A Comparison Between Common and Reinforcement Learning-Based Supply Air Temperature Reset Strategies with Varying Occupant Temperature Preferences

Hussein ELEHWANY[1]*, Burak GUNAY[2], Mohamed OUF[3], Nunzio COTRUFO[4], Jean-Simon VENNE[5], Junfeng WEN[6]

[1]Carleton University, Department of Civil and
Environmental Engineering, Ottawa, Ontario, Canada
613-520-5784, husseinelehwany@cmail.carleton.ca

[2]Carleton University, Department of Civil and
Environmental Engineering, Ottawa, Ontario, Canada
613-520-5784, burakgunay@cunet.carleton.ca

[3]Concordia University, Department of Building, Civil and
Environmental Engineering, Montreal, Quebec, Canada
514-848-2424, mohamed.ouf@concordia.ca

[4]BrainBox AI, Montreal, Quebec, Canada
888-585-2630, n.cotrufo@brainboxai.com

[5]BrainBox AI, Montreal, Quebec, Canada
888-585-2630, js.venne@brainboxai.com

[6]Carleton University, School of Computer Science,
Ottawa, Ontario, Canada
613-520-4333, junfengwen@cunet.carleton.ca

* Corresponding Author

ABSTRACT

The supply air temperature (SAT) of an air handling unit in multi-zone variable air volume systems could impact the energy use significantly. Formerly, buildings used a constant SAT which resulted in high energy consumption due to the increased load on perimeter heaters. Lately, ASHRAE guideline 36 introduced the trim and respond logic (Taylor, 2015) as an improved SAT reset strategy which depends on the feedback of the cooling requests of the zones. However, the trim and respond logic might fail to provide comfort and/or energy savings in case of higher demands, conflicting thermal preferences at different zones and varying occupancy patterns. This study investigates four SAT reset strategies: 1) constant 13℃ SAT, 2) SAT reset based on outdoor air temperature (OAT), 3) trim and respond, and 4) trim and respond combined with OAT reset; with different cases of varying zone setpoints. It also introduces a deep Q-network (DQN) reinforcement learning (RL) algorithm for SAT reset and compares its performance with the other strategies. All the cases are simulated using EnergyPlus. The objective is to address the shortcomings of the currently adopted methods in industry and to show the potential of reinforcement learning in HVAC controls. The results show that the common SAT reset strategies do not perform well with cases of varying setpoint leading to either higher energy cost or decrease in occupant comfort, while the DQN-based method provided a better alternative. These findings establish a basis for future work that would focus on developing a multi-agent occupant centric control (OCC) method that takes energy and occupant comfort into account by utilizing RL methods.

1. INTRODUCTION

Each One of the most common HVAC systems in commercial buildings is multi-zone variable air volume (VAV) air handling units (AHU) (Taylor, 2018). VAV AHU systems vary the air flow rate supplied to each zone to control the indoor thermal conditions and to provide sufficient ventilation. Originally, air is supplied at a constant temperature however this usually leads to high energy consumption (Torabi et al., 2022). Another common method is resetting the supply air temperature (SAT) based on the outdoor air temperature. ASHRAE Guideline 36 provided a detailed strategy for SAT setpoint reset that takes occupant comfort into consideration and consumes less energy compared to the former methods (ASHRAE, 2021).

Many buildings depend on constant thermostat setpoints and occupancy schedules for their operation. Some buildings that have access to occupancy data and occupants feedback, take average occupant counts to create occupancy schedules and average thermostat setpoints to decide the appropriate heating and cooling setpints. However, occupants comfort preferences exhibit significant variation (Kim et al., 2018) and averaging out occupant preferences compromises the comfort of individuals (Park et al., 2019). The performance of the SAT reset strategy recommended by the ASHRAE Guideline 36 (ASHRAE, 2021) has been evaluated in many studies under uniform setpoints and regular occupancy patterns. Raftery et al., (2018) , for instance, developed an SAT reset strategy that used feedback from cooling requests and energy use, achieving 29% of energy savings. However, there is a research gap in studying the common SAT reset strategies under varying zone setpoints and evaluating their performance in terms of energy use and occupant comfort.

