American Journal of Smart Technology and Solutions (AJSTS)

ISSN: 2837-0295 (ONLINE)

VOLUME 5 ISSUE 2 (2026)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Smart Technology and Solutions (AJSTS)

Volume 5 Issue 2, Year 2026
ISSN: 2837-0295 (Online)
DOI: https://doi.org/10.54536/ajsts.v5i2.7966
https://journals.e-palli.com/home/index.php/ajsts

Enhancing Energy Management in Multi-Zone Buildings Using the On-Policy Reinforcement Learning Algorithm SARSA

Mohamed Abdalla Abd El-Hameed Attia[1], Mohamed Abdelaal Ahmed Abdelaal[1*], Elsayed Abd Elhameed Sallam[1]

Article Information

Received: May 02, 2026

Accepted: June 08, 2026

Published: September 14, 2026

Keywords

Buildings Energy Management, Energy Management, HVAC Control, Multi-Zone Buildings, Reinforcement Learning, Sarsa, Smart Buildings

ABSTRACT

THVAC systems in commercial buildings, particularly open-plan offices, account for substantial energy use. Thermal coupling between adjacent zones complicates control, while conventional rule-based and model-based approaches may not adapt effectively to changing occupancy and environmental conditions. Deep reinforcement learning methods can also impose high computational demands, limiting their practical deployment. This study developed and evaluated a lightweight, computationally inexpensive HVAC control policy for multi-zone open-plan offices using the on-policy reinforcement learning algorithm SARSA. A tabular SARSA controller was implemented in an EnergyPlus–Python co-simulation of a six-zone office model. The agent was trained using realistic occupancy schedules, TMY3 weather data, and 5-minute control intervals. A multi-objective reward function balanced energy efficiency, thermal comfort, and switching frequency. Performance was benchmarked against rule-based, model predictive control (MPC), and deep Q-network (DQN) controllers across several U.S. climate regions. The SARSA controller reduced annual HVAC energy use by 39.39%, from 18,200 to 11,030 kWh. Thermal comfort violations were limited to 2.85%, with an average temperature deviation of 1.2 °F and a 15-minute recovery time. Switching frequency decreased to 4.1 ON/OFF cycles per zone per day, approximately half that of rule-based control. Across different climates, energy savings ranged from 32.6% to 40.7%. These findings show that tabular SARSA is an effective and computationally efficient alternative to deep reinforcement learning for real-time HVAC control in multi-zone smart buildings.

INTRODUCTION

Energy use in commercial buildings in the United States presents a significant issue, as heating, ventilation, and air conditioning (HVAC) systems account for about 40% of the overall energy consumption (Ahmad *et al.*, 2016). This high energy consumption costs building owners a lot of money, and it also lets out a lot of greenhouse gases. This indicates that managing HVAC systems efficiently is a key approach to helping the world achieve its sustainability goals (Cao *et al.*, 2018). More companies are shifting into open-plan workplaces because they are flexible and make good use of space (Balaji *et al.*, 2013). However, these building designs make the thermal dynamics more convoluted, which makes it much difficult to keep the temperature stable. When two areas are adjacent without any physical barriers between them, significant convective heat transfer occurs. This thermally connects the areas, so that decisions made in one zone may influence the surrounding ones (Ding *et al.*, 2020) . Because of this thermal dependency, regulating HVAC shifts from being a series of separate choices for each zone to a complicated optimization problem that affects multiple zones. Since many companies have adopted hybrid or remote work models, office usage patterns have changed, resulting in lower and more irregular occupancy rates. This, combined with the effects of the COVID-19 pandemic, has further complicated HVAC control (Balaji *et al.*, 2013). This renders the control of HVAC in open-plan settings exceedingly more difficult. Hybrid or remote work models are implemented by many businesses today, which leads to significantly lower and more fluctuating occupancy than when employees work at their workstations full-time (Ghahramani *et al.*, 2015).

HVAC systems are usually programmed to work according to predictable workloads. Consequently, they have a tendency to work on pre-programmed programs even though the occupation of the building is minimal, thereby consuming a lot of energy and lowering the efficiency of the operations (Afroz *et al.*, 2018). We require HVAC control systems that are capable of adjusting to these dynamic trends of space usage and varying indoor environments (Azuatalam *et al.*, 2020). The absence of dynamic changes in the rule-based systems makes them susceptible to quick changes, undermining the energy efficiency and comfort of the occupants (Standard, 1992). The recent developments in the field of artificial intelligence, especially reinforcement learning (RL), would contribute to solving these problematic control issues (Sutton & Barto, 1998). Adaptive control policies may be trained by reinforcement learning, which trains optimal strategies by engaging with the building environment directly, without modeling the system or manually tuning the system to achieve good performance (Gao *et al.*, 2020).

