Deng, Z. and Chen, Q. 2021. “Reinforcement learning of occupant behavior model for cross-building transfer learning to various HVAC control systems,” *Energy and Buildings*, 238: 110860.

Reinforcement learning of occupant behavior model for cross-building transfer learning to various HVAC control systems

Zhipeng Deng[1], Qingyan Chen[1], *
[1]Center for High Performance Buildings (CHPB), School of Mechanical Engineering, Purdue University, 585 Purdue Mall, West Lafayette, IN 47907, USA

*Corresponding author: Qingyan Chen, yanchen@purdue.edu

Abstract

Occupant behavior plays an important role in the evaluation of building performance. However, many contextual factors, such as occupancy, mechanical system and interior design, have a significant impact on occupant behavior. Most previous studies have built data-driven behavior models, which have limited scalability and generalization capability. Our investigation built a policy-based reinforcement learning (RL) model for the behavior of adjusting the thermostat and clothing level. Occupant behavior was modelled as a Markov decision process (MDP). The action and state space in the MDP contained occupant behavior and various impact parameters. The goal of the occupant behavior was a more comfortable environment, and we modelled the reward for the adjustment action as the absolute difference in the thermal sensation vote (TSV) before and after the action. We used Q-learning to train the RL model in MATLAB and validated the model with collected data. After training, the model predicted the behavior of adjusting the thermostat set point with R^2 from 0.75 to 0.8, and the mean absolute error (MAE) was less than 1.1 °C (2 °F) in an office building. This study also transferred the behavior knowledge of the RL model to other office buildings with different HVAC control systems. The transfer learning model predicted the occupant behavior with R^2 from 0.73 to 0.8, and the MAE was less than 1.1 °C (2 °F) most of the time. Going from office buildings to residential buildings, the transfer learning model also had an R^2 over 0.6. Therefore, the RL model combined with transfer learning was able to predict the building occupant behavior accurately with good scalability, and without the need for data collection.

Keywords
Thermal comfort, machine learning, artificial neural network, air temperature, thermostat set point, Q-learning, building performance simulation

1. Introduction

In the United States, buildings account for 41% of primary energy use, mainly for maintaining a comfortable and healthy indoor environment [1]. Unfortunately, current methods for simulating building energy consumption are often inaccurate, and the error can be as high as 150% to 250% [2, 3]. Discrepancies between the simulated and actual energy consumption may arise from various occupant behavior in buildings [4, 5]. Therefore, it is important to estimate the impact of occupant behavior on building energy consumption [6].

Occupant behavior in buildings refers to occupants' movements and their interactions with building components such as thermostats, windows, lights, blinds and internal equipment [7]. The existing methods for exploring the effects of occupant behavior on energy consumption were mostly based on building performance simulations [8]. In these simulations, modelling occupant behavior is challenging due to its complexity [9, 10, 11]. Previous studies have tried to predict the energy consumption in commercial and residential buildings with the use of various occupant behavior models. These models can be divided into three categories: data-driven, physics-based and hybrid models.

In the data-driven category, many researchers have built linear regression models [12], logistic regression models [13, 14], statistical models [15-16], and artificial neural network (ANN) models [17]. To be specific, Andersen [12] and Fabi [13] collected data on occupants' heating set-points in dwellings and predicted the thermal preference along with indoor environmental quality and heating demand. Langevin's model [14] used heating set-point data from a one-year field study in an air-conditioned office building. Sun and Hong [16] used a simulation approach to estimate energy savings for five common types of occupant behavior in a real office building across four typical climates. Deng and Chen [17] collected data in an office building for one year to predict occupant behavior in regard to thermostat and clothing level by means of an ANN model. In these studies, the models considered different variables that affect occupant behavior in buildings. However, the generalization capabilities of these data-driven models were not good [18], since the occupant behavior differed from building to building. Some review papers [19, 20] have discussed contextual factors that cause occupant behavior to vary greatly, such as room occupancy, availability and accessibility of an HVAC system, and interior design. The authors observed that it was difficult to apply an occupant behavior model developed for one building to another building. Hong et al. also indicated that, because a large number of data-driven behavior models emerged in scattered locations around the world, they lack standardization and consistency and cannot easily be compared one with another [21]. Moreover, all the data-driven models require sufficient data for training, but the estimation of building energy and modelling of occupant behavior are done mostly during the early design stages, when collecting occupant behavior data is impossible [22]. It is hard to build a data-driven occupant behavior model without data or satisfactory generalization capability.

As for the physics-based models, a review by Jia et al. [23] pointed out that occupant behavior modelling has progressed from deterministic or static to more detailed and complex. Therefore, many researchers have based their models on the causal relationships of occupant behavior. The driving factors of occupant behavior can be divided into three main types: environmentally related, time related and random factors [20, 24]. Hong et al. developed a DNAS (drivers, needs, actions, systems) framework that standardized the representation of energy-related occupant behavior in buildings [21]. Many researchers have adopted this framework for their behavior studies. For example, dynamic Bayesian networks by Tijani et al. [25] simulated the occupant behavior in office buildings as it relates to indoor air quality. The advantage of Bayesian network model was in its representation of occupant behavior as probabilistic cause-effect relationships based on prior knowledge. D'Oca et al. [26] built a knowledge discovery database for window-operating behavior in 16 offices. Zhou et al. [27] used an action-based Markov chain approach to predict window-operating actions in office spaces. They found that the Markov chain reflected the actual behavior accurately in an open-plan office and was therefore a beneficial supplemental module for energy simulation software. The Markov chain model depends on the previous state to predict the probability of an event occurring. This characteristic is useful for representing individuals' actions and motivations [9]. In addition, many researchers have built other kinds of models for different building types and scenarios. For instance, hidden Markov models [23, 28] were used to simulate occupant behavior with unobservable hidden states, and thus these models could be employed under very complicated conditions. Survival models [29] could feature different occupant types to mimic variations in control behavior. Meanwhile, a decision tree model [30, 31] regarded occupant decisions and possible behavior as branched graphical classification. This model was straightforward, but complex causal factors in real situations might give rise to too many branches. In recent years, more complex agent-based models [32-34] have yielded good predictions of occupant behavior with individual differences among occupants. In short, physics-based occupant behavior models with physical meaning have exhibited better generalization capability than data-driven models. Hence, the present study used a Markov decision process (MDP) to model occupant behavior and build a logic-based reinforcement learning model to explore the model's scalability.

Reinforcement learning (RL) is a machine learning area concerned with the ways in which agents take actions to maximize certain rewards [35]. Off-policy RL can use historical data for training without interacting with the environment. In contrast, policy-based reinforcement learning does not require previous training data because it creates its own experience via random explorations of the environment. As such, this way of learning can obtain rules and knowledge not limited to specific conditions but adaptable to various scenarios. It has been applied successfully to a range of fields, including robot control [36] and playing Go [37]. In the built environment, the RL model has been used to improve building energy efficiency and management when the reward is defined as minimizing building energy consumption [38-40]. For instance, Zhang et al. [38] used deep reinforcement learning to control a radiant heating system in an existing office building and achieved a 16.7% reduction in heating demand. A multi-agent reinforcement learning framework by Kazmi et al. [39] achieved a 20% reduction in the energy required for the hot water systems in over 50 houses. Liang [40] modelled an HVAC scheduling system control as an MDP, and the model did not require prior knowledge of the building thermal dynamics model. Similarly, when the reward is the thermal comfort level of occupants, the RL model can be used to control the thermal comfort and HVAC system in buildings [41, 42]. For example, Yoon et al. [43] built performance-based comfort control for cooling while minimizing the energy consumption. Ruelens and coauthors [44] used model-free

RL for a heat-pump thermostat. Their learning agent reduced the energy consumption by 4–9% during 100 winter days and by 9–11% during 80 summer days. Azuatalam et al. [45] applied RL to the optimal control of whole-building HVAC systems while harnessing RL's demand response capabilities. Similarly, Chen [46] and Ding [47] developed novel deep RL for reducing the training data set and training time. Meanwhile, several previous studies used the RL model for advanced building control [43, 48, 49] and lighting control [50]. In addition, there have been some integrated applications. For example, Valladares et al. [51] used the RL model with a probability of reward combination to improve both the thermal comfort and indoor air quality in buildings. The RL model developed by Brandi et al. [52] optimized indoor temperature control and heating energy consumption in buildings. Ding et al. [53] also employed a novel deep RL framework for optimal control of building subsystems, including HVAC, lighting, blind and window. Hence, RL can be used to model the HVAC system for both thermal comfort and energy management. Physics-based and model-free RL also have the potential to model occupant behavior without data since the logic is very similar. Therefore, this research built an RL model for thermostat set point and clothing level adjustment behavior based on the correlation between thermal sensation and thermally influenced occupant behavior [17].

For modeling of the occupant behavior in buildings with limited information and no data, transfer learning was a feasible approach [18]. The transfer learning method stores knowledge about one problem and then applies it to a related problem. It has been used for cross-building [54, 55], cross-home [56] and even cross-city [57] energy modelling. For instance, Mocanu et al. [58] transferred a building energy prediction to a new building in a smart grid. Ribeiro et al. [59] used various machine learning methods to predict school building energy and transfer the prediction to other new schools. Gao et al. [60] built a transfer learning model for thermal comfort prediction in multiple cities. Xu et al. [61] conducted transfer learning for HVAC control between buildings with different sizes, numbers of thermal zones, materials, layouts, air conditioner types, and ambient weather conditions. They found that this approach significantly reduced the training time and energy cost. Therefore, based on the potential of transfer learning, we used it to transfer knowledge about occupant behavior from one building to other buildings.

The purpose of the present study was to build an RL occupant behavior model for thermostat and clothing level adjustment in a particular building, and transfer the model to other buildings with different HVAC control systems. For this purpose, we first built an MDP of the occupant behavior and used a thermal sensation model to build the rewards. We then trained the RL model with the use of Q-learning. Next, we used transfer learning to explore the occupant behavior in several other buildings. We also validated the RL occupant behavior model and the transferred model with data collected from various buildings. Finally, we analyzed the simulated building energy performance with the use of the RL model and the transferred model.

2. Methods

To develop an occupant behavior model, we first modeled the occupant behavior as an MDP and developed the RL model on the basis of this process. Subsequently, we trained the model with the use of a Q-learning algorithm. Next, we transferred the knowledge of the occupant behavior model from one building with manual control to other buildings with thermostat setback and occupancy control systems. Finally, we validated the transfer learning model with collected data. Fig. 1 summarizes the methods and models in this study.