Recently, occupant-centric controls (OCC) have emerged as an approach that uses measurement from building, occupancy data and occupant feedback to provide a more comfortable and energy efficient control scheme (Nagy et al., 2023). Reinforcement learning (RL) is a promising machine learning method for OCC as it does not require a detailed physical model, such as the model-free RL methods (Wang & Hong, 2020). Moreover, RL depends on continuous interactions with the environment to reach an optimal solution without the need for expert knowledge (Han et al., 2019).

Some studies utilized RL methods to address the SAT setpoint reset problem. (Fang et al., 2022) used a deep Q-network (DQN) RL algorithm to control the SAT and supply water temperature setpoints, while (Jia et al., 2019) used a policy gradient RL method for SAT setpoint control as well. (Lu et al., 2023) implemented an optimization-based controller and deep RL controller and compared their performance with ASHRAE Guideline 36 sequence of operation.

However, no studies were found in the literature that analyzed the RL-based SAT methods under cases of varying setpoints. Therefore, this paper has the following objectives:

- Compare the four common SAT reset strategies with cases of varying setpoints and evaluate their energy usage and occupant comfort performance.
- Introduce a DQN RL-based SAT reset strategy as a potential alternative and compare its performance with cases of non-uniform setpoints.

2. METHODOLOGY

2.1 Simulation Environment

A nine-zone model of a single floor is developed and simulated using EnergyPlus. Each zone is 5m-by-5m and all exterior walls have 3m-by-1.5m windows with a window-to-wall ratio of 30%, as shown in Figure 1. The floor and the ceiling are adiabatic as the heat exchange between floors is neglected assuming that different floors have similar conditions. The whole floor is served by an Air Handling Unit (AHU) and each zone has a Variable Air Volume (VAV) unit with reheat coils. All zones are equipped with baseboard heaters as well. The minimum outdoor air ventilation rate is specified based on ASHRAE Standard 62.1 (ASHRAE, 2019) as 2.5 L/s per person and 0.3 L/s per square meter. The heating and cooling elements are available throughout the year and the AHU operates between 04:00 and 19:00 on weekdays. A setback and setup temperatures of 18℃ and 27℃ are implemented. The model is equipped with a differential dry bulb economizer with a limit set in accordance to ASHRAE Standard 90.1 (ASHRAE, 2022). Moreover, dual maximum logic for the maximum air flow setpoints is enabled in the model by setting the “damper heating action” of the VAVs to “reverse”. All simulations are done with weather conditions of Ottawa, Canada in 2019. Table 1 summarizes the model parameters.

In order to have consistency with all the simulated cases, the auto-sizing function of EnergyPlus is disabled and the capacities of the heating and cooling elements, and the baseboard heaters are fixed, as well as the maximum and minimum air flow rates, in accordance with ASHRAE Standard 62.1 (ASHRAE, 2019). The sizes of the heating and cooling elements as well as the maximum air flow rates are increased to make sure that the HVAC system is capable of meeting extreme temperature setpoints.

Different SAT setpoint reset strategies are implemented into EnergyPlus with aid of the Python API that allows continuous communication between the simulation and the python script. The Python API allows reading information at every time step such as indoor temperatures. Python functions are used to implement the different SAT reset strategies. At each simulation step, the API writes the calculated SAT values to an EnergyPlus schedule, that is assigned to the supply air temperature of the AHU.