The heating, ventilation, and air conditioning (HVAC) systems are some of the biggest direct consumers of energy in the built environment, with direct impact on the cost of operation, comfort of the occupants, and

[1] Computers and Control Department, Faculty of Engineering, Tanta University, Tanta, Egypt
[*] Corresponding author's e-mail: MohamedAbdelaalAhmedAbdelaal@outlook.com

emission of greenhouse gases (Asim *et al.*, 2022). The building industry and construction industry form about 34-40 % of the final energy consumption in the world, as well as an equivalent proportion of the carbon emission, which means that the building systems must be efficient. The HVAC equipment can be up to 40-50 % of the total building energy in this industry, especially in commercial buildings, where cooling and ventilating loads are extremely high due to occupancy, inside loads, and external climatic conditions (Guo *et al.*, 2019). The increase in the HVAC energy use also increases the peak electrical loads, overloading grids and compelling the augmentation in the application of sources of carbon-intensive generation, particularly in regions of large temperature swings (Ojadi *et al.*, 2024). The traditional regulation of HVAC relies, to a considerable extent, on rule-based algorithms with classical feedback regulators such as Proportional-Integral-Derivative (PID) regulators. Even though these methods are easy and highly powerful, they are not adaptable to non-steady state due to the variation of occupancy, weather, and thermal properties of buildings. This tends to result in a lack of efficient control and, therefore, energy and discomfort wastage (Pokharel). To address these inadequacies, more advanced model-based methods, including Model Predictive Control (MPC), have been thought of at the cost of high calculations, and also reliance on quality system models, which are hard to acquire in real-life building contexts (Kouvaritakis & Cannon, 2016).

Reinforcement Learning (RL) is a type of machine learning approach in which an agent learn control policies through action on the environment, and has been a promising approach to HVAC optimization. The RL can handle non-linear dynamics and high-dimensional state space, which enables it to optimize the policies to make trade-offs on the goals of energy consumption and thermal comfort (Sierla *et al.*, 2022). In contrast to supervised learning, RL devices do not require control sequence labels; they are just enhanced decision plans by optimizing them in terms of accumulated rewards. This is befitting to HVAC control concerns that are always linked with trade-offs of energy consumption, comfort, and equipment maintenance, which is multi-objective. It has been found through literature that RL demonstrated both success and failure in the construction of HVAC systems. Deep reinforcement learning models have been shown to achieve 30-40% energy savings over rule-based control in a wide variety of climate regions, and often discover comfort at a low-cost energy level. According to other RL research, the effect is a 2-14 % energy reduction and better thermal comfort and faster convergence compared to other traditional methods. The hybrid combination with expert knowledge has also achieved about 13.3 % of lowering HVAC energy usage, along with a massive decline in training time (Sivamayil *et al.*, 2023).

Although RL has been shown to introduce considerable benefits in the HVAC sector, its practical application in HVAC remains limited due to the presence of several challenges associated with the implementation of RL in this sector, including the volumes of data needed, the safety of online discovery, and compatibility issues with an existing building automation system. Furthermore, field studies show that under the consideration of deployment cost and reliability, a basic average of 13-16 % real-world energy savings can be achieved with the advanced controllers like MPC and RL (Jiang *et al.*, 2021). Overall, reinforcement learning is a more adaptive and scalable, more data-driven variant of classical HVAC control and can be employed to generate effective strategies reducing energy consumption costs and guaranteeing comfortable interiors. Both the efficiency of the algorithms and the safety limit and hybrid learning plans should be further researched to transform RL into a real-world building management system, which is non-simulated (Jiang *et al.*, 2021).

Therefore, this study aimed to suggest the SARSA reinforcement learning algorithm, which is a lightweight and robust algorithm that acts as a real-time HVAC control alternative in open-plan multi-zone offices. A well-designed reward function that can meet three critical goals, including sustaining thermal comfort, improving energy efficiency, and reducing system switching frequency, guides the learning process.