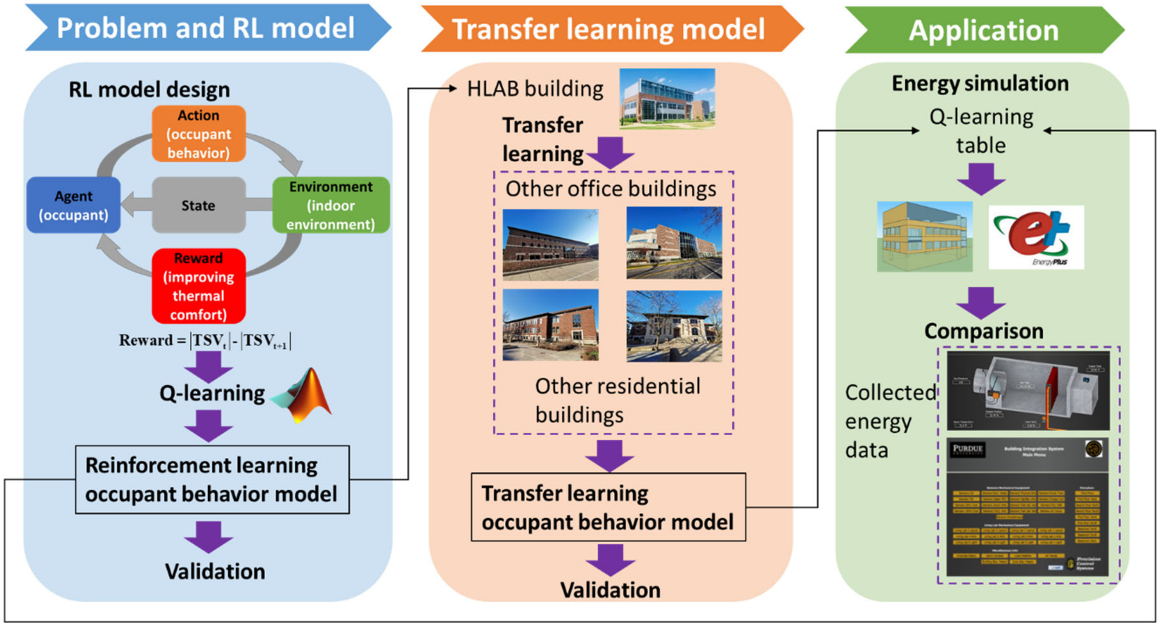


Fig. 1. Flow chart of methods in this study, including the reinforcement learning occupant behavior model, transfer learning model and energy simulation

2.1 Framework of reinforcement learning model

As shown in Fig. 2, in the RL model, an agent can gather information directly from the environment of different states, and then take actions inside and compare the results of these actions via the reward function. This cycle is repeated over time, until the agent has enough experience to correctly choose the actions that yield the maximum reward. Thus, through interaction with an environment and repeated actions, the RL model can evaluate the consequences of actions by learning from past experience. As for the building occupants, the decision to take an action in a specific indoor environment is a similar process to that of the RL model. The MDP is used to describe an environment for reinforcement learning, because the indoor environment and thermal comfort are fully observable. In this study, the occupant behavior was modelled as a decision-making process in which the policy-based RL was used. The building occupant, the occupant behavior, the indoor environment and the improving thermal comfort level are the agent, action, state and reward, respectively, in the model. In each state, the logic of occupant behavior is to proactively seek more comfortable conditions in the indoor environment [11]. Numerous factors are related to the occupant behavior, and we will introduce them in detail in the following sections.

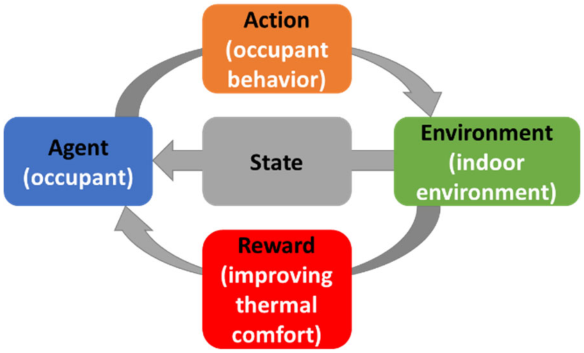


Fig. 2. Illustration of the RL model with agent, action space, environment space and rewards.

We modeled the occupant behavior in offices as an MDP, as shown in Fig. 3. In the initial state, the agent had many possible choices of behavior, such as adjusting the thermostat set point by various degrees or adjusting the clothing level. For every action, there was a corresponding feedback reward, such as improvement or deterioration of thermal comfort. The agent took an action to enter a follow-up environment, and this process kept going. The time step size for action prediction was 15 minutes. We took the actual occupant behavior occurrence into consideration, because there was a certain delay in the occurrence of the behavior, and the occupant did not act immediately when feeling uncomfortable. We also assumed that the action could take effect in the subsequent time step if the HVAC system was in normal operation. Note that in Fig. 3 we have listed only some possible actions. There may be others, such as reducing the clothing level and making a more extreme adjustment to the thermostat set point. These additional actions are represented by an ellipsis.

The MDP in this study entailed the following specifications:

Environment space: The state contains information about the indoor environment that occupants use in deciding on the proper action. In this research, the state space included room air temperature, room air relative humidity, thermostat set point, clothing level of occupants, metabolic rate, room occupancy and time of day. Although there are many other factors [20, 24] that impact occupant behavior, we neglected them in order to simplify the structure of the RL model. Here we assumed that the thermal sensation of occupants was not impacted by the time of day. Therefore, time was not included in the TSV and reward calculation. An exception was the transfer learning model for setback and occupancy control in Section 2.3, which moved to a nighttime state at certain times. Generally, time functioned as a label, and it did not contain a numerical value that might influence the RL model and training. In summary, the state space can be expressed as

$$S = \{T_{air}, RH_{air}, T_{setpoint}, Clo, Met, occupancy, time\} \quad (1)$$

Action space: The action is the occupant behavior that is performed with the goal of more comfortable conditions. In this research, the action space included raising or lowering the thermostat set point by different degrees, or maintaining the same set point; putting on, keeping the same, or taking off clothes; and arriving. The action space can be expressed as

$$A = \left\{ A_{raise}, A_{keep}, A_{lower}, A_{put\ on}, A_{keep}, A_{take\ off} \right\} \tag{2}$$

where the first three actions $A_{raise}, A_{keep}, A_{lower}$ represent adjustments to the thermostat set point, and the last three actions $A_{put\ on}, A_{keep}, A_{take\ off}$ represent adjustments to the clothing level.

Reward function: The goal of the action is a higher thermal comfort level for the occupants. Therefore, in this research, the reward was modelled as the absolute difference between the initial TSV before the action and the final TSV after the action, which can be expressed as

$$R = \left| TSV_t \right| - \left| TSV_{t+1} \right| \tag{3}$$

where subscripts *t* and *t+1* represent the current and next time steps, respectively. It is clear that in order to maximize the reward *R*, $TSV_{t+1} = 0$, which means that the desired thermal sensation is neutral after the occupant behavior occurs.

In this research, we predicted the TSV in offices with the use of an ANN model [17, 62] that expresses TSV as a function of four input parameters as:

$$\text{TSV} = f\,(\text{air temperature, relative humidity, clothing insulation, metabolic rate}) \tag{4}$$

where *f* represents the function of the ANN model. We assumed that the mean radiation temperature was the same as the air temperature, and the air velocity was less than 0.2 m/s. To develop the ANN model, we collected data from over 25 occupants in an office building during the four seasons of 2017. The number of collected data points for training the model was about 5,000. The model had three layers, and there were ten neurons in the hidden layer. We used the Levenberg-Marquardt algorithm to train the model, and it predicted the TSV with a mean absolute error (MAE) of 0.43 after training.

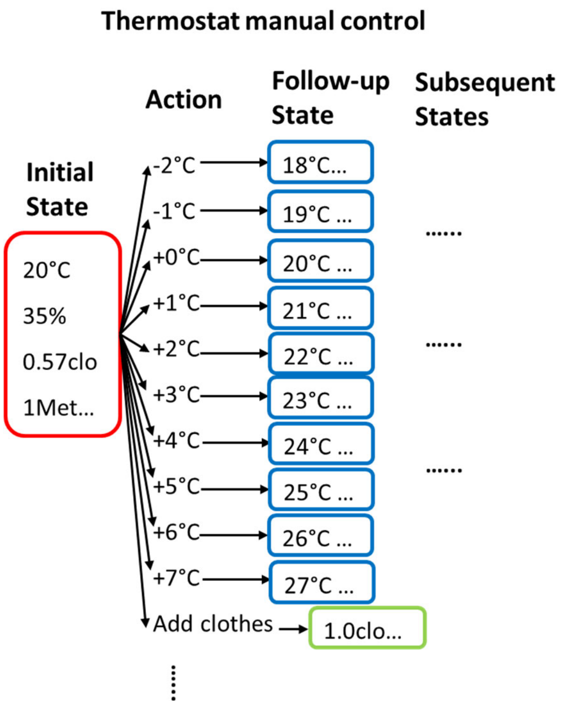


Fig. 3. MDP for the occupant behavior of thermostat set point manual control and clothing level adjustment. Each state space includes numerous parameters, as expressed by Eq. (1), and the figure displays only the key parameters. The initial state is followed by many actions, follow-up states and possible subsequent states. In addition to what is shown in the figure, further possibilities are indicated by an ellipsis.

For buildings without a thermal comfort model, predicted mean vote (PMV) [63] can also be used to model the reward, which is expressed as

$$R = |PMV_t| - |PMV_{t+1}| \tag{5}$$

As above, maximizing the reward R requires that $PMV_{t+1} = 0$.

Reward modelling in the RL model for multi-occupant offices with multiple agents [64] was different from that for single-occupant offices. For multi-occupant offices, the modelling was divided into two categories. In one category, the reward of a dominant occupant was maximized. Here, one occupant near the thermostat would adjust the thermostat dominantly, and the others in the room would compromise with this occupant's preference, as is the case in some workplaces [17, 65]. Thus, the reward was for the dominant individual and can be expressed as

$$R = |TSV_{t,dominant}| \text{-} |TSV_{t+1,dominant}| \tag{6}$$

During data collection, we also found that in some offices all the occupants had equal control of the thermostat [17]. Therefore, in our other multi-occupant office category, the average reward for all occupants was maximized. The reward was averaged as

$$R = \frac{1}{n}\sum_{i}\left(\left|TSV_{t,i}\right| - \left|TSV_{t+1,i}\right|\right) \tag{7}$$

where *n* is the number of occupants in the room, and *i* represents different occupants. For a single-occupant office where only the dominant occupant was in the room, the two categories of reward modelling were the same as Eq. (6) and (7).

2.2 Q-learning

After designing the model framework, we needed to train the RL model. One of the available training methods is Q-learning. Here "Q" means "quality," a policy function of an action taken in a given state. It can be expressed as the following mapping:

$$Q : S \times A \rightarrow R \tag{8}$$

Q-learning is a model-free RL algorithm for learning a policy that tells an agent which actions to take under various circumstances [66]. This learning method has been widely used for training RL models [43, 49, 51, 67, 68]. With the state space, action space and reward modelling described in Section 2.1, we used the Q-learning algorithm to update the quality. The updating equation for Q-learning can be expressed as

$$Q_{new}\left(s_t, a_t\right) = Q_{old}\left(s_t, a_t\right) + \alpha \cdot \left[r_t + \gamma \cdot \max_{a} Q\left(s_{t+1}, a\right) - Q_{old}\left(s_t, a_t\right)\right] \tag{9}$$

where Q is the quality, s the state, a the action, α the learning rate, r the reward, γ the discount factor, and $\max_{a} Q\left(s_{t+1}, a\right)$ the estimation of optimal future value. According to this equation, as the training begins, the quality is initialized to arbitrary or uniform values. Then, at each episode t of the training process, the agent in state s_t selects an action a_t with a reward r_t and an estimated future reward for future actions. After the action, the agent enters a new state s_{t+1}. When the maximized reward is confirmed, the optimal action is learned and the quality Q is updated. In this process, the RL model gradually learns to take actions in a certain environment, and we can obtain a Q-learning table of states by various actions. Q-learning is similar to the actual decision process for occupant behavior in buildings.