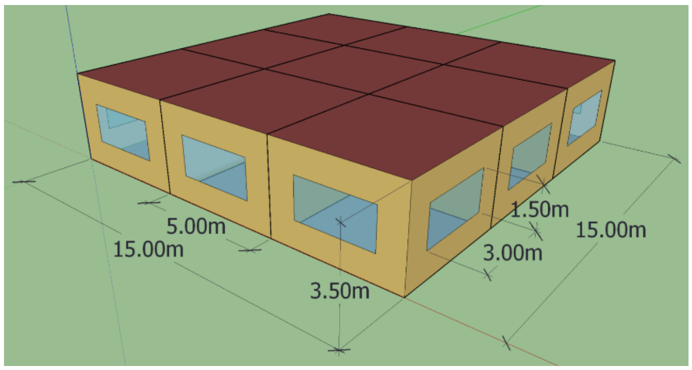

Figure 1: The simulated EnergyPlus model consisting of nine zones.

Table 1: Model parameters.

Parameter	Value
Exterior wall's R value	4 m^2K/W
Window's U-value	2 W/m^2K
Window's SHGC	0.5
Infiltration rate	0.25 L/sm^2
Occupants per zone	1
Lighting load	8.5 W/m^2
Equipment load	7.5 W/m^2

2.2 Common Supply Air Temperature Reset Strategies

The following four SAT setpoint reset strategies are studied in this paper.

- **Constant supply air temperature:** This strategy depends on supplying air at a constant low temperature, usually 13℃, to be able to meet the cooling demands. However, this method leads to unnecessarily high energy consumption because it depends mainly on the heating elements at the zone level to meet the heating demands during the winter season.
- **Supply air temperature reset based on outdoor air temperature:** The SAT varies between SAT_{min} and SAT_{max} depending on the outdoor air temperature (OAT) as shown in Figure 2(a). The SAT follows a linear relation between the OAT_{min} and OAT_{max}. While the default SAT_{max} in ASHRAE Guideline 36 is 18℃, this study uses a higher SAT_{max} of 20℃ to improve the utilization of the AHU's heating energy with low-density occupancy profiles.
- **Trim and respond:** The ASHRAE Guideline 36 introduced the trim and respond logic as a SAT setpoint reset strategy (ASHRAE, 2021). It increases the SAT by a "trim" amount at every time step until reaching SAT_{max}. If cooling is required by some zones, the system triggers cooling requests. Beyond a certain limit of ignored requests, the SAT is reduced by a "respond" amount. The respond amount is proportional to the difference between the cooling requests and the ignored requests, with a maximum allowable respond limit. The SAT is allowed to fluctuate between SAT_{min} and SAT_{max}, as

shown in Figure 2(b). In this study, the trim, respond and maximum respond amounts are chosen to be 0.25℃, -0.5℃, and -1.5℃, respectively. The ASHRAE Guideline 36 has a two-minute time step, as opposed to the five-minute time step of this study, therefore the trim and respond amounts are scaled accordingly to match the new time step. Equation (1) shows how the cooling requests are calculated.

- **Trim and respond with outdoor air temperature:** ASHRAE Guideline 36 recommends an SAT reset strategy that combines trim and respond logic and the OAT reset strategy, as shown in Figure 2(c).

$$cooling\ requests = \begin{cases} 3, & if\ T_{in} > T_{clgsp} + 3℃ \\ 2, & if\ T_{in} > T_{clgsp} + 2℃ \\ 1, & if\ cooling\ loop > 95\%, until < 85\% \end{cases} \tag{1}$$

