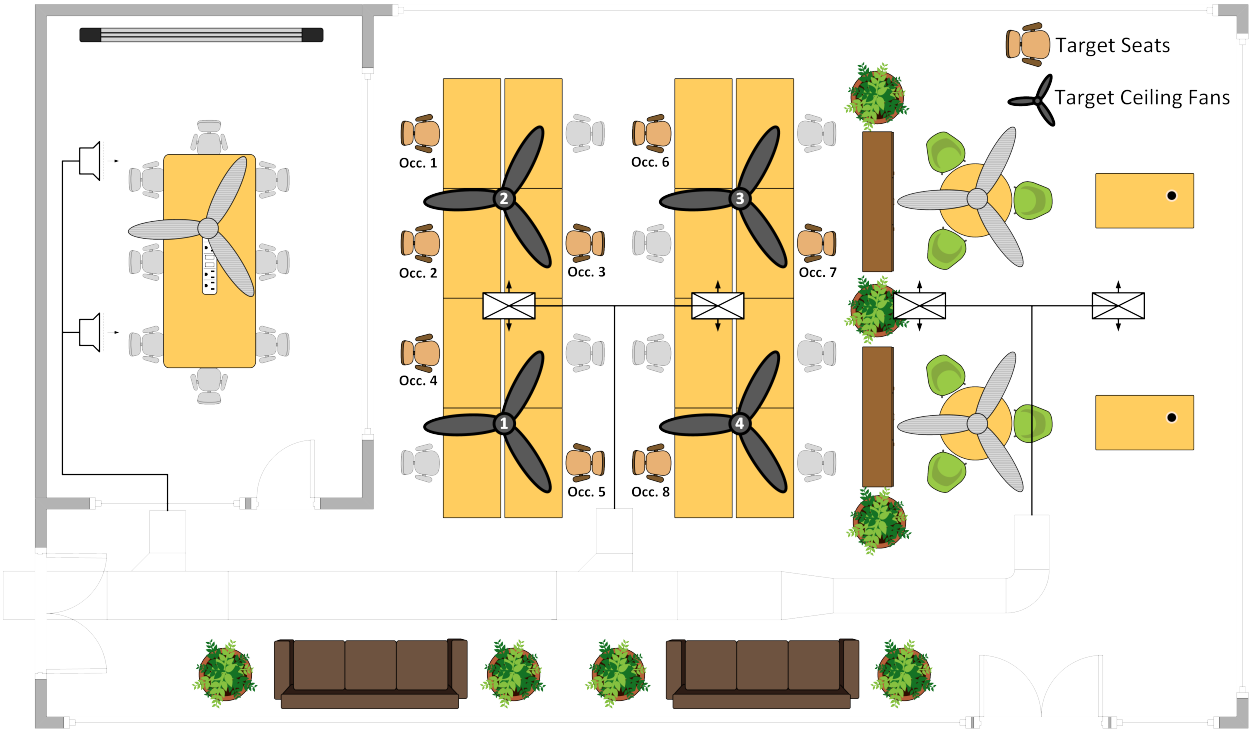


Figure 4: Plan view of the open office used in this study to evaluate the control performance of the proposed DRL control framework. The office seating plan and the location of the ceiling fans are illustrated.

Approved by the Institutional Review Board (IRB), an open office located on the 3rd floor of the SDE4 building was selected to conduct the experiment. As illustrated in Figure 4, it is 164m^2 and provides workspaces for 16 resident occupants. However, due to the safety measures for the COVID-19 pandemic [72], only eight occupants participated in the experiment since other people could not commit to coming to the office regularly. Participants were students and research staff from different countries between 20 to 40 years old. The experiment area has four ceiling fans, each serving four desks. We assumed that the air movement at each desk is mainly determined by the closest ceiling fan.

3.2. DRL Control for the Hybrid Cooling System

3.2.1. State Design

Each state observation consists of 10 items shown in Table 2. Min-Max normalization is applied to constrain their values between 0 and 1 (Equation 6). ob_{min} and ob_{max} are the lower and upper bounds of the state items. The real-time state variables can be accessed via the BAS. Note that only the latest data stored in the BAS from the current control time step was used, so weather forecasting is not required for this study.

$$ob_{norm} = \frac{ob - ob_{min}}{ob_{max} - ob_{min}} \tag{6}$$

Table 2
State items and the corresponding upper and lower limits for the BDQ agent

No.	Item	ob_{min}	ob_{max}
1	Room Temperature (°C)	24	29
2-5	Modes for Ceiling Fan 1 to 4*	0	5
6	Hour of the Day	0	23
7	Outdoor Air Temperature (°C)	21.5	34.5
8	Outdoor Solar Radiation (W/m^2)	0	1,150
9	Outdoor Relative Humidity (%)	40	100
10	Occupancy Flag**	0	1

* Each ceiling fan can be independently operated at 6 speed modes. M0 is off and M5 is the highest speed.
** The occupancy flag is a binary variable indicating the occupied hours. The flag is 1 for the period 10:00 am – 6:00 pm of weekdays.

3.2.2. Action Design

The action space is 5-dimensional, including the room temperature setpoint and four ceiling fans' speed mode. The temperature setpoint can take seven discrete values at 0.5°C interval: {26°C, 26.5°C, 27°C, 27.5°C, 28°C, 28.5°C, 35°C}, and each ceiling fan can take 6 actions: {M0 (off), M1, M2, M3, M4, M5}. Consequently, the action space totals up to 9,072. Formally, for an environment with an N-dimensional action space and n_d discrete sub-actions for each action dimension d, there are a total of $\prod_{d=1}^{N} n_d$ possible actions to be considered if using conventional DRL algorithms. This will rapidly encounter the issue of exponential explosion for the domains with large action spaces.

In particular, 35°C is the setback setpoint during the occupied hours, during which the supply airflow of the DOAS is kept at the lowest limit (10%). This value complies with the minimum outdoor air requirements in the local building code. Additionally, the BDQ agent can be extended to solve problems with continuous action spaces; however, finer room temperature setpoint resolutions are often unnecessary in real buildings and may even lead to conflicting comfort feedback from untrained occupants. Jung and Jazizadeh [73] compared three control setpoint resolutions, i.e., 0.1°C, 0.5°C, and 1°C, and found that finer resolutions did not show any statistically significant impact on the average occupant satisfaction because ordinary occupants often cannot tell slight temperature differences. Accordingly, a relatively coarse resolution of 0.5°C was used in this study.

3.2.3. Reward Design

The reward design is empirically determined to achieve better training convergence rate and balance between thermal comfort and HVAC energy consumption, which is calculated by:

$$\overline{DP} = \sum_{k=1}^{K} \left(c_{ij}^{k} * Pres^{k} \right) / \sum_{k=1}^{K} Pres^{k} \tag{7}$$

$$R = -\begin{cases} \eta * \left(\overline{DP} + \lambda \right)^{\beta} + E_{HVAC} | Occ.Flag = 1 \\ \eta * \left(\overline{DP} + \lambda \right)^{\beta} + \delta * E_{HVAC} | Occ.Flag = 0 \end{cases} \tag{8}$$

Where $\overline{DP}$ denotes the mean discomfort penalty for the occupants who are currently present in the office. K is the number of total occupants. c_{ij}^{k} is the thermal discomfort penalty for the k^{th} occupant at the i^{th} room temperature and the j^{th} fan speed mode, which can be looked up from the personalized comfort matrix. $Pres^{k}$ is the binary presence indicator for the k^{th} occupant, and it is 1 when the corresponding occupant is present. $\overline{DP}$ is computed as the arithmetic mean of the discomfort penalties considering the current occupant presence.. E_{HVAC} is the total HVAC energy consumption (kWh) during the last control time step. The detailed calculation of the energy consumption associated with this experiment space can be found in Section 3.6. η and δ are tunable hyperparameters. Specifically, η balances the relative weights between the energy use and the thermal comfort, and δ penalizes any excessive energy consumption during unoccupied hours when the thermal comfort is not the main concern. Lastly, λ and β are used to re-scale the mean discomfort penalty.

We first tuned η to ensure that the thermal discomfort penalty and energy consumption are of similar magnitude during the occupied hours ($Occ.Flag = 1$). Since $\overline{DP}$ took a much smaller absolute value than E_{HVAC} in this study, η had to be large, and a slight change in $\overline{DP}$ would significantly affect the total reward R. As mentioned in Section 2.3, the discomfort penalties from the personalized comfort matrix take integer values, which intends to align with the standard comfort votes. Consequently, $\overline{DP}$ could fluctuate drastically when the number of present occupants is small. For example, when a participant reports that the indoor environment is 'unacceptable' (-2), the change in the mean penalty value would be -0.67 if three occupants were present and -0.25 if eight occupants were present. This discrepancy could lead to a very unsmooth reward function, making the agent easily get stuck in a local optimum. To deal with it, we found that applying a transformation (with scaling factors λ and β) to $\overline{DP}$ was effective in smoothing out the changes in the rewards. Additionally, β also helped further penalize large discomfort penalties. The hyperparameters of the reward were empirically fine-tuned on the virtual environment.

3.3. Personalized Comfort Matrix Initialization and Online Update

In this section, an example is provided to demonstrate the development of the personalized comfort matrix. The two controllable environment variables are the room temperature and air velocity in this experiment. The proposed personalized comfort matrix takes them as inputs and maps the thermal discomfort level (discomfort penalties) to the RL reward calculation. Following [71], two short comfort questionnaires were used in this study via the Google Form platform. The preliminary comfort survey was carried out to initialize the personalized comfort matrices for the eight participants. The online comfort survey was disseminated every 60 minutes only to the occupants present in the office during the online deployment. The feedback collected was used to update the comfort matrix at the end of the day. The participants were allowed to answer the questionnaires voluntarily, and reimbursements were made upon completion of 75% of them. We recorded the room temperature and fan speed mode at the time when survey responses were received. Since there are no occupancy sensors in this building, we manually detected the occupants' presence every 30 minutes from 10:00 am and 6:00 pm daily. The initialization and online update process are detailed below, as illustrated in Figure 5.