LITERATURE REVIEW

Past studies on HVAC control cover various approaches with their own contributions and limitations to the research. The use of PID controllers became common in early studies based on their simplicity and rapidity in correcting errors (Homod *et al.*, 2009). The main emphasis of these works was to tune the parameters of PID to achieve thermal comfort in predetermined conditions; however, they had little ability to adapt to changing occupancy conditions or thermal interactions. A cascade PID controller that regulates the temperature of HVAC in a multi-zone architectural setting was used. The PID-based predictive controllers were developed in an attempt to optimise thermal comfort in commercial buildings. The methodology involved the integration of the conventional PID control with forecasting. Building Control Management (BCM) systems are a development of simple PID controllers that include logic based on rules and multiple sensor inputs to manage different subsystems in the building. Such systems have a variety of centralized monitoring features, unification of various building subsystems, rudimentary scheduling features, and a few adaptive features (Freire *et al.*, 2008).

A mid-sized commercial building was installed with a comprehensive BCM system that shows enhanced coordination of HVAC subsystems in operation as opposed to standalone controllers and achieves 12 % energy savings by coordinated operation and optimized equipment sequencing (Bengea *et al.*, 2014). A combined strategy of HVAC management and energy storage optimization of large-scale buildings was made. They also used thermal mass and a battery energy storage system in their methodology, with 20 % energy reduction using a coordinated control approach that used both thermal and electrical storage. With the introduction of Reinforcement Learning, advanced HVAC control methods are available that can acquire complex policies via interaction with the environment (Bianchini *et al.*, 2019).

These approaches use different RL algorithms to estimate the optimal control policies to allow them to solve control problems of high complexity and respond to changing conditions. Deep Q-Networks (DQN) were applied to multi-zone building control, and it is shown that they have a significant improvement over conventional control strategies. The model employed a three-layer deep neural network to approximate Q-values for discrete HVAC actions across multiple zones, achieving 25% energy savings relative to rule-of-thumb systems and maintaining thermal comfort within ASHRAE guidelines (Wei *et al.*, 2017). A multi-agent deep reinforcement learning framework to address the scalability problem of single-agent methods in residential HVAC control was introduced. This approach utilized several RL agents, each of which controlled separate zones and communicated with each other via a standard communication protocol. The approach achieved adaptive control policies that respond to varying occupancy patterns, resulting in a 30% energy reduction compared to thermostat-based control, and demonstrated superior performance during demand response events. More recently, a limited number of studies have investigated on-policy RL methods, such as SARSA, for HVAC control (Li & Du, 2023). These works demonstrated the potential of on-policy algorithms to provide more stable learning by updating values according to the current policy, producing smoother and more predictable control signals. Nonetheless, existing literature predominantly focused on theoretical or small-scale experiments, frequently neglecting the complexities of open-plan multi-zone buildings with thermal coupling and rarely addressing computational feasibility for deployment in typical BEMS platforms (Li & Du, 2023).

A comparative analysis of various reinforcement learning approaches for HVAC optimization was presented, demonstrating the potential of RL to enhance both energy efficiency and occupant comfort. However, this study simplified the building model by aggregating it into a single thermal zone, thereby overlooking the complex thermal interactions typical in open-plan multi-zone office environments. Consequently, its results cannot be applied to natural multi-zone cases where inter-zone heat transfer is an important factor in control performance (Niemann & Schmitz, 2020).

MATERIALS AND METHODS

SARSA Algorithm Fundamentals

SARSA is an on-policy temporal difference learning algorithm that is a part of the model-free reinforcement learning methods (Sutton & Barto, 1998). SARSA was named after the quintuple of information in the update rule S, A, R, S, and A. In contrast to off-policy algorithms like Q-learning, SARSA will update the action-value function depending on what action the current policy is actually performing, which is more conservative and practical (Sutton & Barto, 1998). SARSA was developed initially as a generalization of the temporal difference learning algorithms and has been well-known to have the property of stability and convergence in stochastic settings. This algorithm is especially appropriate in systems where predictability and safety are significant factors, including building control systems where unpredictable reactions might cause discomfort to the occupants or damage to equipment (Watkins & Dayan, 1992).

The SARSA algorithm is mainly based on the value update mechanism, which is based on the principle of the temporal difference of learning. The algorithm has a function of action value Q(s, a) which approximates the cumulative reward expected on doing action a in state s and proceeding with the current policy thereafter (Fu *et al.*, 2018). The update rule of SARSA is stated as:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)] \qquad (1)$$

where Q(st, at) is the current approximation of the action-value function at state s and action a, 0 is the learning rate parameter that controls the rate of learning and adaptation, r is the immediate reward obtained after acting a in state s, 0 is the discount factor that shows the importance of future rewards, st+1 is the next state obtained after acting a, and at+1 is the following action in state st+1 chosen by the policy (Fu *et al.*, 2018).