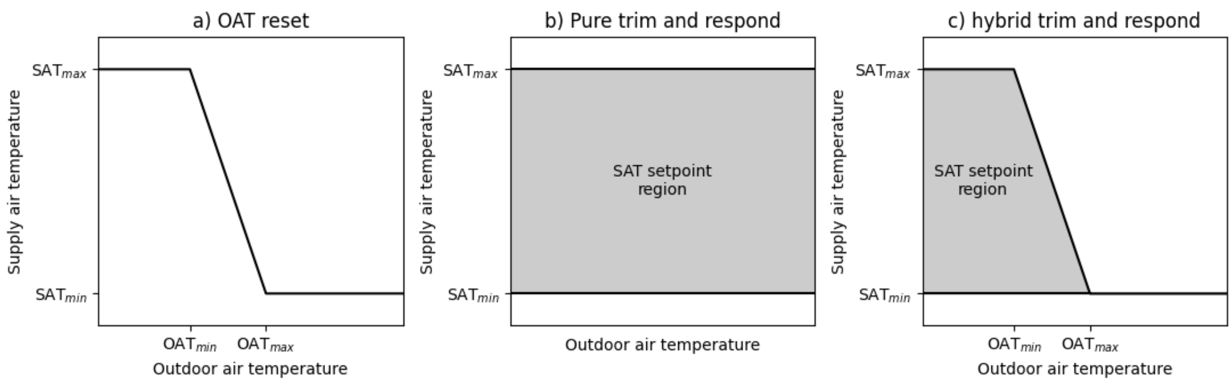

Figure 2: Supply air temperature setpoint reset based on a) OAT reset, b) pure trim and respond, and c) hybrid trim and respond.

2.3 DQN-based SAT Reset Strategy

RL based control methods interact with the environment to reach an optimal solution, called the optimal policy. An RL agent receives observations (states) from the environment and takes actions that affect the environment. At every step a reward is calculated based on the taken action and the reached state. Each state gets a value based on the received immediate reward and the value of states visited in the future. In tabular RL methods, a table contains the values of all state-action pairs and is updated during the training process. Tabular RL methods are more effective with problems of limited state space. Since this study involves a multi-zone building, the number of states could be too many, therefore non-tabular methods, such as DQN (Mnih et al., 2013), are more effective. DQN uses artificial neural networks (ANN) to approximate the value function of the state-action pairs. The DQN SAT reset problem is formulated as follows:

- **States:** The state variables are the indoor temperatures of the nine zones, the outdoor air temperature, and the time of the day. The indoor temperatures and the OAT are defined as continuous variables, while the time of the day is a discrete variable. During the development of the algorithm, it was found that the OAT has a significant role in improving the learning process as it helps the agent to distinguish between the seasons. Furthermore, including time of the day in the state space improved the stability of the algorithm.
- **Actions:** The agent is allowed to control the SAT setpoint by choosing one of the following actions: (13, 14, 15, 16, 17, 18, 19, 20℃).
- **Reward:** Equation (2) is used to calculate the reward at every time step. α and β are factors that penalize discomfort and energy use, respectively. α is fixed while β is tuned to achieve a desired performance. The energy use is the sum of all the energy meters normalized per unit of floor area. T_{dev} quantifies the amount of temperature deviation from the comfort range defined by the heating setpoint ($T_{htgsp,i}$) and cooling setpoint ($T_{clgsp,i}$) as shown in Equation (3). The sum of all the deviations from all the zones is used, giving equal weight to each zone.

- **Exploration:** The algorithm was allowed to randomly choose any action (explore) with a probability ε. The value of ε starts at 90% at the beginning of the simulation and decays exponentially reaching 1%.

$$reward = -\alpha * T_{dev} - \beta * energy\ use \tag{2}$$

$$T_{dev} = \begin{cases} \sum_{i=1}^{9} \left|T_{in,i} - T_{htgsp,i}\right| & if\ T_{in,i} < T_{htgsp,i} \\ \sum_{i=1}^{9} \left|T_{in,i} - T_{clgsp,i}\right| & if\ T_{in,i} > T_{clgsp,i} \\ 0 & otherwise \end{cases} \tag{3}$$

Table 2: DQN algorithm parameters

Parameter	Value
Learning rate	$1e^{-4}$
Discount factor, γ	0.99
Exploration rate, ε	$0.01 + (0.9 - 0.01)e^{-0.001t}$
Batch size	128
ANN layers	3
Nodes in ANN hidden layer	128
Optimizer	AdamW

Table 2 lists the parameters used with the DQN algorithm. The discount factor, which ranges from zero to one, is 0.99 to put emphasis on future rewards. The DQN algorithm consists of two ANNs: a policy net and a target net. The policy net is used to choose the optimal actions and is updated during the training process. The target net is used to estimate the best possible value for the upcoming state and is updated at a slower rate than the policy net. This ensures a stable training process (Lillicrap et al., 2015). Moreover, the DQN algorithm stores all the history to a buffer and samples a random batch (here 128 samples) from the gathered experiences at every time step to train the ANN.