A one-time preliminary comfort survey and the PMV model were used to initialize the personalized comfort matrix for the agent pre-training. As shown in Figure 12a in Appendix A, this 4-question survey collected information about occupants' typical clothing type (Q1), overall thermal preference (Q2), overall air movement preference (Q3), and the general opinion about their current indoor environment (Q4). The 3-step initialization procedures are as follows:

Step 1. The on-site air velocity measurement was taken at different ceiling fan modes. Along with the occupant's clothing type (Q1), PMV values were calculated at different combinations of room temperature and fan mode,

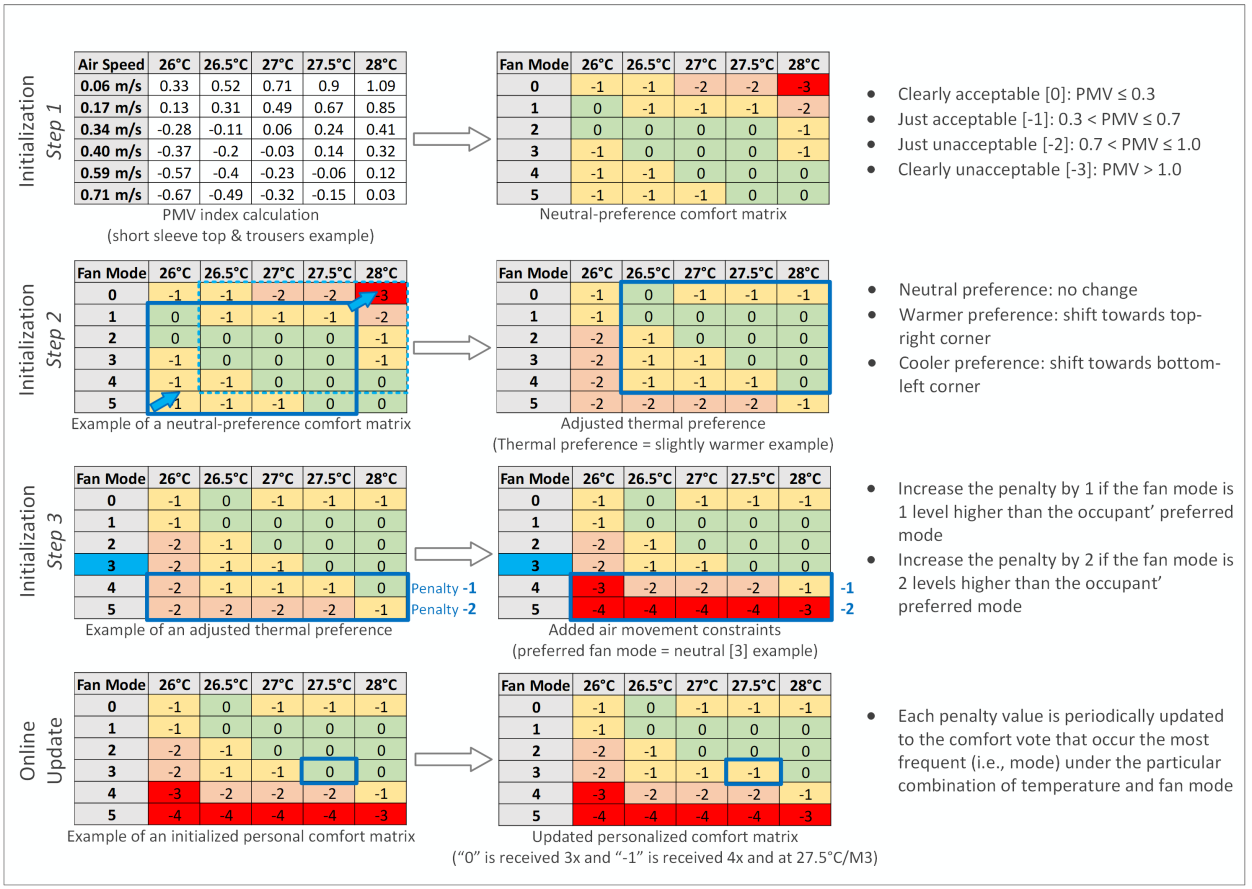


Figure 5: Initialization and online update process for the personalized comfort matrix.

which were then translated into 4-point discomfort penalties. The ASHRAE Standard 55 recommends that the PMV index is within ±0.5 limits. However, it will lead to a wide thermal acceptable range in the comfort matrix. To ensure a higher level of satisfaction, we considered a more stringent limit and started penalizing PMV values over ±0.3. At this point, a neutral-preference personalized comfort matrix was developed.

Step 2. Based on the occupants' perceptions of the current indoor environment (Q4), neutral-preference matrices were shifted towards the top-right corner if they prefer warmer or towards the bottom-left corner if they prefer cooler. This experiment assumed that one scale difference in Q4 can be compensated for by one level of change in both the temperature setpoint and fan mode. In other words, the comfort matrices would be shifted only in the diagonal direction.

Step 3. Fan mode constraints were incorporated to account for the difference in occupants' tolerance for elevated air movement. Specifically, an additional penalty of 1 or 2 will be added for all environmental conditions in the comfort matrices where the fan mode is higher than the occupants' preferred fan mode in Q3. After that, the initialized personalized comfort matrices were prepared.

Another 5-question online comfort survey was employed to collect real-time thermal comfort feedback and to continuously update the personalized comfort matrices (Figure 12b in Appendix A). Q1 asked the time length of stay in the office. Since we only consider the thermal comfort in steady-state conditions, transient comfort votes (occupants sitting for less than 15 minutes) were excluded. Q2 was a standardized question for 7-point Bedford scale comfort descriptors. The online update of the personalized comfort matrices was based on the Q3 and Q4 votes, which sought to understand the occupants' acceptance of their current thermal environment and air movement. Note that occupants may give distinct comfort votes under the same indoor conditions. To alleviate this uncertainty and avoid unstable changes in the control strategy, the actual votes collected were not directly used for the agent training. Instead, each discomfort

Table 3
BDQ agent training setups and hyperparameters

Parameters	Value	Parameters	Value
Network width of 1^{st} hidden layer in share module	256 units	Prioritized replay buffer size	10^3
Network width of 2^{nd} hidden layer in share module	128 units	Virtual environment simulation time step	1 second
Network width of sub-action branches	64 units	Control time step	15 minutes
Network width of state value branch	64 units	Reward discount factor γ	0.9
Nonlinear activation function	Relu	Batch size	64
Optimizer	Adam	Target network update frequency	288
Weights initialization method	Xavier	η in the reward	6000
Learning rate	5×10^{-4}	δ in the reward	10
Gradient clip method	L2-norm	λ in the reward	-0.4
Gradient clip size	10	β in the reward	2

penalty value in the comfort matrix was periodically updated to the comfort vote that occurred the most frequent (i.e., mode) under the particular combination of room temperature and fan mode. In this way, the agent interacted with those personalized comfort matrices rather than the actual real-time votes, resulting in a more stable learning process.

3.4. BDQ-based Agent Training Setup

Prior to any interaction with the actual building, the proposed BDQ agent was pre-trained in the virtual environment with three years (2018, 2019 & 2020) of Singapore's actual meteorological year (AMY) weather data, of which the first two years of data was for training and the remaining year of data was for testing. Table 3 summarized the agent training setups and hyperparameters used. Since the action space was 5-dimensional in this experiment, the BDQ agent had five network branches. The size of each network layer is indicated in Figure 3. To encourage the agent to explore the environment, we sampled actions from a Gaussian distribution with its mean at the greedy actions. Following a linear schedule, the standard deviations started at 0.4 and decreased to 0.05 after the second year.

This experiment involved eight occupants with a total of 2^8 different occupant presence states. However, in practical applications, since some occupancy states (e.g., 100% occupied) would occur much more frequently than others (e.g., 0% occupied), the control performance would not compromise too much if we only focus on the most frequent occupancy states. Furthermore, long-term historical occupancy data is often unavailable in real-world buildings, making it difficult to train offline agents that take high-granularity occupancy data as their state input. Hence, we simplified the control problem by developing multiple DRL policies, each of which was only responsible for a specific occupancy state. This experiment selected the most frequent 16 out of the 128 possible occupancy states based on a short-term occupancy presence observation. At each control time step, one of the 16 trained policies will be put in charge based on the latest occupancy presence detection. Consequently, by eliminating the dependency of the occupancy presence information as the agent's state inputs, the training process involves updating 16 policies in parallel, which significantly reduces the occupancy data required for offline training.

3.5. BDQ-based Agent Online Deployment

The BDQ agent was deployed for two weeks to control the hybrid cooling system. The deployment setups and the hyperparamters were the same as those used in pre-training, except that the control time step was changed to 30 minutes, as the daily temperature in the tropics is relatively stable and our experiment involves a few manual tasks. At each control step, we manually detected the occupant presence states and selected the appropriate BDQ policy to be put in charge. The agent received the latest readings of state variables from the BAS and chose the optimal actions accordingly. The sub-action for temperature setpoint was sent directly back to the BAS, while the sub-action for ceiling fan mode was executed through a local router. The personalized comfort matrices were updated daily at the end of the day based on the online comfort survey votes, and the virtual environment was re-calibrated automatedly once a week using the operating data from the past week. As aforementioned, the model calibration is performed only at the early stage until the reply buffer is filled with measurement data.

Since we only maintained the top 16 BDQ policies, there may be occasional occurrences of occupancy presence states that are not covered. In these cases, we would always use the available policy corresponding to a similar

occupancy state. Indeed, according to the historical data, the selected 16 most frequently occurring occupancy states had already covered the actual occupancy states for 86% of the time.

3.6. Performance Evaluation

The control performance of the proposed DRL framework was evaluated against the baseline control strategy (27°C with fixed fan speed at M3). The performance metrics are the total HVAC energy consumption, total thermal acceptability rate, and total air movement acceptability rate. The total HVAC energy (E_{HVAC}) of the hybrid cooling system consists of the district cooling plant energy, DOAS fan energy, and ceiling fan energy. The DOAS unit that serves the experiment space also supplies outdoor air to other spaces. While the plant and DOAS energy were not metered at the room level, we used Equation 9 and Equation 10 to estimate the energy consumed by this room. Specifically, the cooling plant energy (E_{Plant}) was computed using the enthalpy difference between the room air and supply air (H_{Room} and $H_{SupplyAir}$). The DOAS fan energy (E_{DOAS}) was assumed to be proportional to the ratio of the room's supply airflow (V_{Room}) to the system's total supply airflow (V_{System}). ρ_{Air} denotes the air density at room temperature and η_{Plant} denotes the coefficient of performance (COP) of the chilled water plant.

$$E_{Plant} = \rho_{Air}(H_{Room} - H_{SupplyAir})V_{Room}/\eta_{Plant} \tag{9}$$

$$E_{DOAS} = E_{System}(V_{Room}/V_{System}) \tag{10}$$

The proposed and baseline control deployments were conducted over a consecutive 4-week time period. The baseline control strategy was run three days before, two days in between, and one day after the DRL-based agent deployment to account for any changes in occupants' schedules or the seasonal divergence of thermal satisfaction. In practice, it is infeasible to operate the proposed control and the baseline control simultaneously, making it difficult to compare them under different outdoor weather conditions and occupancy states. Following [11], to ensure a fair comparison, for each day of the DRL control, we took the energy consumption data and compared it to the baseline day that had the most similar average outdoor temperature and solar radiation during the occupied hours.