The learning rate and discount factor could impact the learning process. In this study, we selected a learning rate of 0.3 and discount factor of 1. We used a table of states by various actions because the choices of actions in the MDP were discrete for adjusting the thermostat by different degrees or clothing insulation to certain values. Thus, the discount factor had little impact on the Q-learning result. As for the learning rate, we will provide training results for learning rate variations in Section 3.2. We used the MATLAB 2020a Reinforcement Learning Toolbox [69] to build and train the RL model.

2.3 Transfer learning

After designing and training the RL occupant behavior model, we sought to transfer the model to other buildings with limited information and even with no data. As shown in Fig. 4(a), an ANN model, one of the data-driven models, has a layered structure with input, hidden and output layers. The training process for the ANN model uses data to update the values of coefficients in the hidden layer. Therefore, the model can only be used for similar buildings with available data. In previous attempts to apply the model directly to other buildings, the performance was usually not good [18, 21]. In those studies, transfer learning of the ANN model grabbed layers of neural network weights and trained the model again with new data. Prediction for different buildings with transferring data-driven models requires the data to retrain. Additionally, the meanings of the coefficients inside the models are still unclear to researchers. Therefore, the information in the hidden layer cannot be transferred or used for other buildings. However, as shown in Fig. 4(b), the policy-based RL occupant behavior model is a logical model with physical meaning, and thus it can be partially transferred to other buildings. We transferred the higher-level rules of the RL model, i.e., the logic of thermal actions, the pursuit of thermal comfort from one building to another building. We could do this because even for different buildings and HVAC control systems, the logic of occupant behavior that seeks more comfortable conditions remained the same. Therefore, the feasible actions and rewards of the RL model were similar for different buildings. For example, we built an RL occupant behavior model for a building with manual thermostat control. In other buildings with thermostat setback or occupancy control, occupants might adjust the thermostat set point in different ways. When they left the room or during the night, the building automation system could reset the thermostat set point to save energy. When the occupants reentered the room, they could adjust the set point and override the system operation. The occupants' overriding of the automation systems might indicate their dissatisfaction [70]. As such, there was a "night state" before the occupants' arrival in the morning, when the set point and air temperature were different, as depicted in Fig. 4(b). After the occupants' arrival or in the morning, the state space entered the normal initial state. Thus, the transfer learning model structure was similar to original model with possible actions and rewards in the daytime. We could therefore transfer a portion of the parameters in the action space and the rewards to other buildings. Even without data for these buildings, we could still model and predict the occupant behavior.

For residential buildings, large-scale collection of occupant behavior data has usually been more difficult, because such buildings are generally not equipped with building automation system (BAS) [17]. The use of questionnaire surveys to gather data has been reported as time-consuming and limited in accuracy [23]. Under this circumstance, building a model by transfer learning was a feasible approach. Similarly, we also transferred the RL occupant behavior model for office buildings to residential buildings. The occupant behavior of manual thermostat control was the same in both types of buildings, but the improved thermal comfort level and reward for actions were different [17]. Moreover, there were other factors that distinguished the occupant behavior in office buildings from that in residential buildings [71, 72]. Therefore, we needed to modify the state space and reward in the transfer learning model for residential buildings.

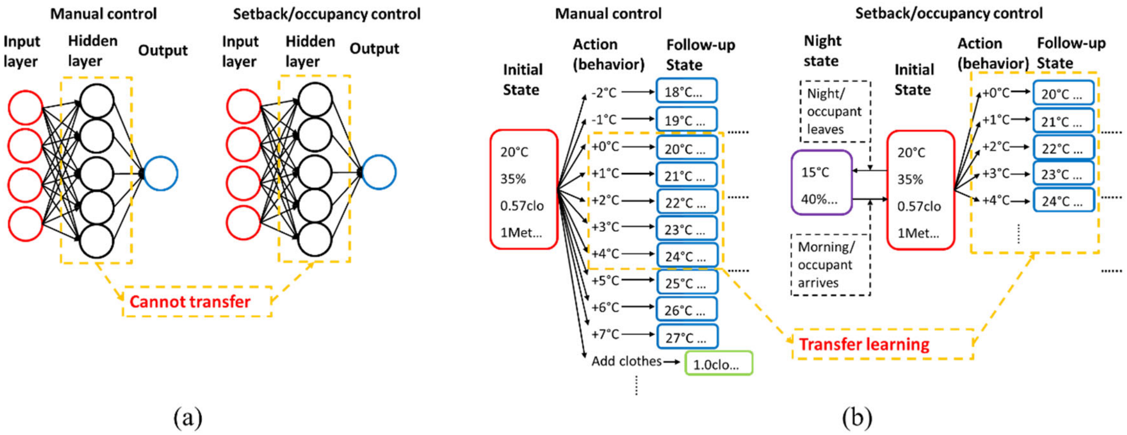


Fig. 4. Transfer of the occupant behavior model for manual control to other buildings with thermostat setback or occupancy control: (a) the data-driven ANN model cannot be transferred because of the coefficient values in the hidden layer; (b) the policy-based RL model can be transferred, and portions of the action and state space are the same.

For residential buildings, a previous study [17] found that the comfort zone of a building was 1.7 °C (3 °F) higher in summer, and 1.7 °C (3 °F) lower in winter, than the ASHRAE comfort zone [73]. Therefore, we were able to use this information to transfer the thermal sensation and occupant behavior model from the office building to residential buildings. Since the shape of the thermal comfort zone was similar, whereas the impact of air temperature on thermal comfort and occupant behavior was different [17], the logical RL behavior model could be partially transferred. The MDP for manual control of the thermostat was the same in the office building and residential buildings. We transferred the RL occupant behavior model with the use of PMV to calculate the reward as

$$R = \left|PMV_{Residence\_i}\right| - \left|PMV_{Residence\_f}\right| \tag{10}$$

Here, the PMV in the residence was defined differently from the traditional PMV model because of the different comfort zone. With the 3 °F difference in winter and summer, it was calculated as

$$PMV_{Residence\_winter} = PMV(T_{air} + 3, RH, T_r, V, Clo, Met) \tag{11}$$

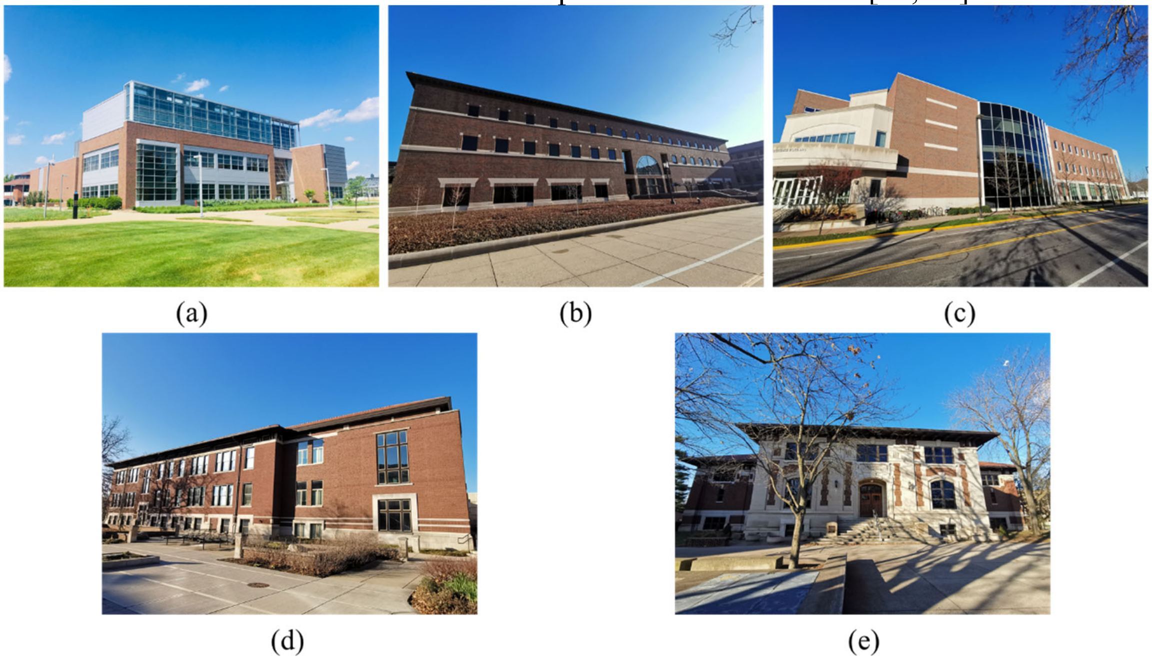
$$PMV_{Residence\_summer} = PMV(T_{air} - 3, RH, T_r, V, Clo, Met) \tag{12}$$

where the *PMV* function represents the traditional way of calculating PMV with six parameters.

2.4 Data collection for model validation

In order to validate the RL model, this study collected indoor air temperature, relative humidity, thermostat set point, lighting occupancy, clothing level of occupants, and data on the occupant behavior of adjusting the thermostat, from the BAS in 20 offices in the Ray W. Herrick Laboratories (HLAB) building at Purdue University in 2018, as shown in Fig. 5 (a). Half of the offices were multi-occupant student offices, and the rest were single-occupant faculty offices. The building used a variable air volume (VAV) system for heating and cooling. Each office had an independent VAV box and a thermostat (Siemens 544-760A) that enabled the BAS to control the air temperature in the room. We downloaded the indoor environment data of room air temperature and thermostat set point from the BAS. In addition, we used a questionnaire to record the clothing level of the occupants and their clothing-adjustment behavior in the HLAB building.

We also gathered room air temperature, relative humidity, thermostat set point and lighting occupancy data in four other office buildings on the Purdue University campus in three seasons of 2018, as shown in Fig. 5(b)-(e). Each building contained more than 100 offices. The HVAC systems in these buildings were similar to those in the HLAB building. However, the HVAC control strategies in the four buildings differed from that in the HLAB building. The HVAC system operated constantly in the HLAB building, and the occupants could adjust the thermostat set point manually. The LWSN building, by contrast, used a thermostat setback that overrode the manual control at night, from 11 PM to 6 AM. Meanwhile, the MSEE, HAAS and STAN buildings used occupancy control for the HVAC system in each room in addition to manual control. Table 1 provides the data collection information for each building, including the number of offices in which data was collected, the HVAC control type, the data collection interval, and the types of data that were collected. The details of the data collection process can be found in [17, 74].