2.4 Varying Setpoint Cases

Three different cases of varying temperature setpoints, across the zones, are used to compare the SAT reset strategies. One or two zones are set to have an extreme setpoint while the rest of the zones have a 22℃ setpoint. Additionally, a base case with a uniform 22℃ setpoint across all zones is used a reference. The following are all the simulated cases:

- Base case with uniform 22℃ setpoint.
- One zone with 18℃ setpoint.
- One zone with 25℃ setpoint.
- Two zones with different setpoint, 18℃ and 25℃.

The setpoints are chosen based on a dataset of an office building located in Ottawa, Canada. After analyzing the temperature setpoints of all the zones, it was found that the maximum and minimum setpoints were 18℃ and 25℃. The dataset is obtained from a study by (Hobson et al., 2021). A 2℃ deadband between heating and cooling setpoints, is allowed for simulation. For example, if the zone setpoint is 22℃, the heating and cooling setpoints are 21℃ and 23℃ , respectively.

2.5 Evaluation Metrics

Different metrics are required to evaluate the energy and comfort performance of the SAT reset strategies. For energy evaluation, the energy usage of each heating or cooling element is used and normalized per floor area, as well as fan energy. To evaluate discomfort due to temperature fluctuations, the total duration when indoor temperature deviates from the heating and cooling setpoints is quantified for each zone. The sum of duration of discomfort across all zones is calculated, as shown in Equation (4). Since the duration of discomfort can vary depending on the building's operational hours and number of zones, it is necessary to compare it against a base case.

$$Duration\ of\ discomfort = \sum_{i=1}^{n\ zones} hours\ of\ temperature\ deviation \qquad (4)$$

3. RESULTS

In order to tune β, the penalty factor for energy use in the reward signal, a preanalysis was conducted. The factor for the temperature deviation in the reward signal, α, was fixed as one while β was varied from 0.1 to 1000 in a logarithmic manner as shown in Figure 3. It was found that below ten there was no significant improvement in comfort or increase in energy use, hence in the simulations the value of β was set to ten to maximize occupant comfort. However, these results show the flexibility of DQN algorithm as it can be tuned to put more or less emphasis on energy use and occupant comfort.

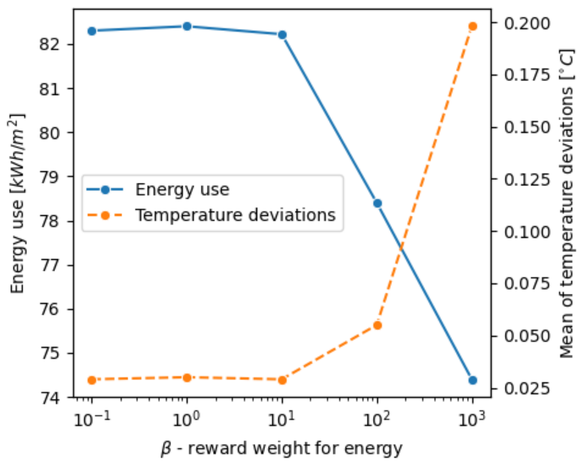

Figure 3: The effect of varying β on the energy and temperature deviation terms in the reward signal.