4. Results

4.1. Training with the Virtual Environment

Figure 6 compares the convergence of the BDQ and the Double DQN algorithm in the pre-training. The BDQ agent achieved training convergence in less than a year, which was significantly faster than the time duration of the conventional DRL algorithm, indicating that the BDQ agent handled high-dimensional action spaces more efficiently in this multivariate control task. Note that rewards fluctuated between episodes, especially in the early phase. This fluctuation is because not only is the reward highly related to the outdoor weather, but random actions (determined by a Gaussian distribution) were taken over some timesteps to encourage the agent to explore the environment. In particular, the exploration probability was linearly annealed during the training.

Based on preliminary sensitivity analysis, four parameters were selected to calibrate the virtual environment: floor thermal conductivity, leakage area, internal thermal mass, and effective window area. Root mean squared error (RMSE) of room temperature at 15-minute resolution was adopted as the objective function. The automated model re-calibration took place after the first week of deployment, during which the RMSE was reduced from 0.91°C to 0.41°C, indicating a more representative thermal dynamics. Figure 7 compares the predicted thermal responses obtained before and after the GP-BO based calibration. As per ASHRAE Guideline 14 [74], the re-calibrated model had a Mean Bias Error (MBE) and Coefficient of Variation of the Root Mean Squared Error (CVRMSE) of 0.7% and 1.5%, respectively.

4.2. Analysis of Control Actions for a Typical Day

Figure 8 presents an overview of the 5-dimensional control actions and the corresponding occupant presence for a typical day. Four different colors were used to map the occupants to their closest ceiling fans, as shown in the bottom part of the figure. It is apparent that the BDQ agent tried to maximize the rewards by pushing the room temperature up to around 28°C. To further reduce the energy consumption, the agent used the setback setpoint for the minimum outdoor airflow rate, allowing the room temperature to float towards the end of the occupancy. This controlling behavior

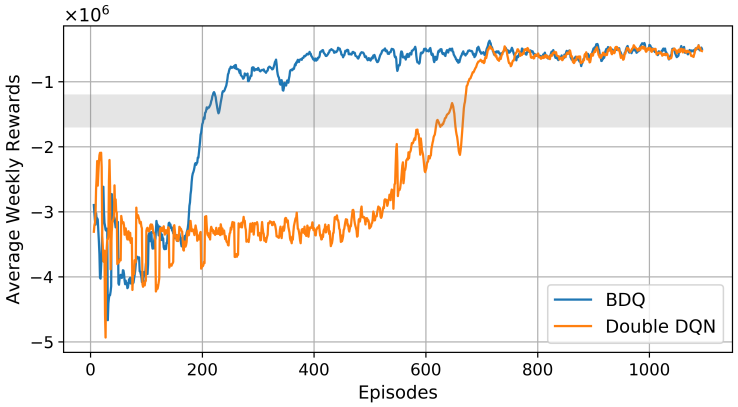


Figure 6: Comparison of the convergence performance between BDQ and Double DQN. The reward for the baseline control strategy is shown in the gray shaded region. Each natural day is an episode. To match the baseline performance, it took 7 months in simulation for the BDQ and 21 months in simulation for the Double DQN.

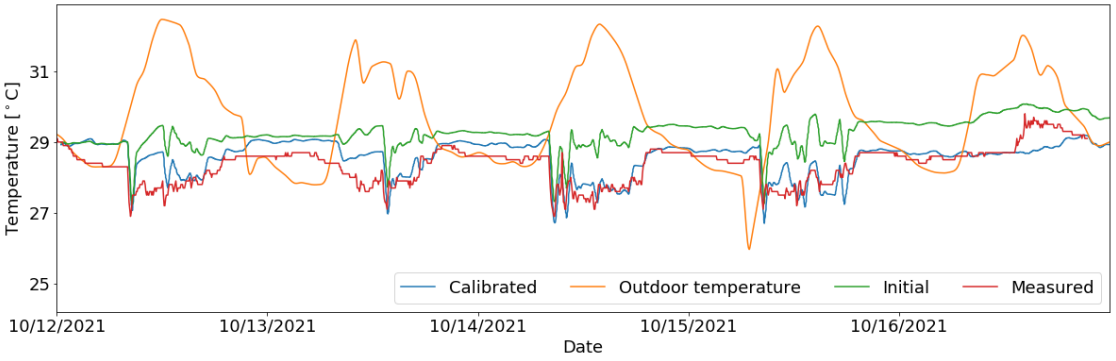


Figure 7: Predicted room temperature on the first week of experiment using the virtual environment. Due to a missing data issue on the first day (10/11/2021), we included the following Saturday (10/16/2021) data into the calibration.

happened almost every day. The length of time to use the setback setpoint was usually 30 or 60 minutes and was highly dependent on outdoor conditions. Through the interactions with the high-fidelity virtual environment, the agent appeared to learn that using the setback setpoint can maximize the reward. This is because the energy consumption is minimal at this setpoint, and the room temperature does not rise too quickly to cause discomfort due to the thermal lag. Additionally, the room temperature fluctuated by about 0.3°C throughout the day. Given that this is an open-plan office, this discrepancy could be attributed to the constant movement of occupants, the opening and closing of doors, or the sensor measurement errors.

Depending on the occupants' presence, the ceiling fans' mode was dynamically adjusted throughout the day to compensate for the elevated room temperature. Figure 13 presents the initial and final personalized comfort matrix for each participant is presented. Occupant 2 had a warm preference; thus, fan 2 was operated under either M0 or M1 in the morning and early afternoon when this occupant was alone under the fan. Since occupant 1 and 2 shared the same fan and occupant 1 had a neutral preference, fan 2 was increased to M2 after occupant 1 arrived at the office around 16:00. Additionally, the agent temporarily turned the temperature setpoint down to 27.5°C at 13:00 upon the arrival of occupant 4, who preferred a neutral to cool environment. The setpoint was raised again to 28°C as soon as occupant 2 returned around 13:30. Fan 3 and fan 4 were operated at M3 most of the time because both occupant 7 and 8 had similar neutral to cool thermal preferences. Note that occupant 3 and 6 were not present at the office on this day. These control actions suggest that the BDQ agent is capable of dynamically choosing appropriate actions based on occupant-specific thermal comfort.

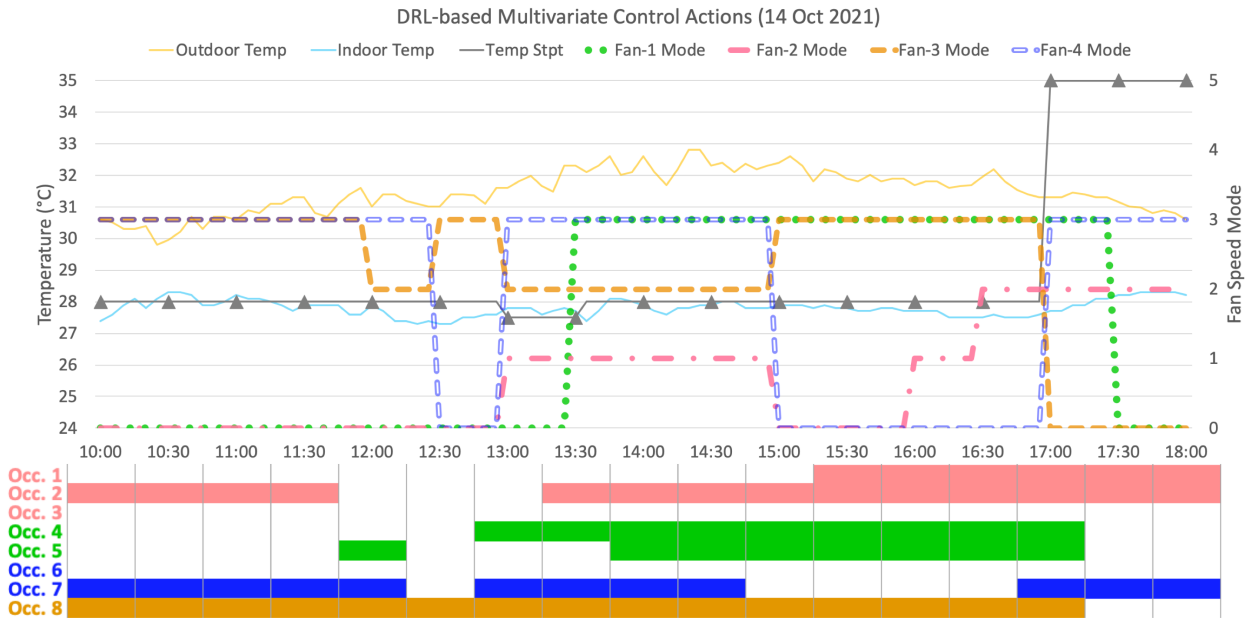


Figure 8: Visualization of control actions for a typical day. The solid gray line with triangular markers displays the temperature setpoints throughout the day. Ceiling fan 1 (green dotted line) served occupants 4 and 5, ceiling fan 2 (dash-dotted pink line) served occupants 1, 2 and 3, ceiling fan 3 (dashed orange line) served occupants 8, and ceiling fan 4 (dashed blue line) served occupants 6 and 7.

4.3. Comfort Performance Evaluation based on Real-time Votes

Before online learning begins, we compared the pre-training performance of the proposed BDQ agent with two benchmarks in terms of the thermal and air movement acceptability as shown in Figure 9. In addition to the baseline control, `RL+PMV` is also considered because the PMV is widely used as the indicator of thermal comfort in RL studies on building control. `RL Day 1` is the first day of deployment of the proposed BDQ agent in the actual building prior to any online update. Hence, the difference in control performance between `RL Day 1` and `RL+PMV` can only be caused by the different thermal comfort information used for offline training. Although the baseline strategy achieved a rather good total thermal acceptability rate of 88%, the proposed BDQ agent slightly outperformed it on day 1. More importantly, training the agent using the PMV index as the thermal comfort metric led to the worst control performance. Initializing the comfort matrix with just a one-time comfort survey improves the total thermal acceptability by 25% over the PMV method. This result is expected because the standard PMV calculation does not consider any air movement limits. Consequently, the agent pushed the energy reduction too aggressively by using the maximum temperature and fan speed available at all times. Consistent with previous literature, this finding supports that the PMV index is often insufficient for individuals because thermal neutrality is not the same as occupant satisfaction. Thanks to the one-time preliminary comfort survey, the differences in personal thermal comfort preferences can be captured prior to the deployment, which greatly facilitates the learning process.