(a) (b) (c)

(d) (e)

Fig. 5. Photographs of the buildings used for data collection: (a) HLAB building, (b) MSEE building, (c) LWSN building, (d) STAN building and (e) HAAS building

Table 1. Data collection information for each building

Building	Offices for data collection	HVAC control type	Data collection interval	Collected data
HLAB	20	Manual control	5 min	Room lighting status Number of room occupants Room air temperature and RH Thermostat set point Room CO_2 concentration Clothing level Room supply-air flow rate Room supply-air temperature
LWSN	106	Manual control +thermostat setback	10 min	Room lighting status Number of room occupants Room air temperature and RH Thermostat set point Clothing level
MSEE	99	Manual control +occupancy control	15 min	
STAN	122	Manual control +occupancy control	15 min	
HAAS	48	Manual control +occupancy control	15 min	

2.5 Building energy simulation with RL model

The purpose of constructing the RL occupant behavior model was to evaluate the impact of occupant behavior on building energy performance. Therefore, we also implemented the RL occupant behavior model in EnergyPlus. We utilized SketchUp to construct the building geometry model in Fig. 6, and then used the model in the EnergyPlus simulations. Table 2 lists the structural and material properties used for the building envelope in the simulations. The structural information was obtained from the HLAB building construction drawings and documents.

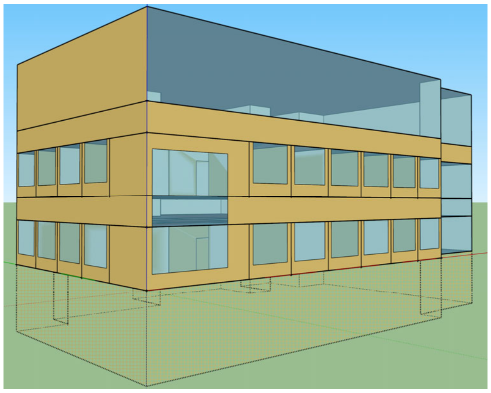


Fig. 6. Geometric model of the HLAB building for EnergyPlus simulations

Table 2. Structural and material properties of the HLAB building for the simulations

Construction component	Layers (from exterior to interior)	Thickness (mm)	Conductivity (W/m K)	Density (kg/m^3)	Specific heat (J/kgK)

Exterior window	Clear float glass	6	0.99	2528	880
	Air cavity	13	0.026	1.225	1010
	Clear float glass	6	0.99	2528	880
Exterior wall 1	Brick	92.1	0.89	1920	790
	Air cavity	60.3	0.026	1.225	1010
	Rigid insulation	50.8	0.03	43	1210
	Exterior sheathing	12.7	0.07	400	1300
	CFMF stud	152.4	0.062	57.26	964
	Gypsum board	15.9	0.16	800	1090
Exterior wall 2	Aluminum panel	50.8	45.28	7824	500
	Rigid insulation	50.8	0.03	43	1210
	Exterior sheathing	12.7	0.07	400	1300
	CFMF stud	152.4	0.062	57.26	964
	Gypsum board	15.9	0.16	800	1090
Interior gypsum wall	Gypsum board	15.9	0.16	800	1090
	Metal stud	92.1	0.06	118	1048
	Gypsum board	15.9	0.16	800	1090
Interior glass wall/door	Glass	6	0.99	2528	880
Interior wood door	Wood	44.45	0.15	608	1630

Fig. 7 depicts the simulation process with the RL occupant behavior model. When the simulation starts, the program first checks whether or not the office is occupied, since the behavior occurs only when there is an occupant inside the office. If so, the agent decides on the action to the next time step based on the Q-learning table. Next, the energy simulation program decides whether or not to adjust the thermostat set point or the clothing level of the occupants. The building energy use will correspond to this decision. Moving to the next time step, the program checks whether or not the simulation time has ended; if not, it again checks if the room is occupied. To obtain a reasonable variation range, we performed the simulation 200 times and analyzed the results [74].

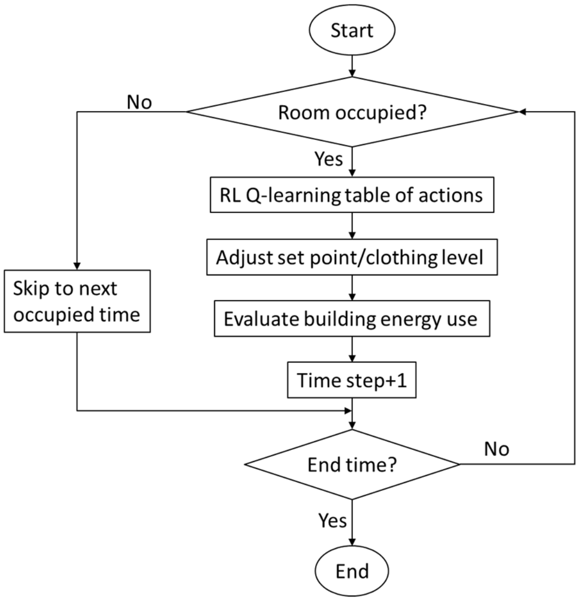
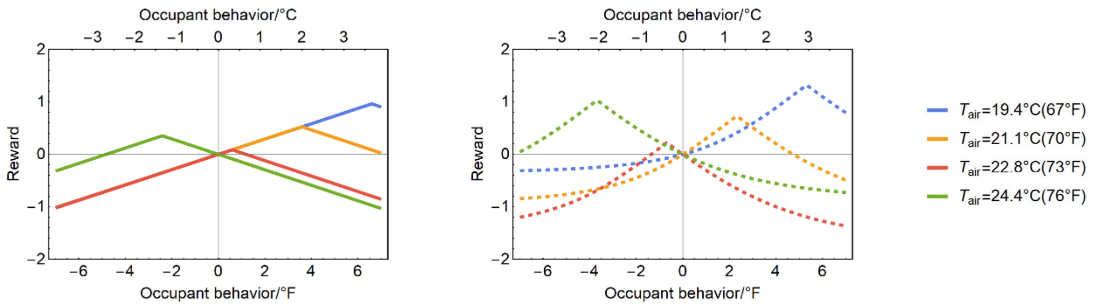

Start
Room occupied?
No
Yes
RL Q-learning table of actions
Adjust set point/clothing level
Evaluate building energy use
Time step+1
Skip to next occupied time
End time?
No
Yes
End

Fig. 7. Building energy simulation process incorporating the RL occupant behavior model and Q-learning table of actions

3 Results

3.1 Results of modelling the reward for action

Fig. 8 shows the result of reward modelling when the PMV model and the thermal comfort ANN model were used with Eqs. (3)–(5). The figure depicts the relationship between occupant behavior and the corresponding rewards in various air temperatures when other parameters were the same. For example, when the air temperature was 19.4 °C (67 °F), the occupant might feel cool in winter. Thus, the reward for raising the thermostat set point was positive most of the time, until the occurrence of overheating caused by an excessive adjustment. For each state, there was one occupant behavior of set point adjustment that led to the maximum reward. The reward situation was similar when the air temperature was high and the occupant lowered the set point. When the air temperature was about 22.8 °C (73 °F), the occupant already felt nearly neutral. In this case, either raising or lowering the set point would lead to a negative reward, and the optimal occupant behavior was to make no adjustment. We used this quantified logic to build the RL model.

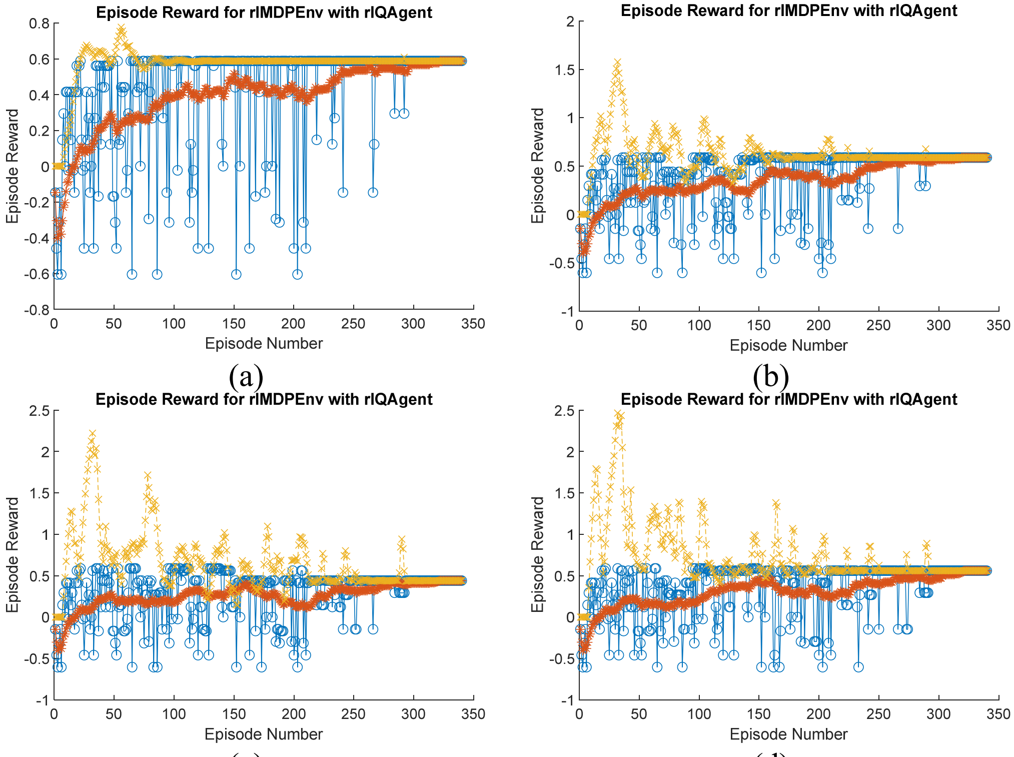


(a) (b)

Fig. 8. Reward value modelled for different air temperatures in winter by using (a) the PMV model and (b) the thermal comfort ANN model.

3.2 Results of the RL occupant behavior model

Fig. 9 depicts the training process for the RL model with the use of Q-learning. The blue, red, and orange curves represent the episode reward, the average reward in nearby episodes, and the quality, respectively. Initially, at the beginning of the training process, the RL model knew nothing about the relationship between the environment, states and actions. Thus, it could only take random actions to explore the relationship, and it received varying rewards. As a result, the episode reward was very low. As the learning process went on, the RL model tried various actions to find a way of maximizing the reward. The quality was updated with the use of Eq. (9). In the examples shown in Fig. 9, the thermostat set point and air temperature were 22.8 °C (73 °F), and the occupant was wearing summer clothing. After training over 300 episodes, the RL model learned to take the action at this state that maximized the reward at 0.61. Fig. 9 also shows that an overly high learning rate made the learning process very unstable, and the quality fluctuated during the training. Meanwhile, a low learning rate would slow down the training process.