Figure 4 (a) to (d) show the effect of different SAT reset strategies on the energy consumption of the heating and cooling elements, and the fans, and compare between the four studied cases, which are: 1) all zones with uniform 22℃ setpoint, 2) one zone with 18℃ setpoint, 3) one zone with 25℃ setpoint, and 4) two zones with different setpoints, 18℃ and 25℃. The constant 13℃ SAT led to the highest energy use with all cases. Supplying air at 13℃ put extra load on the cooling coils to bring the mixed air temperature in the AHU to such low temperature. Moreover, ventilating with such cold air during the heating season increased the load on the perimeter heaters at the zones. It is, also, observed how the 13℃ was accompanied with the greatest increase in energy use with the cases of extreme setpoints with comparison to the base case. The OAT reset and the hybrid T&R showed similar behaviour in energy use, with slight increase with the cases of varying setpoint compared to the base case. The hybrid T&R behaved similar to the OAT reset because its maximum allowable SAT followed the same linear relation as the OAT reset. On the other hand, the pure T&R strategy provided the highest possible SAT as it depends solely on the cooling requests to reduce the SAT. Therefore, the pure T&R method was the most energy efficient.

Figure 4 (e) to (h) illustrate the duration of discomfort that resulted from the different SAT reset strategies. The results were plotted on log-scale to account for the significantly large difference in the results. The constant SAT and the OAT reset led to a significant increase in discomfort with the cases of varying setpoints compared to the base case. Since the constant SAT and the OAT reset methods lack feedback about occupant comfort, they resulted in the poorest occupant comfort. While the pure and hybrid T&R strategies have shown better performance, there was an increase in the duration of discomfort with at least one order of magnitude in the cases of varying setpoints compared to the base case. With the trim and respond methods, the first cooling request is triggered when the VAV damper exceeds 95% of their maximum position, and the second and third cooling requests are triggered when indoor temperature deviates from the setpoints with 2℃ and 3℃, respectively. This reactive nature of the T&R logic leads to continuous oscillations of indoor temperature around the cooling setpoint, hence deviating from the comfort range. The results showed that this is more prone to happen with cases of unusual temperature setpoints.

The DQN-based SAT reset achieved a better balance between energy use and comfort. Although it consumed slightly more energy than the OAT reset and both T&R methods, it achieved less duration of discomfort with a lower energy usage than the constant 13℃ SAT. The DQN algorithm also showed robustness with the cases of varying setpoints. Using the same algorithm parameters, the DQN managed to maintain the duration of discomfort with the cases of varying setpoints at the same order of magnitude as the base case. The advantage of the DQN method was that it depends on feedback from both energy and comfort. Moreover, the proactive nature of RL meant that the algorithm learnt from its previous experience, hence avoiding the repeated occurrence of unfavorable conditions, like the repeated triggering of cooling requests with T&R logic. The increased energy use with the DQN method is because it tends to keep the zone temperatures within the comfort range.