411 valid real-time comfort votes were collected during the experiment period. Figure 10 illustrates the thermal and air movement acceptability rate during the online deployment. Due to the small daily sample size, we combined the votes for the second week (`RL Day 6` to `RL Day 10`) of the DRL control for a fair comparison with the baseline. The total thermal acceptability rate was improved by 11% to 99%. Additionally, there was a gradual rise in the ratio of 'Clearly Acceptable' votes for both thermal and air movement performance. The pink line in Figure 10 indicates the p-value of the Chi-square test for the daily 'Clearly Acceptable' rate compared to the one in the baseline. The results show that the proposed BDQ agent achieved a 31% increase in clear thermal acceptability rate and the improvement was statistically significant with 95% confidence since day 5 of the deployment. Furthermore, a strong positive correlation ($r(8) = .88$, $p < .01$) was observed between the thermal and air movement acceptance votes, which implies that the air movement is a critical factor in the overall thermal comfort of this hybrid cooling system.

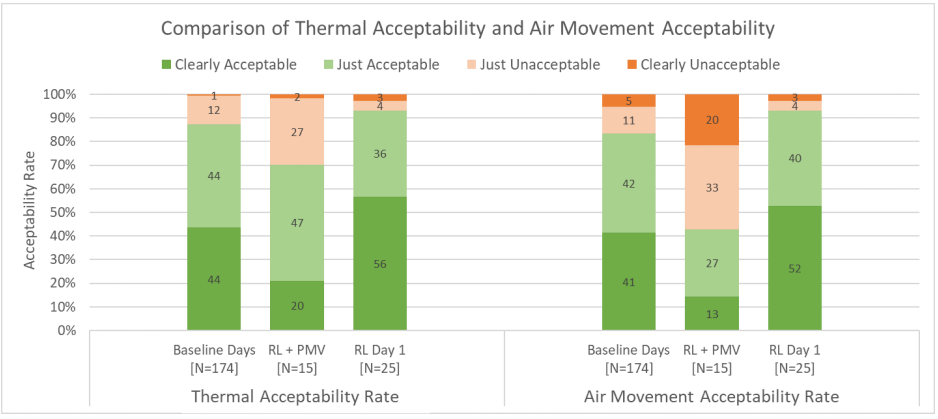


Figure 9: Comparison of comfort results prior to the online training. `Baseline Days` denotes the aggregate of all experiment days operated under the baseline strategy (27°C and fan speed M3). `RL+PMV` denotes the BDQ agent pre-trained with the standard PMV model. `RL Day 1` denotes the BDQ agent pre-trained with the personalized comfort matrix.

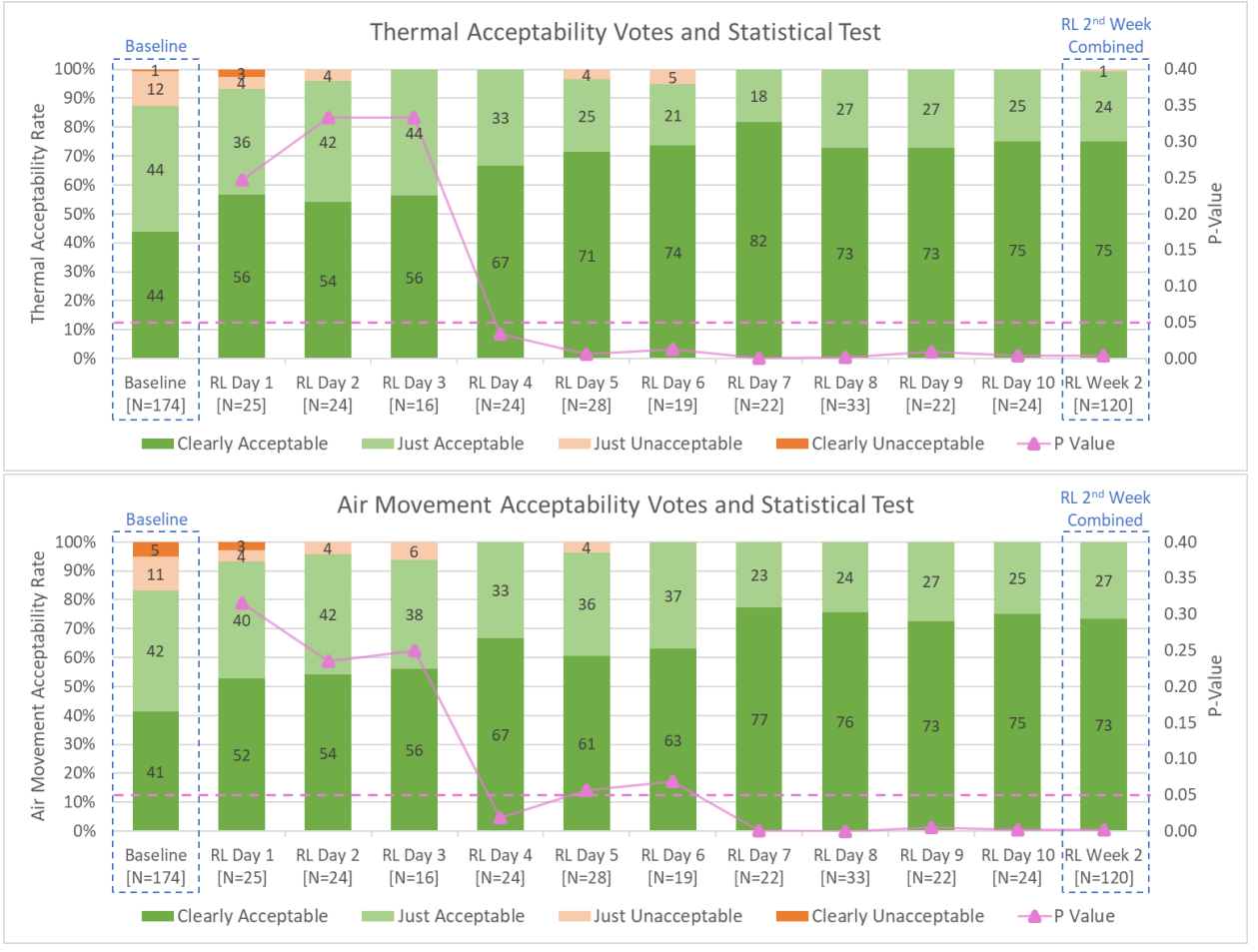


Figure 10: Comparison of comfort results during the online deployment. `RL Day 1` to `RL Day 10` denote the respective experiment days controlled by the proposed BDQ agent. `RL Week 2` denotes the aggregate of the BDQ agent deployment days in the second week. The number of valid comfort votes is also indicated.

4.4. Energy Performance Evaluation Results

Table 4 summarizes the daily energy consumption and the corresponding outdoor weather information. During the experimental period from 10:00 am to 6:00 pm, the mean normalized baseline energy consumption was 15.3 kWh. By contrast, the BDQ agent consumed 13.1 kWh of HVAC energy per day on average, which yielded 13.9% of HVAC energy reduction. Given that the baseline control strategy in the SDE4 building has already achieved net-zero energy, the energy performance of the proposed control framework is very significant. Additionally, the ceiling fans consumed less than 3% of the total HVAC energy in both cases, indicating that raising room temperature and compensating for it by increasing air movement is a very promising energy-efficient measure without compromising the thermal comfort.

Closer inspection of Table 4 shows that there is a significant positive correlation ($r(10) = .87$, $p < .01$) between the HVAC energy consumption and the mean solar radiation. Since there is not much daily variation in mean outdoor temperature in Singapore's climate, its outdoor solar radiation greatly affects the daily mean HVAC energy consumption. By further analyzing the hourly cooling demand profile, we observed that the hotter the outdoor weather, the more energy reduction could be achieved through the DRL control. However, when the outdoor temperature was below 28°C and solar radiation was below 300 W/m^2, the DOAS unit mostly operated at its lowest airflow rate, regardless of the temperature setpoint used. Lastly, monday (baseline day 1 and RL day 4 & 9) tended to consume more energy than other weekdays. The difference was due to the thermal mass of the building that built up heat over the weekend, increasing the cooling demand on Monday.

Table 4
Comparison of energy consumption between baseline and BDQ agent.

Day	Number of survey responses	Mean solar radiation (Watt/m^2)	Mean outdoor temperature (°C)	Mean indoor temperature (°C)	Energy Consumption (kWh)			
					Plant	PAHU	Ceiling fans	Total HVAC energy
Baseline control								
1	36	331.7	29.8	26.8	10.2	4.3	0.2	14.7
2	37	658.4	31.3	26.9	9.2	4.0	0.2	13.4
3	32	549.9	30.9	27.0	9.5	4.0	0.4	13.9
4	18	260.8	29.8	27.1	13.0	5.9	0.2	19.1
5	32	530.2	30.8	27.1	13.4	6.1	0.4	19.9
6	18	168.7	27.3	27.1	6.5	4.6	0.3	11.4
Norm.*	31	500.5	30.5	27.0	10.3	4.7	0.3	15.3
DRL-based control								
1	25	713.5	31.2	27.8	10.3	4.1	0.2	14.6
2	24	599.2	31.7	28.0	9.1	3.5	0.4	13
3	16	519.5	31.0	28.1	9.5	3.5	0.2	13.2
4	24	588.6	31.5	27.8	10.3	4.0	0.2	14.5
5	28	683.3	31.3	28.0	10.6	4.2	0.4	15.2
6	19	84.4	25.3	28.0	7.6	2.4	0.2	10.2
7	22	485.6	30.1	27.8	8.3	2.6	0.4	11.3
8	33	433.5	29.9	27.7	8.8	2.7	0.4	11.9
9	22	553.5	30.5	27.9	10.2	3.6	0.2	14
10	24	495.0	32.1	27.9	9.6	3.6	0.2	13.4
Mean**	24	515.6	30.4	27.9	9.4	3.4	0.3	13.1

* The mean energy consumption of the baseline control deployment days after performing the weather normalization.
** The mean energy consumption of the BDQ agent deployment days.