(a)

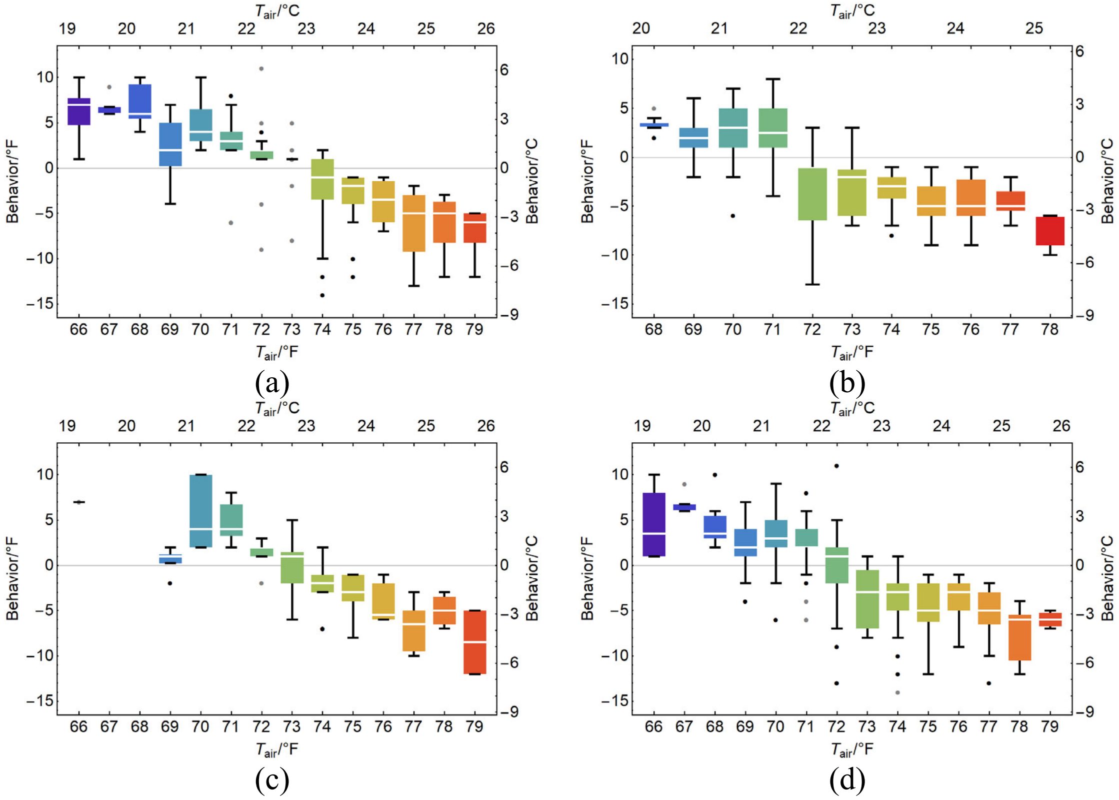
(b)

(c)

(d)

Fig. 9. Training of the RL model with the use of Q-learning as the number of episodes increases. The blue, orange, and yellow curves represent the episode reward, the average reward in nearby episodes, and the quality, respectively. (a) learning rate = 0.1; (b) learning rate = 0.3; (c) learning rate = 0.5; (d) learning rate = 0.7.

The trained RL model would always predict the same occupant behavior in the same state and environment, which was unrealistic. Actual office occupant behavior is influenced by many other factors that we did not build into the RL model [24, 28]. Considering all these factors would have led to an overly complex behavior model. A previous study [11] pointed out that behavior models should not only represent deterministic events but also be described by stochastic laws. Additionally, different thermal preferences on the part of occupants would also cause their behavior to differ. Fig. 10 displays the distribution of collected thermostat set point adjustment behavior at different air temperatures in the HLAB offices. In the box-and-whisker charts, the boxes, whiskers and dots represent the standard deviation, upper and lower bounds, and outliers of the occupant behavior, respectively. The air temperature and occupant behavior had a clear negative correlation. The figure indicates that even at the same air temperature and similar states, the variation range of collected occupant behavior was over ±1.1 °C (2°F) in both single- and multi-occupant offices in different seasons. Under these conditions, the rewards of different actions did not differ greatly, but the RL model always pursued the action that absolutely maximized the reward. For example, the RL model might predict the occupant behavior of raising the set point by 5 °F, while raising it by 4 °F or 6 °F would also be reasonable behavior in a real scenario. Therefore, based on the results in Fig. 10, we added a randomness of -2 °F to +2 °F into the RL model for the final decision to make it more reasonable.



(a)

(b)

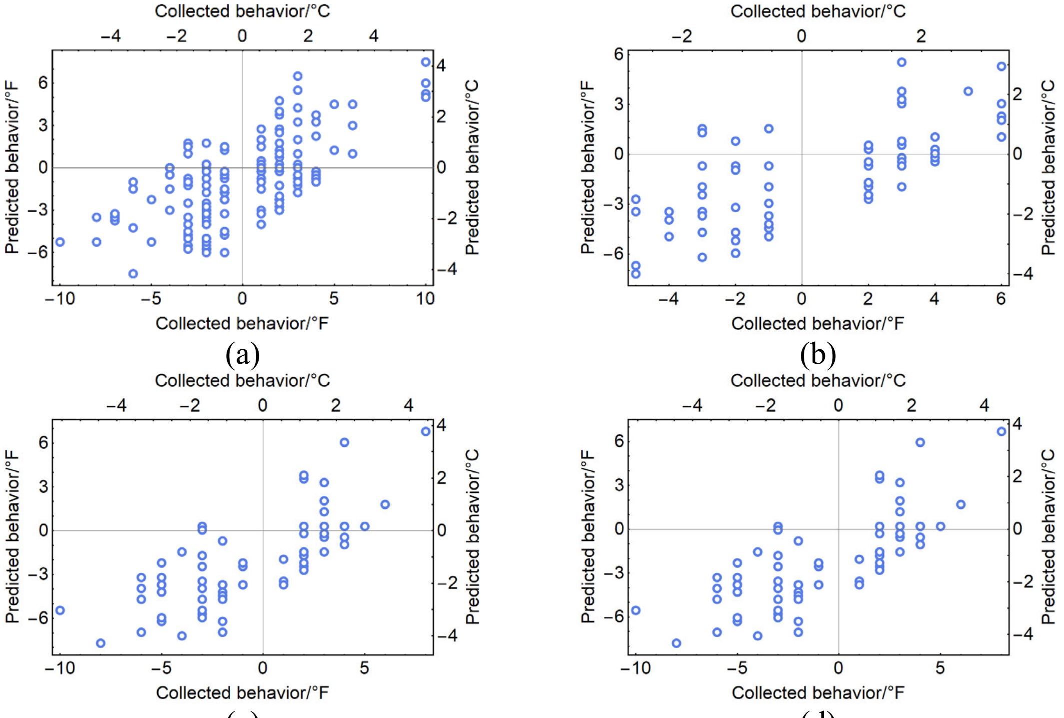
(c)

(d)

Fig. 10. The distribution of thermostat set point adjustment by occupants in: (a) single-occupant offices, (b) multi-occupant offices, (c) winter with Clo = 1, and (d) summer with Clo = 0.57.

3.3 Validation of the RL model

We validated the RL model with the use of data collected in 2018 after adding the randomness for the final decision. Fig. 11 compares the collected occupant behavior with the RL model prediction for HLAB offices in four seasons in 2018. For most of the time, the RL prediction results matched the collected data. Table 3 lists all the prediction results for R^2 and MAE. The R^2 was around 0.7–0.8, and the mean absolute error (MAE) was around 1.5–1.9 °F. The overall R^2 and MAE were 0.79 and 1.68 °F, respectively. We removed some data as outliers when the HVAC system was under maintenance and the occupant lost control. We also compared the performance of the RL model for single- and multi-occupant offices. For single-occupant offices, the R^2 was 0.8 and the MAE was 1.5 °F. For multi-occupant offices, the R^2 was 0.78 and the MAE was 1.8 °F. The prediction results for multi-occupant offices were not as good as for single-occupant offices. In previous studies, a prediction R^2 of 0.8 was deemed acceptable for an occupant behavior model [74]. Hence, the model performance of the RL model was reasonable.



(a) (b)

(c) (d)

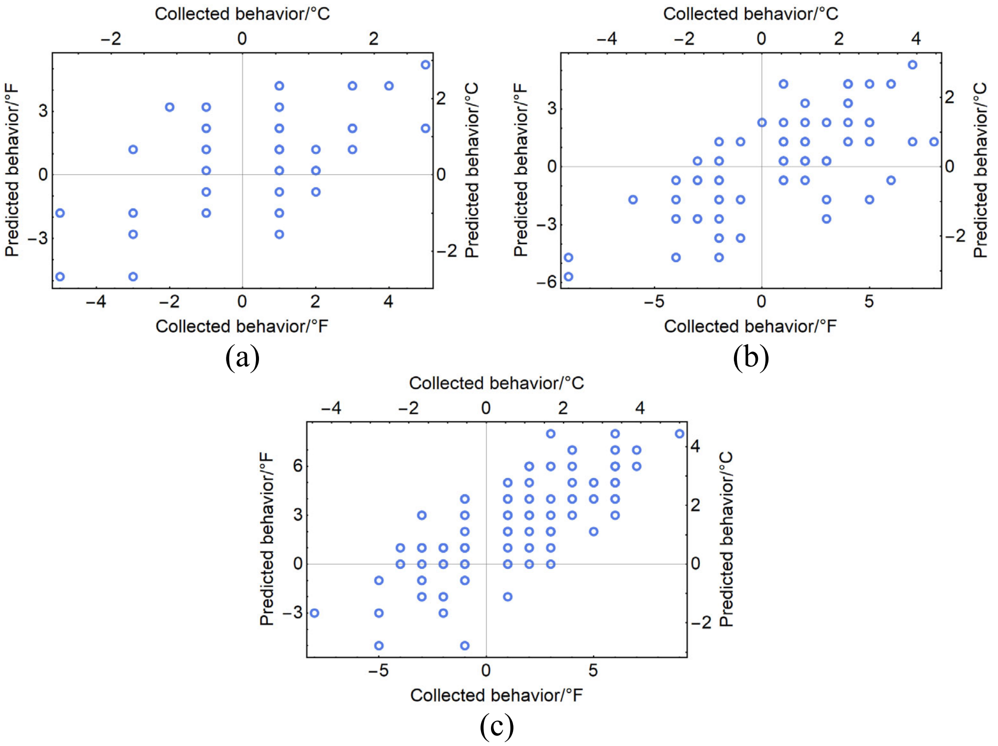
Fig. 11. Comparison of collected data on the occupant behavior of adjusting the thermostat set point and the RL model prediction for HLAB offices in 2018: (a) winter, (b) spring, (c) summer, and (d) fall.

Table 3. Prediction performance of the RL model for the HLAB offices

	R^2	MAE
Winter 2018	0.75	1.6
Spring 2018	0.79	1.9
Summer 2018	0.79	1.5
Fall 2018	0.81	1.7
Overall	0.79	1.68

3.4 Results of transfer learning model

After validating the RL model for the HLAB offices, we used the transfer learning model to predict occupant behavior in four other office buildings on the Purdue University campus. Fig. 12 shows the collected occupant behavior data and the RL model prediction in three seasons. The overall R^2 was 0.7, and the MAE was 1.7 °F. The results were not as good as the model validation results for the same building, presented in Section 3.3, but it was a feasible method for predicting occupant behavior for the different buildings without data.



(a) (b)

(c)

Fig. 12. Comparison between collected behavior data and behavior predicted by the RL model in four other Purdue University office buildings in 2018 in (a) summer, (b) fall, and (c) winter.