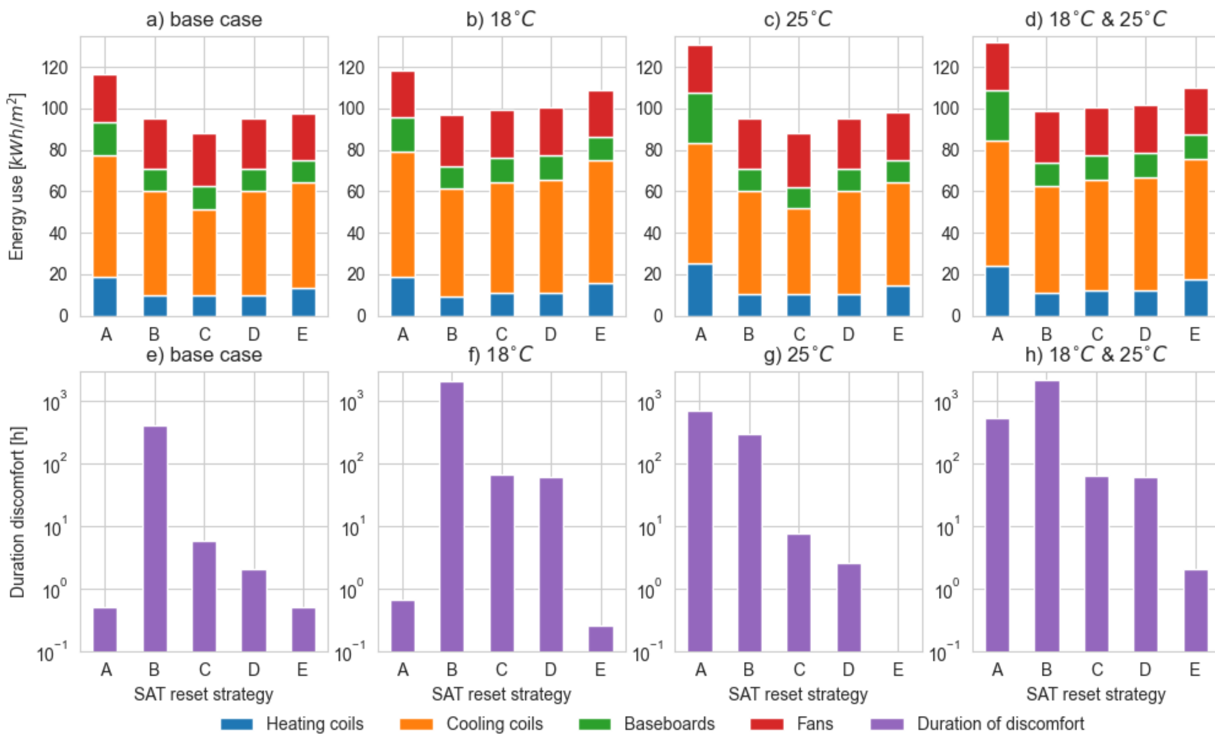

Figure 4: Subplots (a) to (d) show the annual results energy use of heating and cooling elements as well as fans, and subplots (e) to (f) show the duration of discomfort. The compared SAT reset strategies are A: 13℃ SAT, B: OAT reset, C: pure T&R, D: hybrid T&R, and E: DQN.

4. CONCLUSION

This paper studies the behaviour of four common SAT setpoint reset strategies with cases of varying setpoint across the zones. It also compared their performance with a DQN-based SAT reset strategy. The constant SAT at 13℃ led to the highest energy consumption throughout the year, while the OAT reset strategy led the highest level of discomfort. The OAT reset and both T&R methods caused higher discomfort in the cases of varying setpoint compared to the base case with uniform setpoints.

The DQN method achieved better comfort results in comparison with the pure and hybrid trim and respond, and the OAT reset methods with lower energy consumption than the constant 13℃ setpoint. The DQN method also showed more robustness with different cases of varying setpoints.

For future work, the DQN algorithm would be improved by normalizing the energy use of different components to make sure that the reward signal weighs these components equally. Furthermore, it is planned to study the DQN SAT setpoint reset method with more cases of varying and stochastic thermal preferences, as well as different occupancy patterns. This would pave the way for developing a multi-agent RL-based OCC algorithm that utilizes zone-level and system-level information to provide a robust sequence of operation that maintains occupant comfort with minimal energy requirements.

REFERENCES

ASHRAE. (2019). *Standard 62.1: Ventilation for Acceptable Indoor Air Quality*. American Society of Heating, Refrigerating, and Air-Conditioning.

ASHRAE. (2021). *Guideline 36: High-performance sequences of operation for hvac systems.* American Society of Heating, Refrigerating, and Air-Conditioning.

ASHRAE. (2022). *Standard 90.1: Energy standard for buildings except low-rise residential buildings.* American Society of Heating, Refrigerating, and Air-Conditioning.