4.5. Analysis of Online Update for Personalized Comfort Matrix

As shown in Figure 13, the differences between the initial and final personalized comfort matrices are generally small, except for occupant 1 and 2. The comfort matrix for occupants 3 and 8 has not even changed. These results suggest that the proposed simple comfort matrix initialized with the one-time survey effectively captured most of the personal comfort information in this experiment.

5. Discussion

5.1. Advantage of BDQ-based Agent

According to the learning curve shown in Figure 6, both the BDQ agent and the conventional DRL agent eventually reached a similar level of reward, indicating that they converged to similar final performance levels for this task. However, the BDQ agent significantly outperformed its counterpart in terms of sample-efficiency. Specifically, the BDQ agent not only found the path to accelerate the learning progress earlier but also learned at a faster rate with a steeper learning curve. This difference may become more pronounced when the action dimensionality is further increased. Since there are always discrepancies between simulated and actual environments, and the thermal dynamics of actual environments may change over time, sample-efficient agents will provide better adaptive performance during online training. Moreover, we observed that the BDQ agent converged to the optimal policy in a stable manner. A possible explanation for this might be that the use of the shared decision module enabled a form of centralized coordination.

5.2. Mismatch between Thermal Acceptability and Thermal Comfort

The online comfort survey included both the thermal acceptability and the Bedford thermal comfort scales. Interestingly, there was a discrepancy between the occupants' responses to the two questions. According to the dedicated thermal comfort studies [75], the thermal comfort votes falling in the categories of 'Comfortable', 'Comfortably Cool/Warm', or slightly 'Cool/Warm' should be perceived as thermal acceptable. As can be seen from Figure 11, the thermal comfort votes received throughout the BDQ deployment period except for `RL Day 1` were within the acceptable categories. However, this observation was not in good agreement with the votes obtained by the thermal acceptability scale (as shown in Figure 10), and the discrepancy is more significant if we compare the rate of 'Comfortable' and 'clearly acceptable'. The reason for this is unclear but may be related to the elevated airflow. Additionally, some participants may be unfamiliar with the definitions of terms used in the questionnaire. It is important to note that this study considers total thermal acceptability as the comfort performance metric. However, if we evaluate the comfort performance by the Bedford scale, the improvement due to the proposed DRL control was 16% (77% to 93%).

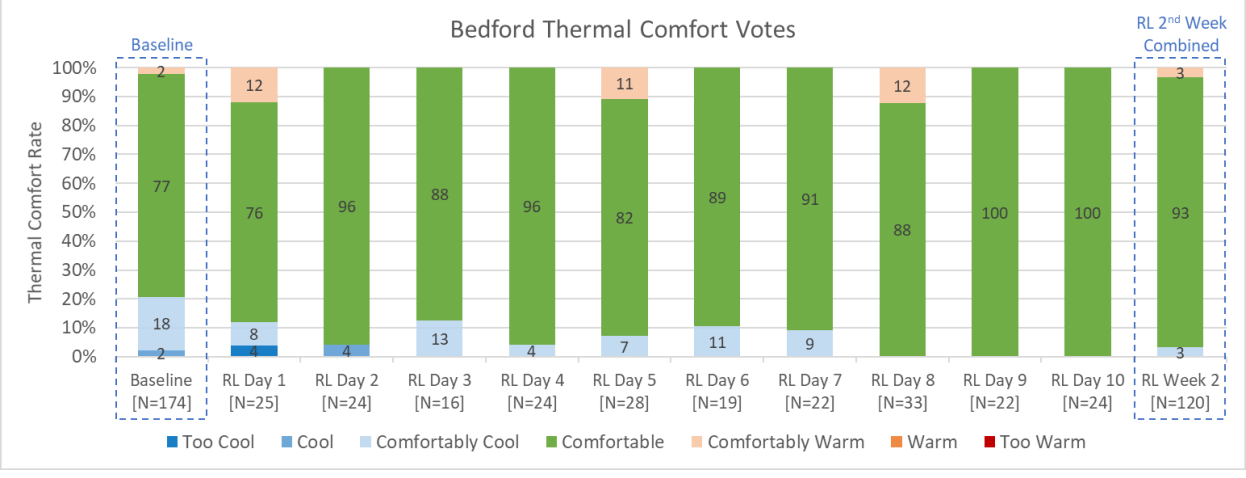


Figure 11: Bedford thermal comfort votes during the online deployment.

5.3. Justification of Energy Reduction

The 13.9% HVAC energy reduction was determined by the measured data. To consider a wider range of weather conditions, we also compared the annual energy consumption between the proposed BDQ agent and the baseline in the virtual environment. We applied the same schedules as in the 4-week experiment period for the occupancy, lighting and plug load to the simulation model. Additionally, the discomfort penalty values from the final personalized comfort matrices on day 10 were used to mimic the real-time comfort votes for the control. Hence, the comfort matrices were not updated throughout the simulation. Based on the aforementioned assumptions, the estimated HVAC energy savings

were 9.5% for Singapore's 2020 AMY weather dataset, which was less than the experiment results. This inconsistency could be due to potential errors that occurred during weather normalization calculations. Another possible explanation for this may be the mismatch between the actual and the simulated environment despite the use of advanced model calibration techniques. The model assumes well-mixed air, which is often not the case in reality, mainly because occupants and electrical equipment are unevenly distributed across the space. We speculate that since the exhaust from the experimental room depends entirely on pressure difference (no mechanical exhaust in the system), the higher supply airflow rate in the baseline control may lead to a short air circuit and higher leakage rate, resulting in wasted cooling energy.

5.4. Importance of Considering Personalized Comfort in HVAC Control

The findings presented in Section 4.3 are in agreement with those obtained by previous thermal comfort research that controlling buildings based on average occupants is insufficient for comfort [50, 47]. However, personalized comfort is often not considered in studies on advanced HVAC controls. In this experimental setup that allowed for increased airflow, a simple one-time comfort survey significantly helped improve the thermal control performance compared to using the PMV index alone. Therefore, it is worth exploring how to integrate personalized comfort information in the control loop to facilitate the practical application of advanced HVAC controls in buildings. The proposed tabular-based personal comfort modeling method with periodical updates showed a promising thermal acceptability result, which has the potential to be used in a broader range of applications. Further research should be undertaken to investigate the robustness and generalization of this method.

5.5. Necessity of a Virtual Environment for Agent Pre-training

Virtual environments are useful for training the RL agent due to the realism and flexibility that they offer. There is no doubt that high-fidelity building models are expensive to create. On the one hand, they usually require extensive domain knowledge about the dynamics of complex building systems. On the other hand, high granularity data needed to calibrate or identify those models is often difficult to acquire. A previous study showed that RL agents can be pre-trained using imitation learning, thus eliminating the need for a high-fidelity model [43]. However, a number of studies have shown that the performance of RL algorithms is very sensitive to numerous hyperparameters [76, 77], such as the learning rate, reward decay factor, and network architecture. Without a proper virtual environment to pre-train the agent offline, fine-tuning these hyperparameters only during online deployment can be significantly inefficient and even lead to catastrophic consequences due to unsafe exploration.

Furthermore, Wang and Hong [38] pointed out that simulation models do not have to be very accurate because the pre-trained policy will be further improved when interacting with the actual environment. However, more research should be undertaken to investigate what are the minimum performance requirements for a satisfactory virtual environment.

5.6. Potential Implementation Strategy in Practice

The proposed DRL based occupant-centric control framework has demonstrated promising control performance in the experiment setting. However, its online learning phase heavily relied on the high temporal granularity comfort survey (i.e., every 60 minutes), which is challenging to carry out in practice. By analyzing the daily changes in the personalized comfort matrices, we found that 5 of the 8 comfort matrices had few new updates after the 4th day of the online deployment, as the corresponding occupants voted for 'Clearly Acceptable' for most of the time. Additionally, there were no updates for all the matrices on the 10th day. This implies that frequent comfort surveys may not be necessary for most occupants. Therefore, instead of using a fixed survey schedule, we can consider asking occupants to voluntarily submit comfort surveys only when they feel uncomfortable. This can be done by placing a QR (quick response) code linked to the comfort questionnaire on their desks. Alternatively, we could adopt the sparse sampling technique to develop a parsimonious survey schedule [78]. As wearables are becoming ubiquitous, it will be easier to collect personal comfort data in natural environments [79]. In both ways, the frequency of the need for comfort surveys would be significantly reduced, making practical implementations of the proposed control framework feasible and scalable in occupants' natural office environments.

5.7. Limitations

The main weakness of this study was the lack of long-term demonstration of the agent's online training performance directly using the BAS data. Since this study relied on repeated comfort surveys and survey fatigue started to appear at

the end of our experiment, the agent was deployed online for only two weeks. Consequently, the online agent training was based on the periodically re-calibrated simulation model. Another potential problem is that occupant presence data was required at a high granularity in this control framework, which is expensive and difficult to obtain. We assumed that the occupancy presence was perfectly known by using manual observations in this experiment, while fault-free occupancy sensing is rare in real buildings. Last but not least, maintaining multiple BDQ policies is inefficient and may lead to sub-optimal performance because these policies are trained independently and a few policies are updated more frequently than others. Consequently, switching between policies may lead to unstable or inconsistent control strategies if the occupancy state changes frequently.

6. Conclusion and Future Work

This study proposes a practical DRL-based framework for multivariate occupant-centric HVAC control. Our approach enables a linear growth of the total number of network outputs with the increasing HVAC action dimensionality, which improves the learning efficiency for controlling HVAC systems consisting of multiple subsystems. Additionally, the personalized thermal comfort matrix that maps the discomfort level onto the RL reward is presented, which can be easily integrated into human-in-the-loop building operations. The proposed agent was deployed in an actual office to validate the control performance, dynamically controlling its room temperature setpoint and ceiling fans' speed mode.

Compared to the baseline optimal static control strategy, the proposed agent achieved a 13.9% cooling energy reduction and an 11% total thermal acceptability improvement based on the deployment results. The offline training in simulation showed that the BDQ-based agent was significantly more sample-efficient than the standard Double DQN algorithm. More importantly, it has been revealed that initializing the comfort matrix based on a one-time comfort survey is effective for capturing individual differences in thermal comfort, as this brought 25% thermal acceptability improvement over the PMV method to the proposed control framework.

This is the first DRL-based multivariate HVAC control study that has been implemented in a real building. It highlights the importance of considering individual differences in the thermal comfort for building operations. As future work, to improve the scalability and robustness of the control framework, we will explore effective approaches to encode occupant presence information in state inputs instead of maintaining multiple policies. Besides, since developing a high-fidelity virtual environment is expensive and an overly accurate model may not provide obvious benefits for the agent pre-training, further investigation is needed to determine the minimum performance requirements for a satisfactory virtual environment. Moreover, this study employed the automated model calibration to address the discrepancies between the simulated and real environments. It is worth investigating more sim-to-real transfer techniques for efficient and generalized policy transfer. Last but not least, it is necessary to extend the BDQ agent to multi-zone building controls using cooperative multi-agent techniques.