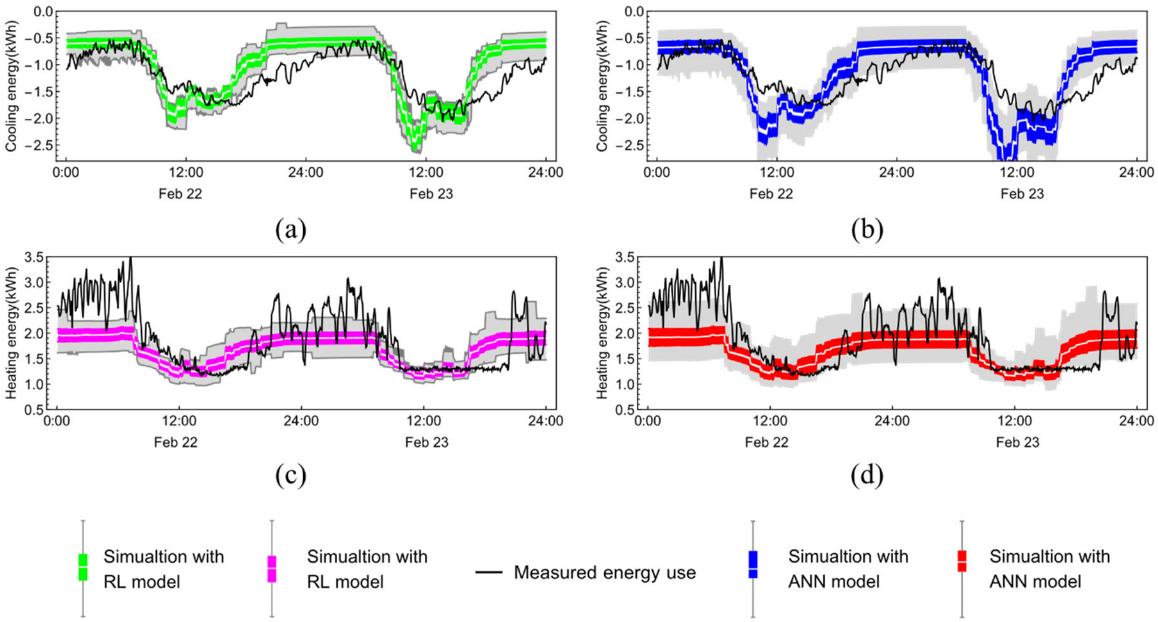
We also used the defined reward in Eqs. (10)–(12) to train the RL model again for residential buildings. Table 4 shows the prediction performance of the transfer learning model. In the residential buildings, the R^2 was between 0.6 and 0.7 in the four seasons, and the MAE varied from 2.1 °F to 2.9 °F. The results were worse than for the transfer learning in the other four office buildings. The reason was that the cross-type prediction was more difficult than cross-building prediction. In the residential buildings, there were many factors that impacted the occupant behavior differently than in the office buildings [71, 72] but were not considered in the current RL model. One feasible way to further improve the transfer learning model would be to introduce more impact factors in the state space, in addition to re-modeling the reward function. Furthermore, the quality and quantity of collected data in the residential buildings were not as good as in the office buildings because we used questionnaire surveys in the former. Recording accurate occupant behavior data with corresponding environmental parameters and incorporating the impact factors are directions for improvement in further studies of residential buildings.

Table 4. Prediction performance of the transfer learning model from the HLAB building to residential buildings

Season	R^2	MAE
Winter	0.67	2.1
Spring	0.61	2.9
Summer	0.69	2.3
Fall	0.67	2.7

3.5 Energy analysis with the RL occupant behavior model

After using the transfer learning model to predict occupant behavior in different buildings, we compared the collected heating and cooling energy use data and the simulation with the RL model in the HLAB building, for two days in winter. In Fig. 13, the box-and-whisker charts represent the simulation results with the use of the RL model and the ANN model. The black curve represents the measured data. For most of the time, the measured energy fluctuated within the lower and upper bounds predicted by the RL model. However, the variation range predicted by the RL model was narrower than that predicted by the ANN model. Table 3 lists the average heating and cooling loads and standard deviations for different seasons in one year. The reason for the difference between models was that the logic of the RL model was to improve the thermal comfort level of occupants. Therefore, the predicted occupant behavior was mostly reasonable. The model could not simulate illogical and extreme behavior such as adjusting the thermostat set point to the highest or lowest value for quick heating or cooling [74]. Such behavior can waste a lot of energy.



(a) (b)

(c) (d)

Fig. 13. Comparison of the collected heating and cooling energy use data and the simulation of manual thermostat control with the RL model in the HLAB building for two days in winter.

Table 5. Comparison of measured data with the heating and cooling loads (kWh) simulated by the ANN and RL models in four seasons.

Load		Winter	Spring	Summer	Fall
Heating	Measurement	3396	2833	2102	3183
	Simulation using ANN model	3526±108	2925±110	2275±35	3298±68
	Simulation using RL model	3084±67	2948±41	2239±27	3067±24
Cooling	Measurement	857	2261	2725	1205
	Simulation using ANN model	902±170	2006±115	2597±42	1136±90
	Simulation using RL model	863±72	1812±56	2570±30	974±30

We also used the transfer learning RL model to predict the energy use with thermostat setback and occupancy control. Fig. 14 shows all the energy simulation results in summer. The measurement and simulation using actual behavior exhibited little divergence. Thermostat setback and occupancy control could reduce energy use by about 30% and 70%, respectively. The average energy simulation results using the RL model were almost the same as with the ANN model, but the variation was less with the former model; this finding was similar to the results in Table 3. Hence, it is feasible to use the transfer learning RL model to predict the energy use in other buildings with various HVAC control systems.

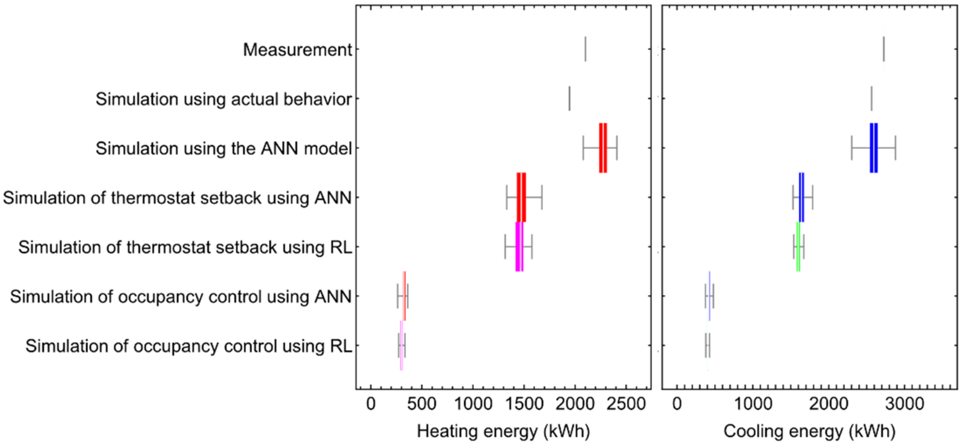


Fig. 14. Comparison of the measured heating and cooling loads and the results simulated by different models with thermostat setback and occupancy control in summer.

4 Discussion

In this study, we built an RL model to predict comfort-related occupant behavior in office buildings, and validated the model with collected data. We also used transfer learning for cross-building occupant behavior modelling. Although various impact factors were modelled in state space, including indoor air temperature and relative humidity, room occupancy and time, we neglected factors such as gender [75], cultural background [76], and age [4]. To improve the model's performance and widen its applicability, we need to determine the quantitative relationship between these factors and the occupant behavior for reward modelling in future studies. In the MDP, the time step size for occupant behavior prediction was 15 minutes. Thus, the impact of occupant behavior on the HVAC system and indoor environment was not immediate; rather, it was somewhat delayed. We assumed that the action could take effect in the subsequent time step if the HVAC system was in normal operation. Actually, based on the collected data and observation [17], after adjusting their behavior, the occupants tended to wait for a while, being aware of the HVAC response time. Even though the neutral TSV had not been reached, no occupant behavior occurred during this waiting time. If an occupant waited for a long time, such as 3–4 time steps, and still did not feel neutral, then there may have been issues with the HVAC control system or air handing units. In this case, the occupant behavior would be very complicated and personalized, including complaining and making another adjustment, this time to an extreme high or low set point. To improve the learning process and model performance, possible rewards could account for abnormal HVAC operations with longer response time and more time steps. Improving thermal comfort and energy efficiency behavior modelling is a potential direction for our future research.

In this study, we assumed that the occupant behavior and TSV decisions were based on the current indoor environment. This assumption was similar to those in the most recognized PMV thermal comfort model. According to the adaptive thermal comfort model, the outdoor climate and past thermal history may influence occupants' thermal preference and behavior. This could explain some of the prediction discrepancy exhibited by the current RL occupant behavior model, which was a limitation in the current study. Furthermore, the adaptive thermal comfort model has usually been applied to naturally ventilated rooms. In this study, the buildings were all mechanically ventilated. If we assumed adaptive thermal comfort and considered the outdoor climate and past thermal history, we could still build the MDP and introduce these factors in the state and reward. In this case, the model would be more complex. We could apply the adaptive thermal comfort theory and use historical states in the RL model to improve the prediction result as a future research direction. In the present study, we defined the reward as the difference between initial and final TSV as shown in Eqs. (5)–(7). Such definition was result-oriented and path-independent, because the middle terms could be canceled if there were many adjustment behaviors. Thus, the occupants could find the set point that maximized the cumulative reward in different ways, which increased the variation in occupant behavior. However, this study considered only comfort-related occupant behavior and not energy-related behavior in offices. This was because the cost of maintaining a comfortable environment in an office is typically not on the minds of occupants [17]. For simulation of energy-saving occupant behavior in other kinds of buildings, the RL model would also require energy parameters for the state space and reward modelling, such as heating and cooling rates and air change rate [77]. Finally, the RL model and transfer learning in this study exhibited good generalization capability and scalability. These models also have potential for other kinds of occupant behavior, such as interactions with windows [24], shades [19], lighting [78] and other indoor appliances.

With the RL model, we tried to model and predict the occupant behavior without collecting data but rather by building a policy-based MDP. We also used transfer learning to obtain the occupant behavior in other office buildings and in residential buildings with different HVAC systems and very limited information. This cross-building occupant behavior transfer was extremely difficult in the data-driven models. Therefore, the generalization capability of the RL and transfer learning models was better than that of the regression models. Meanwhile, the better generalization capability of the RL model may indicate a lesser ability to make predictions for specific buildings. As a result, the prediction accuracy of the RL model may not be as good as that of the data-driven models.

5 Conclusion

This study built and validated an RL occupant behavior model for an office building and transferred it to other buildings with thermostat setback and occupancy control. We also compared the energy use simulated by the RL model with measured data and predictions by the ANN model for the HLAB offices and four other office buildings on the Purdue University campus. This investigation led to the following conclusions:

1. The policy-based RL occupant behavior model trained by Q-learning was able to learn the logic of occupant behavior and predict the behavior accurately. The results for prediction of set point adjustment exhibited an R^2 around 0.8 and MAE less than 2 °F.
2. Transfer learning successfully transferred the logic and part of the occupant behavior model structure to other buildings with different HVAC control systems, such as thermostat setback and occupancy control. We also transferred the RL model from office buildings to residential buildings with a modification to the impact of air temperature on occupant behavior. The prediction performance was

good, with R^2 above 0.6 and MSE less than 2 °F. These transfer learning models did not require data collection. Unlike data-driven models, the transfer learning RL model had physical meaning and strong generalization capability.