Fang, X., Gong, G., Li, G., Chun, L., Peng, P., Li, W., Shi, X., & Chen, X. (2022). Deep reinforcement learning optimal control strategy for temperature setpoint real-time reset in multi-zone building HVAC system. *Applied Thermal Engineering*, *212*, 118552. https://doi.org/10.1016/j.applthermaleng.2022.118552

Han, M., May, R., Zhang, X., Wang, X., Pan, S., Yan, D., Jin, Y., & Xu, L. (2019). A review of reinforcement learning methodologies for controlling occupant comfort in buildings. *Sustainable Cities and Society*, *51*(July), 101748. https://doi.org/10.1016/j.scs.2019.101748

Hobson, B. W., Huchuk, B., Gunay, H. B., & O'Brien, W. (2021). A workflow for evaluating occupant-centric controls using building simulation. *Journal of Building Performance Simulation*, 1–19. https://doi.org/10.1080/19401493.2021.1894485

Jia, R., Jin, M., Sun, K., Hong, T., & Spanos, C. (2019). Advanced Building Control via Deep Reinforcement Learning. *Energy Procedia*, *158*, 6158–6163. https://doi.org/10.1016/j.egypro.2019.01.494

Kim, J., Schiavon, S., & Brager, G. (2018). Personal comfort models – A new paradigm in thermal comfort for occupant-centric environmental control. *Building and Environment*, *132*, 114–124. https://doi.org/10.1016/j.buildenv.2018.01.023

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2015). *Continuous control with deep reinforcement learning*. https://doi.org/10.48550/ARXIV.1509.02971

Lu, X., Fu, Y., & O'Neill, Z. (2023). Benchmarking high performance HVAC Rule-Based controls with advanced intelligent Controllers: A case study in a Multi-Zone system in Modelica. *Energy and Buildings*, *284*, 112854. https://doi.org/10.1016/j.enbuild.2023.112854

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). *Playing Atari with Deep Reinforcement Learning* (arXiv:1312.5602). arXiv. http://arxiv.org/abs/1312.5602

Nagy, Z., Gunay, B., Miller, C., Hahn, J., Ouf, M. M., Lee, S., Hobson, B. W., Abuimara, T., Bandurski, K., André, M., Lorenz, C.-L., Crosby, S., Dong, B., Jiang, Z., Peng, Y., Favero, M., Park, J. Y., Nweye, K., Nojedehi, P., … Vellei, M. (2023). Ten questions concerning occupant-centric control and operations. *Building and Environment*, *242*, 110518. https://doi.org/10.1016/j.buildenv.2023.110518

Park, J. Y., Ouf, M. M., Gunay, B., Peng, Y., O'Brien, W., Kjærgaard, M. B., & Nagy, Z. (2019). A critical review of field implementations of occupant-centric building controls. *Building and Environment*, *165*, 106351. https://doi.org/10.1016/j.buildenv.2019.106351

Raftery, P., Li, S., Jin, B., Ting, M., Paliaga, G., & Cheng, H. (2018). Evaluation of a cost-responsive supply air temperature reset strategy in an office building. *Energy and Buildings*, *158*, 356–370. https://doi.org/10.1016/j.enbuild.2017.10.017

Taylor, S. T. (2018). Making VAV Great Again. *ASHRAE Journal*, *60*(8).

Torabi, N., Gunay, B., O'Brien, W., & Moromisato, R. (2022). Fault-tolerant sequences of operation for VAV AHU systems through building performance simulation. *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 21–29. https://doi.org/10.1145/3563357.3564053

Wang, Z., & Hong, T. (2020). Reinforcement learning for building controls: The opportunities and challenges. *Applied Energy*, *269*(May), 115036. https://doi.org/10.1016/j.apenergy.2020.115036

ACKNOWLEDGMENTS

This research is supported by Brainbox AI Inc and the Natural Sciences and Engineering Research Council of Canada (NSERC). This work was also developed thanks to the excellent research networking provided by IEA EBC Annex 79 "Occupant-Centric Building Design and Operation".