Acknowledgement

This research is supported by Kajima Corporation through its Kajima Technical Research Institute Singapore (KaTRIS) (Project No. A-0008298-00-00).

CRediT authorship contribution statement

Yue Lei: Conceptualization, Methodology, Data Curation, Formal analysis, Investigation, Visualization, Writing - Original Draft, Writing - Review & Editing. **Sicheng Zhan:** Software, Visualization, Writing - Review & Editing. **Eikichi Ono:** Validation, Writing - Review & Editing. **Yuzhen Peng:** Writing - Review & Editing. **Zhiang Zhang:** Writing - Review & Editing. **Takamasa Hasama:** Project administration. **Adrian Chong:** Conceptualization, Supervision, Project administration, Funding acquisition, Writing - Review & Editing.

References

[1] B. P. Center, Annual energy outlook 2020, Energy Information Administration, Washington, DC 12 (2020) 1672–1679.

[2] E. Commission, A renovation wave for europe—greening our buildings, creating jobs, improving lives, 2020.

[3] N. E. Fernandez, S. Katipamula, W. Wang, Y. Xie, M. Zhao, C. D. Corbin, Impacts of commercial building controls on energy savings and peak load reduction, Technical Report, Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2017. doi:10.2172/1400347.

[4] M. Frontczak, S. Schiavon, J. Goins, E. Arens, H. Zhang, P. Wargocki, Quantitative relationships between occupant satisfaction and satisfaction aspects of indoor environmental quality and building design, Indoor air 22 (2012) 119–131. doi:10.1111/j.1600-0668.2011.00745.x.

[5] S. Wang, Z. Ma, Supervisory and optimal control of building hvac systems: A review, Hvac&R Research 14 (2008) 3–32. doi:10.1080/10789669.2008.10390991.

[6] 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, Building and Environment 165 (2019) 106351. doi:10.1016/j.buildenv.2019.106351.

[7] J. Xie, H. Li, C. Li, J. Zhang, M. Luo, Review on occupant-centric thermal comfort sensing, predicting, and controlling, Energy and Buildings 226 (2020) 110392. doi:10.1016/j.enbuild.2020.110392.

[8] X. Dai, J. Liu, X. Zhang, A review of studies applying machine learning models to predict occupancy and window-opening behaviours in smart buildings, Energy and Buildings 223 (2020) 110159. doi:10.1016/j.enbuild.2020.110159.

[9] Z. Afroz, G. Shafiullah, T. Urmee, G. Higgins, Modeling techniques used in building hvac control systems: A review, Renewable and sustainable energy reviews 83 (2018) 64–84. doi:10.1016/j.rser.2017.10.044.

[10] B. Dong, K. P. Lam, A real-time model predictive control for building heating and cooling systems based on the occupancy behavior pattern detection and local weather forecasting, in: Building Simulation, volume 7, Springer, 2014, pp. 89–106. doi:10.1007/s12273-013-0142-7.

[11] D. A. Winkler, A. Yadav, C. Chitu, A. E. Cerpa, Office: Optimization framework for improved comfort & efficiency, in: 2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), IEEE, 2020, pp. 265–276. doi:10.1109/IPSN48710.2020.00030.

[12] N. K. Dhar, N. K. Verma, L. Behera, Adaptive critic-based event-triggered control for hvac system, IEEE Transactions on Industrial Informatics 14 (2017) 178–188. doi:10.1109/TII.2017.2725899.

[13] Y. Peng, A. Rysanek, Z. Nagy, A. Schlüter, Using machine learning techniques for occupancy-prediction-based cooling control in office buildings, Applied energy 211 (2018) 1343–1358. doi:10.1016/j.apenergy.2017.12.002.

[14] Y. Peng, Z. Nagy, A. Schlüter, Temperature-preference learning with neural networks for occupant-centric building indoor climate controls, Building and Environment 154 (2019) 296–308. doi:10.1016/j.buildenv.2019.01.036.

[15] J. Arroyo, C. Manna, F. Spiessens, L. Helsen, Reinforced model predictive control (rl-mpc) for building energy management, Applied Energy 309 (2022) 118346. doi:10.1016/j.apenergy.2021.118346.

[16] 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. doi:10.1145/3061639.3062224.

[17] H. Satyavada, S. Baldi, An integrated control-oriented modelling for hvac performance benchmarking, Journal of Building Engineering 6 (2016) 262–273. doi:10.1016/j.jobe.2016.04.005.

[18] R. Z. Homod, K. S. Gaeid, S. M. Dawood, A. Hatami, K. S. Sahari, Evaluation of energy-saving potential for optimal time response of hvac control system in smart buildings, Applied Energy 271 (2020) 115255. doi:10.1016/j.apenergy.2020.115255.

[19] G. Serale, M. Fiorentini, A. Capozzoli, D. Bernardini, A. Bemporad, Model predictive control (mpc) for enhancing building and hvac system energy efficiency: Problem formulation, applications and opportunities, Energies 11 (2018) 631. doi:10.3390/en11030631.

[20] S. Zhan, A. Chong, Data requirements and performance evaluation of model predictive control in buildings: A modeling perspective, Renewable and Sustainable Energy Reviews 142 (2021) 110835. doi:10.1016/j.rser.2021.110835.

[21] A. Chong, Y. Gu, H. Jia, Calibrating building energy simulation models: A review of the basics to guide future work, Energy and Buildings 253 (2021) 111533. doi:10.1016/j.enbuild.2021.111533.

[22] 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 Transactions on Smart Grid 12 (2020) 407–419. doi:10.1109/TSG.2020.3011739.

[23] V. Hanumaiah, S. Genc, Distributed multi-agent deep reinforcement learning framework for whole-building hvac control, arXiv preprint arXiv:2110.13450 (2021). doi:10.48550/arXiv.2110.13450.

[24] E. Mocanu, D. C. Mocanu, P. H. Nguyen, A. Liotta, M. E. Webber, M. Gibescu, J. G. Slootweg, On-line building energy optimization using deep reinforcement learning, IEEE transactions on smart grid 10 (2018) 3698–3708. doi:10.1109/TSG.2018.2834219.

[25] J. R. Vázquez-Canteli, S. Dey, G. Henze, Z. Nagy, Citylearn: Standardizing research in multi-agent reinforcement learning for demand response and urban energy management, arXiv preprint arXiv:2012.10504 (2020). doi:10.48550/arXiv.2012.10504.

[26] R. Z. Homod, H. Togun, A. K. Hussein, F. N. Al-Mousawi, Z. M. Yaseen, W. Al-Kouz, H. J. Abd, O. A. Alawi, M. Goodarzi, O. A. Hussein, Dynamics analysis of a novel hybrid deep clustering for unsupervised learning by reinforcement of multi-agent to energy saving in intelligent buildings, Applied Energy 313 (2022) 118863. doi:10.1016/j.apenergy.2022.118863.

[27] S. Nagarathinam, V. Menon, A. Vasan, A. Sivasubramaniam, Marco-multi-agent reinforcement learning based control of building hvac systems, in: Proceedings of the Eleventh ACM International Conference on Future Energy Systems, 2020, pp. 57–67. doi:10.1145/3396851.3397694.

[28] T. T. Nguyen, N. D. Nguyen, S. Nahavandi, Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications, IEEE transactions on cybernetics 50 (2020) 3826–3839. doi:10.1109/TCYB.2020.2977374.

[29] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning, arXiv preprint arXiv:1509.02971 (2015). doi:10.48550/arXiv.1509.02971.

[30] B. Sun, P. B. Luh, Q.-S. Jia, B. Yan, Event-based optimization within the lagrangian relaxation framework for energy savings in hvac systems, IEEE Transactions on Automation Science and Engineering 12 (2015) 1396–1406. doi:10.1109/TASE.2015.2455419.

[31] X. Ding, W. Du, A. Cerpa, Octopus: Deep reinforcement learning for holistic smart building control, in: Proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation, 2019, pp. 326–335. doi:10.1145/3360322.3360857.

[32] 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 of Things Journal (2021). doi:10.1109/JIOT.2021.3078462.

[33] H. Kazmi, F. Mehmood, S. Lodeweyckx, J. Driesen, Gigawatt-hour scale savings on a budget of zero: Deep reinforcement learning based optimal control of hot water systems, Energy 144 (2018) 159–168. doi:10.1016/j.energy.2017.12.019.

[34] J. Y. Park, T. Dougherty, H. Fritz, Z. Nagy, Lightlearn: An adaptive and occupant centered controller for lighting based on reinforcement learning, Building and Environment 147 (2019) 397–414. doi:10.1016/j.buildenv.2018.10.028.

[35] G. Dulac-Arnold, N. Levine, D. J. Mankowitz, J. Li, C. Paduraru, S. Gowal, T. Hester, Challenges of real-world reinforcement learning: definitions, benchmarks and analysis, Machine Learning 110 (2021) 2419–2468. doi:10.1007/s10994-021-05961-4.

[36] Y. Chen, L. K. Norford, H. W. Samuelson, A. Malkawi, Optimal control of hvac and window systems for natural ventilation through reinforcement learning, Energy and Buildings 169 (2018) 195–205. doi:10.1016/j.enbuild.2018.03.051.

[37] M. Botvinick, S. Ritter, J. X. Wang, Z. Kurth-Nelson, C. Blundell, D. Hassabis, Reinforcement learning, fast and slow, Trends in cognitive sciences 23 (2019) 408–422. doi:10.1016/j.tics.2019.02.006.

[38] Z. Wang, T. Hong, Reinforcement learning for building controls: The opportunities and challenges, Applied Energy 269 (2020) 115036. doi:10.1016/j.apenergy.2020.115036.

[39] N. Andrews, E. Miller, B. Taylor, R. Lingam, A. Simmons, J. Stowe, P. Waight, Recall bias, mmr, and autism, Archives of disease in childhood 87 (2002) 493–494. doi:10.1136/adc.87.6.493.

[40] S. Katipamula, M. R. Brambley, Methods for fault detection, diagnostics, and prognostics for building systems—a review, part i, Hvac&R Research 11 (2005) 3–25. doi:10.1080/10789669.2005.10391123.