3. The results of energy simulation for thermostat manual control, setback and occupancy control with the use of the RL model were similar to the results with the ANN model. The RL simulation accurately reflected the impact of occupant behavior on building energy use, but the variation predicted by the RL model was less than that predicted by the ANN model.

Acknowledgments

The authors would like to thank Dr. Orkan Kurtulus in the Center for High Performance Buildings at Purdue University for his assistance in setting the building automation system in the HLAB building. We would also like to thank all the occupants of the HLAB offices for their participation and assistance in obtaining the data reported in this study, and Blaine Miller and Chris Sorenson in the Utility Plant Office of Purdue University for providing data in four Purdue buildings. The data collection in this study was approved by Purdue University Institutional Review Board Protocol # 1704019079.

Conflict of Interest

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

References

[1] US Department of Energy, Building energy data. (2011).

[2] De Wilde, Pieter. "The gap between predicted and measured energy performance of buildings: A framework for investigation." Automation in Construction 41 (2014): 40-49. https://doi.org/10.1016/j.autcon.2014.02.009

[3] Zou, Patrick XW, Xiaoxiao Xu, Jay Sanjayan, and Jiayuan Wang. "Review of 10 years research on building energy performance gap: Life-cycle and stakeholder perspectives." Energy and Buildings 178 (2018): 165-181. https://doi.org/10.1016/j.enbuild.2018.08.040

[4] Zhang, Yan, Xuemei Bai, Franklin P. Mills, and John CV Pezzey. "Rethinking the role of occupant behavior in building energy performance: A review." Energy and Buildings 172 (2018): 279-294. https://doi.org/10.1016/j.enbuild.2018.05.017

[5] D'Oca, Simona, Tianzhen Hong, and Jared Langevin. "The human dimensions of energy use in buildings: A review." Renewable and Sustainable Energy Reviews 81 (2018): 731-742. https://doi.org/10.1016/j.rser.2017.08.019

[6] Sun, Kaiyu, and Tianzhen Hong. "A framework for quantifying the impact of occupant behavior on energy savings of energy conservation measures." Energy and Buildings 146 (2017): 383-396. https://doi.org/10.1016/j.enbuild.2017.04.065

[7] Hong, Tianzhen, Sarah C. Taylor-Lange, Simona D'Oca, Da Yan, and Stefano P. Corgnati. "Advances in research and applications of energy-related occupant behavior in

buildings." Energy and Buildings 116 (2016): 694-702. https://doi.org/10.1016/j.enbuild.2015.11.052

[8] Paone, Antonio, and Jean-Philippe Bacher. "The impact of building occupant behavior on energy efficiency and methods to influence it: A review of the state of the art." Energies 11, no. 4 (2018): 953. https://doi.org/10.3390/en11040953

[9] Yan, Da, William O'Brien, Tianzhen Hong, Xiaohang Feng, H. Burak Gunay, Farhang Tahmasebi, and Ardeshir Mahdavi. "Occupant behavior modeling for building performance simulation: Current state and future challenges." Energy and Buildings 107 (2015): 264-278. https://doi.org/10.1016/j.enbuild.2015.08.032

[10] Hong, Tianzhen, Jared Langevin, and Kaiyu Sun. "Building simulation: Ten challenges." In Building Simulation, vol. 11, no. 5, pp. 871-898. Tsinghua University Press, 2018. https://doi.org/10.1007/s12273-018-0444-x

[11] Hong, Tianzhen, Da Yan, Simona D'Oca, and Chien-fei Chen. "Ten questions concerning occupant behavior in buildings: The big picture." Building and Environment 114 (2017): 518-530. https://doi.org/10.1016/j.buildenv.2016.12.006

[12] R.V. Andersen, B.W. Olesen, J. Toftum, "Modelling occupants' heating set-point preferences," in: Building Simulation Conference, 2011, pp. 14–16.

[13] Fabi, Valentina, Rune Vinther Andersen, and Stefano Paolo Corgnati. "Influence of occupant's heating set-point preferences on indoor environmental quality and heating demand in residential buildings." HVAC&R Research 19, no. 5 (2013): 635-645. https://doi.org/ 10.1080/10789669.2013.789372

[14] Langevin, Jared, Jin Wen, and Patrick L. Gurian. "Simulating the human-building interaction: Development and validation of an agent-based model of office occupant behaviors." Building and Environment 88 (2015): 27-45. https://doi.org/10.1016/j.buildenv.2014.11.037

[15] Pfafferott, J., and S. Herkel. "Statistical simulation of user behaviour in low-energy office buildings." Solar Energy 81, no. 5 (2007): 676-682.https://doi.org/10.1016/j.buildenv.2014.11.037

[16] Sun, Kaiyu, and Tianzhen Hong. "A simulation approach to estimate energy savings potential of occupant behavior measures." Energy and Buildings 136 (2017): 43-62. https://doi.org/10.1016/j.enbuild.2016.12.010

[17] Deng, Zhipeng, and Qingyan Chen. "Artificial neural network models using thermal sensations and occupants' behavior for predicting thermal comfort." Energy and Buildings 174 (2018): 587-602. https://doi.org/10.1016/j.enbuild.2018.06.060

[18] Wang, Zhe, and Tianzhen Hong. "Reinforcement learning for building controls: The opportunities and challenges." Applied Energy 269 (2020): 115036. https://doi.org/10.1016/j.apenergy.2020.115036

[19] O'Brien, William, and H. Burak Gunay. "The contextual factors contributing to occupants' adaptive comfort behaviors in offices—A review and proposed modeling framework." Building and Environment 77 (2014): 77-87. https://doi.org/10.1016/j.buildenv.2014.03.024

[20] Stazi, Francesca, Federica Naspi, and Marco D'Orazio. "A literature review on driving factors and contextual events influencing occupants' behaviours in buildings." Building and Environment 118 (2017): 40-66. https://doi.org/10.1016/j.buildenv.2017.03.021

[21] Hong, Tianzhen, Simona D'Oca, William JN Turner, and Sarah C. Taylor-Lange. "An ontology to represent energy-related occupant behavior in buildings. Part I: Introduction to

the DNAs framework." Building and Environment 92 (2015): 764-777. https://doi.org/10.1016/j.buildenv.2015.02.019

[22] O'Brien, William, Isabella Gaetani, Sara Gilani, Salvatore Carlucci, Pieter-Jan Hoes, and Jan Hensen. "International survey on current occupant modelling approaches in building performance simulation." Journal of Building Performance Simulation 10, no. 5-6 (2017): 653-671. https://doi.org/10.1080/19401493.2016.1243731

[23] Jia, Mengda, Ravi S. Srinivasan, and Adeeba A. Raheem. "From occupancy to occupant behavior: An analytical survey of data acquisition technologies, modeling methodologies and simulation coupling mechanisms for building energy efficiency." Renewable and Sustainable Energy Reviews 68 (2017): 525-540. https://doi.org/10.1016/j.rser.2016.10.011

[24] Fabi, Valentina, Rune Vinther Andersen, Stefano Corgnati, and Bjarne W. Olesen. "Occupants' window opening behaviour: A literature review of factors influencing occupant behaviour and models." Building and Environment 58 (2012): 188-198. https://doi.org/10.1016/j.buildenv.2012.07.009

[25] Tijani, Khadija, Stephane Ploix, Benjamin Haas, Julie Dugdale, and Quoc Dung Ngo. "Dynamic Bayesian Networks to simulate occupant behaviours in office buildings related to indoor air quality." arXiv preprint arXiv:1605.05966 (2016). https://arxiv.org/ftp/arxiv/papers/1605/1605.05966.pdf

[26] D'Oca, Simona, Stefano Corgnati, and Tianzhen Hong. "Data mining of occupant behavior in office buildings." Energy Procedia 78 (2015): 585-590. https://doi.org/10.1016/j.egypro.2015.11.022

[27] Zhou, Xin, Tiance Liu, Da Yan, Xing Shi, and Xing Jin. "An action-based Markov chain modeling approach for predicting the window operating behavior in office spaces." In Building Simulation, pp. 1-15. Tsinghua University Press, 2020. https://doi.org/10.1007/s12273-020-0647-9

[28] Andrews, Clinton J., Daniel Yi, Uta Krogmann, Jennifer A. Senick, and Richard E. Wener. "Designing buildings for real occupants: An agent-based approach." IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 41, no. 6 (2011): 1077-1091. https://doi.org/10.1109/TSMCA.2011.2116116

[29] Reinhart, Christoph F. "Lightswitch-2002: A model for manual and automated control of electric lighting and blinds." Solar Energy 77, no. 1 (2004): 15-28. https://doi.org/10.1016/j.solener.2004.04.003

[30] Ryu, Seung Ho, and Hyeun Jun Moon. "Development of an occupancy prediction model using indoor environmental data based on machine learning techniques." Building and Environment 107 (2016): 1-9. https://doi.org/10.1016/j.buildenv.2016.06.039

[31] Zhou, Hao, Lifeng Qiao, Yi Jiang, Hejiang Sun, and Qingyan Chen. "Recognition of air-conditioner operation from indoor air temperature and relative humidity by a data mining approach." Energy and Buildings 111 (2016): 233-241. https://doi.org/10.1016/j.enbuild.2015.11.034

[32] Papadopoulos, Sokratis, and Elie Azar. "Integrating building performance simulation in agent-based modeling using regression surrogate models: A novel human-in-the-loop energy modeling approach." Energy and Buildings 128 (2016): 214-223. https://doi.org/10.1016/j.enbuild.2016.06.079

[33] Azar, Elie, and Carol C. Menassa. "Agent-based modeling of occupants and their impact on energy use in commercial buildings." Journal of Computing in Civil Engineering 26, no. 4 (2012): 506-518. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000158

[34] Lee, Yoon Soo, and Ali M. Malkawi. "Simulating multiple occupant behaviors in buildings: An agent-based modeling approach." Energy and Buildings 69 (2014): 407-416. https://doi.org/10.1016/j.enbuild.2013.11.020

[35] Sutton, R. S., & Barto, A. G. (1998). Introduction to reinforcement learning (Vol. 135). Cambridge: MIT Press.