[41] Y. Bae, S. Bhattacharya, B. Cui, S. Lee, Y. Li, L. Zhang, P. Im, V. Adetola, D. Vrabie, M. Leach, et al., Sensor impacts on building and hvac controls: A critical review for building energy performance, Advances in Applied Energy 4 (2021) 100068. doi:10.1016/j.adapen.2021.100068.

[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 and Buildings 199 (2019) 472–490. doi:10.1016/j.enbuild.2019.07.029.

[43] B. Chen, Z. Cai, M. Bergés, Gnu-rl: A precocial reinforcement learning solution for building hvac control using a differentiable mpc policy, in: Proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation, 2019, pp. 316–325. doi:10.1145/3360322.3360849.

[44] S. Qiu, Z. Li, D. Fan, R. He, X. Dai, Z. Li, Chilled water temperature resetting using model-free reinforcement learning: Engineering application, Energy and Buildings 255 (2022) 111694. doi:10.1016/j.enbuild.2021.111694.

[45] C. Zhang, Z. Zhang, V. Loftness, Bio-sensing and reinforcement learning approaches for occupant-centric control, ASHRAE Transactions 125 (2019).

[46] S. Jung, J. Jeoung, T. Hong, Occupant-centered real-time control of indoor temperature using deep learning algorithms, Building and Environment (2021) 108633. doi:10.1016/j.buildenv.2021.108633.

[47] W. O'Brien, A. Wagner, M. Schweiker, A. Mahdavi, J. Day, M. B. Kjærgaard, S. Carlucci, B. Dong, F. Tahmasebi, D. Yan, et al., Introducing iea ebc annex 79: Key challenges and opportunities in the field of occupant-centric building design and operation, Building and Environment 178 (2020) 106738. doi:10.1016/j.buildenv.2020.106738.

[48] Z. Wang, R. de Dear, M. Luo, B. Lin, Y. He, A. Ghahramani, Y. Zhu, Individual difference in thermal comfort: A literature review, Building and Environment 138 (2018) 181–193. doi:10.1016/j.buildenv.2018.04.040.

[49] E. Ono, K. Mihara, K. P. Lam, A. Chong, The effects of a mismatch between thermal comfort modeling and hvac controls from an occupancy perspective, Building and Environment (2022) 109255. doi:10.1016/j.buildenv.2022.109255.

[50] J. Kim, S. Schiavon, G. Brager, Personal comfort models–a new paradigm in thermal comfort for occupant-centric environmental control, Building and Environment 132 (2018) 114–124. doi:10.1016/j.buildenv.2019.106351.

[51] T. Cheung, S. Schiavon, T. Parkinson, P. Li, G. Brager, Analysis of the accuracy on pmv–ppd model using the ashrae global thermal comfort database ii, Building and Environment 153 (2019) 205–217. doi:10.1016/j.buildenv.2019.01.055.

[52] R. De Dear, G. S. Brager, Developing an adaptive model of thermal comfort and preference (1998).

[53] M. A. Humphreys, J. F. Nicol, The validity of iso-pmv for predicting comfort votes in every-day thermal environments, Energy and buildings 34 (2002) 667–684. doi:10.1016/S0378-7788(02)00018-X.

[54] J. Y. Park, Z. Nagy, Comprehensive analysis of the relationship between thermal comfort and building control research-a data-driven literature review, Renewable and Sustainable Energy Reviews 82 (2018) 2664–2679. doi:10.1016/j.rser.2017.09.102.

[55] R. Z. Homod, K. S. M. Sahari, H. A. Almurib, F. H. Nagi, Rlf and ts fuzzy model identification of indoor thermal comfort based on pmv/ppd, Building and Environment 49 (2012) 141–153. doi:10.1016/j.buildenv.2011.09.012.

[56] M. Wetter, W. Zuo, T. S. Nouidui, X. Pang, Modelica buildings library, Journal of Building Performance Simulation 7 (2014) 253–270. doi:10.1080/19401493.2013.765506.

[57] L. Yang, Z. Nagy, P. Goffin, A. Schlueter, Reinforcement learning for optimal control of low exergy buildings, Applied Energy 156 (2015) 577–586. doi:10.1016/j.apenergy.2015.07.050.

[58] A. Tavakoli, F. Pardo, P. Kormushev, Action branching architectures for deep reinforcement learning, Proceedings of the AAAI Conference on Artificial Intelligence 32 (2018). URL: https://ojs.aaai.org/index.php/AAAI/article/view/11798.

[59] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT press, 2018.

[60] H. Van Hasselt, A. Guez, D. Silver, Deep reinforcement learning with double q-learning, in: Proceedings of the AAAI conference on artificial intelligence, volume 30, 2016. doi:10.1609/aaai.v30i1.10295.

[61] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, N. Freitas, Dueling network architectures for deep reinforcement learning, in: Proceedings of The 33rd International Conference on Machine Learning, PMLR, 2016, pp. 1995–2003.

[62] M. Sewak, Deep reinforcement learning, Springer, 2019.

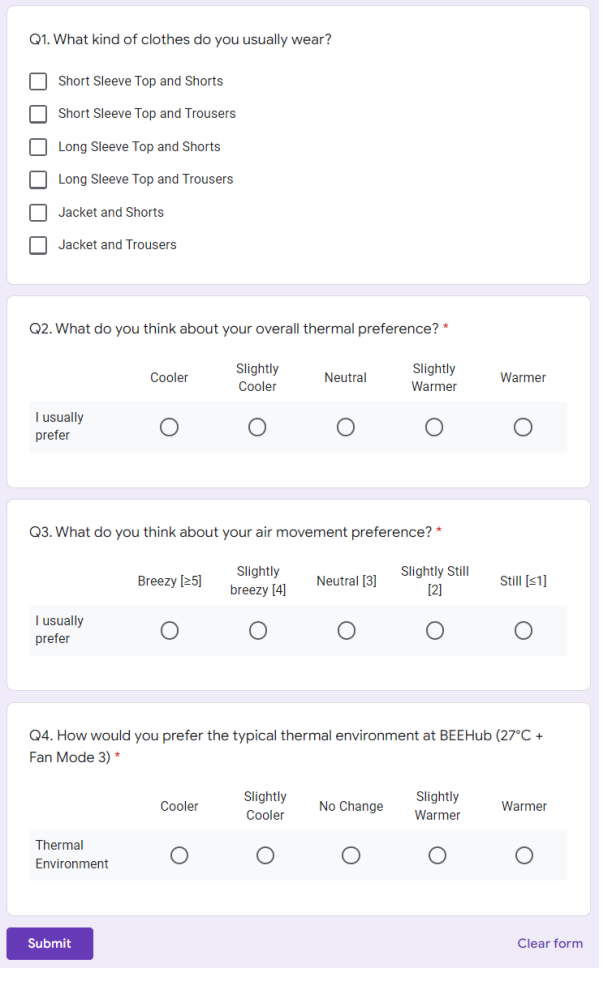
[63] L. Gunnarsen, P. O. Fanger, Adaptation to indoor air pollution, Environment International 18 (1992) 43–54. doi:10.1016/0160-4120(92)90209-M.

[64] L.-J. Lin, Self-improving reactive agents based on reinforcement learning, planning and teaching, Machine learning 8 (1992) 293–321. doi:10.1007/BF00992699.

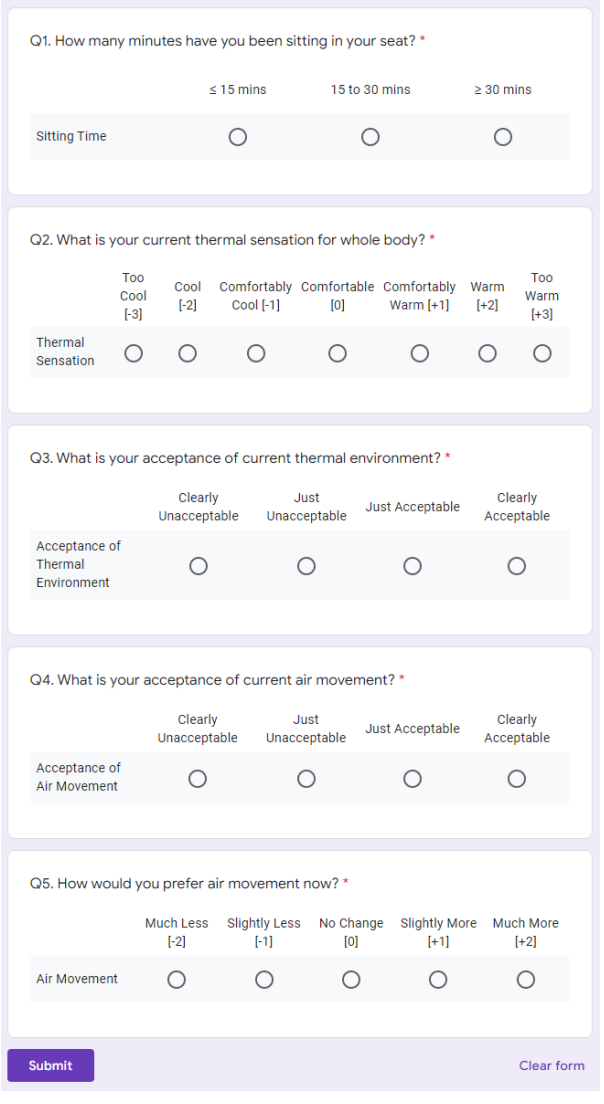
[65] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, W. Zaremba, Openai gym, arXiv preprint arXiv:1606.01540 (2016). doi:10.48550/arXiv.1606.01540.
[66] C. Andersson, J. Åkesson, C. Führer, Pyfmi: A python package for simulation of coupled dynamic models with the functional mock-up interface, Centre for Mathematical Sciences, Lund University Lund, Sweden, 2016.
[67] S. Zhang, R. S. Sutton, A deeper look at experience replay, arXiv preprint arXiv:1712.01275 (2017). doi:10.48550/arXiv.1712.01275.
[68] A. Chong, W. Xu, S. Chao, N.-T. Ngo, Continuous-time bayesian calibration of energy models using bim and energy data, Energy and Buildings 194 (2019) 177–190. doi:10.1016/j.enbuild.2019.04.017.
[69] S. Zhan, G. Wichern, C. Laughman, A. Chakrabarty, Meta-learned bayesian optimization for calibrating building simulation models with multi-source data, in: NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning, 2021. URL: https://www.climatechange.ai/papers/neurips2021/18.
[70] ANSI/ASHRAE, ANSI/ASHRAE Standard 55-2020: thermal environmental conditions for human occupancy, 2020.
[71] K. Mihara, C. Sekhar, Y. Takemasa, B. Lasternas, K. W. Tham, Thermal comfort and energy performance of a dedicated outdoor air system with ceiling fans in hot and humid climate, Energy and Buildings 203 (2019) 109448. doi:10.1016/j.enbuild.2019.109448.
[72] T. P. Velavan, C. G. Meyer, The covid-19 epidemic, Tropical medicine & international health 25 (2020) 278. doi:10.1111/tmi.13383.
[73] W. Jung, F. Jazizadeh, Comparative assessment of hvac control strategies using personal thermal comfort and sensitivity models, Building and Environment 158 (2019) 104–119. doi:10.1016/j.buildenv.2019.04.043.
[74] A. G. ASHRAE, Guideline 14-2014: Measurement of energy, demand, and water savings, American Society of Heating, Refrigerating, and Air Conditioning Engineers, Atlanta, Georgia (2014).
[75] Y. Zhang, R. Zhao, Overall thermal sensation, acceptability and comfort, Building and environment 43 (2008) 44–50. doi:10.1016/j.buildenv.2006.11.036.
[76] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, D. Meger, Deep reinforcement learning that matters, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018. URL: https://ojs.aaai.org/index.php/AAAI/article/view/11694.
[77] T. L. Paine, C. Paduraru, A. Michi, C. Gulcehre, K. Zolna, A. Novikov, Z. Wang, N. de Freitas, Hyperparameter selection for offline reinforcement learning, arXiv preprint arXiv:2007.09055 (2020). doi:10.48550/arXiv.2007.09055.
[78] M. Kearns, Y. Mansour, A. Y. Ng, A sparse sampling algorithm for near-optimal planning in large markov decision processes, Machine learning 49 (2002) 193–208.
[79] R. Solis, A. Pakbin, A. Akbari, B. J. Mortazavi, R. Jafari, A human-centered wearable sensing platform with intelligent automated data annotation capabilities, in: Proceedings of the International Conference on Internet of Things Design and Implementation, 2019, pp. 255–260.