[36] Lillicrap, Timothy P., Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. "Continuous control with deep reinforcement learning." arXiv preprint arXiv:1509.02971 (2015). https://arxiv.org/pdf/1509.02971.pdf

[37] Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert et al. "Mastering the game of go without human knowledge." nature 550, no. 7676 (2017): 354-359. https://doi.org/10.1038/nature24270

[38] Zhang, Zhiang, Adrian Chong, Yuqi Pan, Chenlu Zhang, and Khee Poh Lam. "Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning." Energy and Buildings 199 (2019): 472-490. https://doi.org/10.1016/j.enbuild.2019.07.029

[39] Kazmi, Hussain, Johan Suykens, Attila Balint, and Johan Driesen. "Multi-agent reinforcement learning for modeling and control of thermostatically controlled loads." Applied energy 238 (2019): 1022-1035. https://doi.org/10.1016/j.apenergy.2019.01.140

[40] Yu, Liang, Weiwei Xie, Di Xie, Yulong Zou, Dengyin Zhang, Zhixin Sun, Linghua Zhang, Yue Zhang, and Tao Jiang. "Deep reinforcement learning for smart home energy management." IEEE Internet of Things Journal 7, no. 4 (2019): 2751-2762. https://doi.org/10.1109/JIOT.2019.2957289

[41] Han, Mengjie, Ross May, Xingxing Zhang, Xinru Wang, Song Pan, Yan Da, and Yuan Jin. "A novel reinforcement learning method for improving occupant comfort via window opening and closing." Sustainable Cities and Society (2020): 102247. https://doi.org/10.1016/j.scs.2020.102247

[42] Han, Mengjie, Ross May, Xingxing Zhang, Xinru Wang, Song Pan, Da Yan, Yuan Jin, and Liguo Xu. "A review of reinforcement learning methodologies for controlling occupant comfort in buildings." Sustainable Cities and Society 51 (2019): 101748. https://doi.org/10.1016/j.scs.2019.101748

[43] Yoon, Young Ran, and Hyeun Jun Moon. "Performance based thermal comfort control (PTCC) using deep reinforcement learning for space cooling." Energy and Buildings 203 (2019): 109420. https://doi.org/10.1016/j.enbuild.2019.109420

[44] Ruelens, Frederik, Sandro Iacovella, Bert J. Claessens, and Ronnie Belmans. "Learning agent for a heat-pump thermostat with a set-back strategy using model-free reinforcement learning." Energies 8, no. 8 (2015): 8300-8318. https://doi.org/10.3390/en8088300

[45] Azuatalam, Donald, Wee-Lih Lee, Frits de Nijs, and Ariel Liebman. "Reinforcement learning for whole-building HVAC control and demand response." Energy and AI 2 (2020): 100020. https://doi.org/10.1016/j.egyai.2020.100020

[46] Chen, Bingqing, Zicheng Cai, and Mario 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, pp. 316-325. 2019. https://doi.org/10.1145/3360322.3360849

[47] Ding, Xianzhong, Wan Du, and Alberto E. Cerpa. "MB2C: Model-based deep reinforcement learning for multi-zone building control." In Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, pp. 50-59. 2020. https://doi.org/10.1145/3408308.3427986

[48] Jia, Ruoxi, Ming Jin, Kaiyu Sun, Tianzhen Hong, and Costas Spanos. "Advanced building control via deep reinforcement learning." Energy Procedia 158 (2019): 6158-6163. https://doi.org/10.1016/j.egypro.2019.01.494

[49] Chen, Yujiao, Leslie K. Norford, Holly W. Samuelson, and Ali Malkawi. "Optimal control of HVAC and window systems for natural ventilation through reinforcement learning." Energy and Buildings 169 (2018): 195-205. https://doi.org/10.1016/j.enbuild.2018.03.051

[50] Park, June Young, Thomas Dougherty, Hagen Fritz, and Zoltan Nagy. "LightLearn: An adaptive and occupant centered controller for lighting based on reinforcement learning." Building and Environment 147 (2019): 397-414. https://doi.org/10.1016/j.buildenv.2018.10.028

[51] Valladares, William, Marco Galindo, Jorge Gutiérrez, Wu-Chieh Wu, Kuo-Kai Liao, Jen-Chung Liao, Kuang-Chin Lu, and Chi-Chuan Wang. "Energy optimization associated with thermal comfort and indoor air control via a deep reinforcement learning algorithm." Building and Environment 155 (2019): 105-117. https://doi.org/10.1016/j.buildenv.2019.03.038

[52] Brandi, Silvio, Marco Savino Piscitelli, Marco Martellacci, and Alfonso Capozzoli. "Deep Reinforcement Learning to optimise indoor temperature control and heating energy consumption in buildings." Energy and Buildings (2020): 110225. https://doi.org/10.1016/j.enbuild.2020.110225

[53] Ding, Xianzhong, Wan Du, and Alberto 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, pp. 326-335. 2019. https://doi.org/10.1145/3360322.3360857

[54] Li, Ao, Fu Xiao, Cheng Fan, and Maomao Hu. "Development of an ANN-based building energy model for information-poor buildings using transfer learning." In Building Simulation, pp. 1-13. Tsinghua University Press, 2020. https://doi.org/10.1007/s12273-020-0711-5

[55] Mosaico, Gabriele, Matteo Saviozzi, Federico Silvestro, Andrea Bagnasco, and Andrea Vinci. "Simplified state space building energy model and transfer learning based occupancy estimation for HVAC optimal control." In 2019 IEEE 5th International forum on Research and Technology for Society and Industry (RTSI), pp. 353-358. IEEE, 2019. https://doi.org/10.1109/RTSI.2019.8895544

[56] Ali, SM Murad, Juan Carlos Augusto, and David Windridge. "A survey of user-centred approaches for smart home transfer learning and new user home automation adaptation." Applied Artificial Intelligence 33, no. 8 (2019): 747-774. https://doi.org/10.1080/08839514.2019.1603784

[57] Alam, Mohammad Arif Ul, and Nirmalya Roy. "Unseen activity recognitions: A hierarchical active transfer learning approach." In 2017 IEEE 37th International

Conference on Distributed Computing Systems (ICDCS), pp. 436-446. IEEE, 2017. https://doi.org/10.1109/ICDCS.2017.264

[58] Mocanu, Elena, Phuong H. Nguyen, Wil L. Kling, and Madeleine Gibescu. "Unsupervised energy prediction in a Smart Grid context using reinforcement cross-building transfer learning." Energy and Buildings 116 (2016): 646-655. https://doi.org/10.1016/j.enbuild.2016.01.030

[59] Ribeiro, Mauro, Katarina Grolinger, Hany F. ElYamany, Wilson A. Higashino, and Miriam AM Capretz. "Transfer learning with seasonal and trend adjustment for cross-building energy forecasting." Energy and Buildings 165 (2018): 352-363. https://doi.org/10.1016/j.enbuild.2018.01.034

[60] Gao, Nan, Wei Shao, Mohammad Saiedur Rahaman, Jun Zhai, Klaus David, and Flora D. Salim. "Transfer learning for thermal comfort prediction in multiple cities." arXiv preprint arXiv:2004.14382 (2020). https://arxiv.org/pdf/2004.14382.pdf

[61] Xu, Shichao, Yixuan Wang, Yanzhi Wang, Zheng O'Neill, and Qi Zhu. "One for many: Transfer learning for building HVAC control." In Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, pp. 230-239. 2020. https://doi.org/10.1145/3408308.3427617

[62] Deng, Zhipeng, and Qingyan Chen. "Development and validation of a smart HVAC control system for multi-occupant offices by using occupants' physiological signals from wristband." Energy and Buildings 214 (2020): 109872. https://doi.org/10.1016/j.enbuild.2020.109872

[63] Handbook, A.S.H.R.A.E. "Fundamentals, ASHRAE–American Society of Heating." Ventilating and Air-Conditioning Engineers (2017).

[64] Foerster, Jakob, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. "Learning to communicate with deep multi-agent reinforcement learning." In Advances in neural information processing systems, pp. 2137-2145. 2016.

[65] Klein, Laura, Jun-young Kwak, Geoffrey Kavulya, Farrokh Jazizadeh, Burcin Becerik-Gerber, Pradeep Varakantham, and Milind Tambe. "Coordinating occupant behavior for building energy and comfort management using multi-agent systems." Automation in Construction 22 (2012): 525-536. https://doi.org/10.1016/j.autcon.2011.11.012

[66] Melo, Francisco S. "Convergence of Q-learning: A simple proof." Institute Of Systems and Robotics, Tech. Rep (2001): 1-4.

[67] Yang, Lei, Zoltan Nagy, Philippe Goffin, and Arno Schlueter. "Reinforcement learning for optimal control of low exergy buildings." Applied Energy 156 (2015): 577-586. https://doi.org/10.1016/j.apenergy.2015.07.050

[68] Cheng, Zhijin, Qianchuan Zhao, Fulin Wang, Yi Jiang, Li Xia, and Jinlei Ding. "Satisfaction based Q-learning for integrated lighting and blind control." Energy and Buildings 127 (2016): 43-55. https://doi.org/10.1016/j.enbuild.2016.05.067

[69] https://www.mathworks.com/help/reinforcement-learning/

[70] Gunay, H. Burak, William O'Brien, and Ian Beausoleil-Morrison. "A critical review of observation studies, modeling, and simulation of adaptive occupant behaviors in offices." Building and Environment 70 (2013): 31-47. https://doi.org/10.1016/j.buildenv.2013.07.020

[71] Wei, Shen, Rory Jones, and Pieter De Wilde. "Driving factors for occupant-controlled space heating in residential buildings." Energy and Buildings 70 (2014): 36-44. https://doi.org/10.1016/j.enbuild.2013.11.001

[72] Yu, Zhun, Benjamin CM Fung, Fariborz Haghighat, Hiroshi Yoshino, and Edward Morofsky. "A systematic procedure to study the influence of occupant behavior on building energy consumption." Energy and Buildings 43, no. 6 (2011): 1409-1417. https://doi.org/10.1016/j.enbuild.2011.02.002

[73] Standard, A.S.H.R.A.E. "Standard 55-2010, Thermal environmental conditions for human occupancy." American Society of Heating, Refrigerating and Air Conditioning Engineers (2010).

[74] Deng, Zhipeng, and Qingyan Chen. "Simulating the impact of occupant behavior on energy use of HVAC systems by implementing a behavioral artificial neural network model." Energy and Buildings 198 (2019): 216-227. https://doi.org/10.1016/j.enbuild.2019.06.015

[75] Karjalainen, Sami. "Gender differences in thermal comfort and use of thermostats in everyday thermal environments." Building and Environment 42, no. 4 (2007): 1594-1603. https://doi.org/10.1016/j.buildenv.2006.01.009

[76] Montazami, Azadeh, Mark Gaterell, Fergus Nicol, Mark Lumley, and Chryssa Thoua. "Impact of social background and behaviour on children's thermal comfort." Building and Environment 122 (2017): 422-434. https://doi.org/10.1016/j.buildenv.2017.06.002

[77] Ghahramani, Ali, Kanu Dutta, and Burcin Becerik-Gerber. "Energy trade off analysis of optimized daily temperature setpoints." Journal of Building Engineering 19 (2018): 584-591. https://doi.org/10.1016/j.jobe.2018.06.012

[78] Yan, Da, Xiaohang Feng, Yuan Jin, and Chuang Wang. "The evaluation of stochastic occupant behavior models from an application-oriented perspective: Using the lighting behavior model as a case study." Energy and Buildings 176 (2018): 151-162. https://doi.org/10.1016/j.enbuild.2018.07.037

Highlights

1. Reinforcement learning model for predicting occupant behavior in adjusting thermostat set point and clothing level in an office building.
2. Transfer learning model for transferring occupant behavior from one building to another without data.
3. Transfer learning among buildings of the same type was better than among different types of buildings.
4. The variation range of energy use predicted by the reinforcement learning model was smaller than that predicted by the artificial neural network model.