Appendix

A. Comfort Surveys and Personalized Comfort Matrix

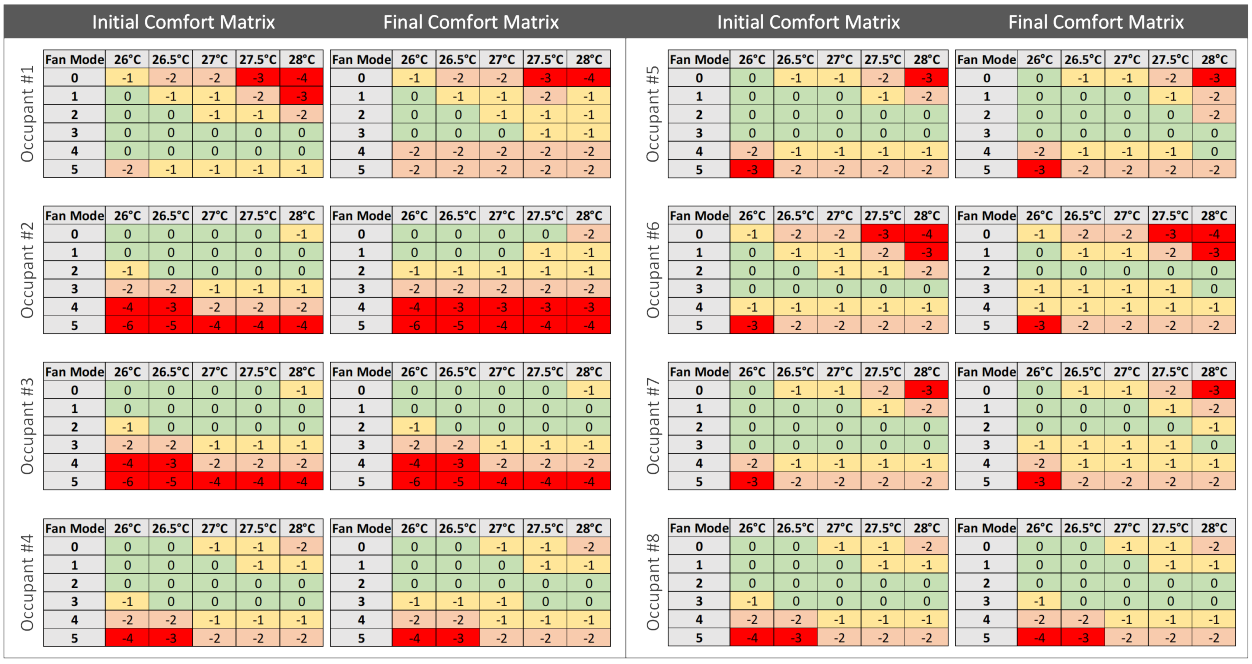


(a) Preliminary comfort survey



(b) Online comfort survey

Figure 12: Comfort surveys used in this study to initialize and online update the personalized comfort matrix.



Occupant #1

Initial Comfort Matrix					
Fan Mode	26°C	26.5°C	27°C	27.5°C	28°C
0	-1	-2	-2	-3	-4
1	0	-1	-1	-2	-3
2	0	0	-1	-1	-2
3	0	0	0	0	0
4	0	0	0	0	0
5	-2	-1	-1	-1	-1

Final Comfort Matrix					
Fan Mode	26°C	26.5°C	27°C	27.5°C	28°C
0	-1	-2	-2	-3	-4
1	0	-1	-1	-2	-1
2	0	0	-1	-1	-1
3	0	0	0	-1	-1
4	-2	-2	-2	-2	-2
5	-2	-2	-2	-2	-2

Occupant #2

Initial Comfort Matrix					
Fan Mode	26°C	26.5°C	27°C	27.5°C	28°C
0	0	0	0	0	-1
1	0	0	0	0	0
2	-1	0	0	0	0
3	-2	-2	-1	-1	-1
4	-4	-3	-2	-2	-2
5	-6	-5	-4	-4	-4

Final Comfort Matrix					
Fan Mode	26°C	26.5°C	27°C	27.5°C	28°C
0	0	0	0	0	-2
1	0	0	0	-1	-1
2	-1	-1	-1	-1	-1
3	-2	-2	-2	-2	-2
4	-4	-3	-3	-3	-3
5	-6	-5	-4	-4	-4

Occupant #3

Initial Comfort Matrix					
Fan Mode	26°C	26.5°C	27°C	27.5°C	28°C
0	0	0	0	0	-1
1	0	0	0	0	0
2	-1	0	0	0	0
3	-2	-2	-1	-1	-1
4	-4	-3	-2	-2	-2
5	-6	-5	-4	-4	-4

Final Comfort Matrix					
Fan Mode	26°C	26.5°C	27°C	27.5°C	28°C
0	0	0	0	0	-1
1	0	0	0	0	0
2	-1	0	0	0	0
3	-2	-2	-1	-1	-1
4	-4	-3	-2	-2	-2
5	-6	-5	-4	-4	-4

Occupant #4

Initial Comfort Matrix					
Fan Mode	26°C	26.5°C	27°C	27.5°C	28°C
0	0	0	-1	-1	-2
1	0	0	0	-1	-1
2	0	0	0	0	0
3	-1	0	0	0	0
4	-2	-2	-1	-1	-1
5	-4	-3	-2	-2	-2

Final Comfort Matrix					
Fan Mode	26°C	26.5°C	27°C	27.5°C	28°C
0	0	0	-1	-1	-2
1	0	0	0	-1	-1
2	0	0	0	0	0
3	-1	-1	-1	0	0
4	-2	-2	-1	-1	-1
5	-4	-3	-2	-2	-2

Occupant #5

Initial Comfort Matrix					
Fan Mode	26°C	26.5°C	27°C	27.5°C	28°C
0	0	-1	-1	-2	-3
1	0	0	0	-1	-2
2	0	0	0	0	0
3	0	0	0	0	0
4	-2	-1	-1	-1	-1
5	-3	-2	-2	-2	-2

Final Comfort Matrix					
Fan Mode	26°C	26.5°C	27°C	27.5°C	28°C
0	0	-1	-1	-2	-3
1	0	0	0	-1	-2
2	0	0	0	0	-2
3	0	0	0	0	0
4	-2	-1	-1	-1	0
5	-3	-2	-2	-2	-2

Occupant #6

Initial Comfort Matrix					
Fan Mode	26°C	26.5°C	27°C	27.5°C	28°C
0	-1	-2	-2	-3	-4
1	0	-1	-1	-2	-3
2	0	0	-1	-1	-2
3	0	0	0	0	0
4	-1	-1	-1	-1	-1
5	-3	-2	-2	-2	-2

Final Comfort Matrix					
Fan Mode	26°C	26.5°C	27°C	27.5°C	28°C
0	-1	-2	-2	-3	-4
1	0	-1	-1	-2	-3
2	0	0	0	0	0
3	-1	-1	-1	-1	0
4	-1	-1	-1	-1	-1
5	-3	-2	-2	-2	-2

Occupant #7

Initial Comfort Matrix					
Fan Mode	26°C	26.5°C	27°C	27.5°C	28°C
0	0	-1	-1	-2	-3
1	0	0	0	-1	-2
2	0	0	0	0	0
3	0	0	0	0	0
4	-2	-1	-1	-1	-1
5	-3	-2	-2	-2	-2

Final Comfort Matrix					
Fan Mode	26°C	26.5°C	27°C	27.5°C	28°C
0	0	-1	-1	-2	-3
1	0	0	0	-1	-2
2	0	0	0	0	-1
3	-1	-1	-1	-1	0
4	-2	-1	-1	-1	-1
5	-3	-2	-2	-2	-2

Occupant #8

Initial Comfort Matrix					
Fan Mode	26°C	26.5°C	27°C	27.5°C	28°C
0	0	0	-1	-1	-2
1	0	0	0	-1	-1
2	0	0	0	0	0
3	-1	0	0	0	0
4	-2	-2	-1	-1	-1
5	-4	-3	-2	-2	-2

Final Comfort Matrix					
Fan Mode	26°C	26.5°C	27°C	27.5°C	28°C
0	0	0	-1	-1	-2
1	0	0	0	-1	-1
2	0	0	0	0	0
3	-1	0	0	0	0
4	-2	-2	-1	-1	-1
5	-4	-3	-2	-2	-2

Figure 13: The initial and final personalized comfort matrix for each participant. The discomfort penalties are highlighted in different colors for better visualization: clearly acceptable ↦ green, just acceptable ↦ yellow, just unacceptable ↦ orange, clearly unacceptable ↦ red.