The term [r + γQ(st+1, at+1) - Q(st, at)] represents the temporal difference error, which measures the difference between the current estimate and the improved estimate based on the observed reward and next state-action value. This error drives the learning process by adjusting the Q-values toward more accurate estimates (Nguyen, 2024). Under appropriate conditions (sufficient exploration and decreasing learning rate), SARSA converges to the optimal action-value function Q*(s, a) with probability 1. Convergence is guaranteed under the following conditions: all state-action pairs are visited infinitely often, the learning rate α satisfies the Robbins-Monro conditions (i.e., $\Sigma\alpha = \infty$ and $\Sigma\alpha^2 < \infty$(, and the policy asymptotically becomes greedy with respect to the optimal action-value function.

SARSA-Based HVAC Control Framework

This section introduces the modeling and algorithmic framework of the SARSA-Based HVAC Control Framework.

Algorithmic Modeling Framework

Proper hyperparameter tuning is essential to ensure the stability and efficiency of the SARSA-based HVAC control framework. In this work, several key parameters were selected based on empirical evaluation to optimize learning performance and convergence behavior. These parameters are summarized as follows:

1. **Learning Rate** (α): The learning rate controls how quickly the algorithm adapts to new information. In practical HVAC control applications, the learning rate (α) is typically selected in the range of 0.1–0.3, with 0.1 selected for stable learning in building control applications. An adaptive schedule may further reduce α over time to ensure convergence, balancing faster adaptation against the risk of oscillations.

2. **Discount Factor** (γ): The discount factor governs the relative weighting of future versus immediate rewards, typically selected in the range 0.9–0.99. A value of 0.95 is selected for HVAC applications, effectively valuing rewards 20 steps ahead at approximately 36% of immediate returns. Higher γ promotes long-term energy savings, whereas lower values prioritize immediate occupant comfort.

3. **Exploration Rate** (ε): A trade-off between exploration and exploitation is ensured by employing an ε-greedy strategy, where a random action is selected with probability ϵ, and the action a=argmax Q(s, a) is selected with probability (1− ϵ). The exploration rate (ε) in ε-greedy action selection balances exploration and exploitation, typically ranging from 0.05 to 0.3. For HVAC applications, ε is often initialized at 1.0 and decayed linearly to 0.05 over training episodes, with a selected operating range around 0.1 to ensure sufficient exploration before converging to policy exploitation.

4. **Policy Update Frequency:** The policy update frequency is structured to reflect building operational cycles, with episodes typically spanning one day (1440 minutes) or one week. Control actions are executed at 5-minute intervals to maintain responsiveness, with Q-values updated at each time step and the policy evaluated at the end of each episode.

Algorithmic Steps for SARSA-Based HVAC Control

This section presents the SARSA-based control algorithm developed for the HVAC system. The proposed framework models each state st encapsulates zone temperatures, setpoints, occupancy indicators, and external ambient conditions. The control action at adjusts system inputs such as supply air flow rates or zone setpoint temperatures. At each episode (typically spanning a day or a week), the algorithm proceeds as follows:

1. Initialize: Set the action-value function Q(s, a) (or its approximator parameters) to zero.

2. Initial State and Action: Obtain the initial state s0 from the building environment and select action a0 using an ε-greedy policy.

3. Time Step Iteration: For each 5-minute interval within the episode:

• Execute Action: Apply at to the HVAC system or simulation, which updates zone temperatures and energy consumption.

• Observe Outcome: Record the new state st+1 and compute the immediate reward.

$$R_t = -\eta_T \sum \left(T_i - T_{target}\right)^2 - \eta_E E_t \quad (2)$$

where Ti is the temperature in zone i, Ttarget is the desired target temperature, Et is the energy consumption at time step t, ηT and ηE are weighting coefficients that balance the trade-off between thermal comfort and energy efficiency.

• Next Action Selection: Choose at+1 from st+1 via ε-greedy policy.

• Q-Value Update: Update the action-value estimate according to equation (1).

• Set Update state and action:
$s_t \leftarrow s_{t+1}$, $a_t \leftarrow a_{t+1}$

4. Policy Evaluation: At the end of each episode, evaluate cumulative energy savings and comfort metrics to monitor learning progression.

Simulation Environment and Implementation

The dataset used for training the proposed SARSA-based HVAC control framework was derived from the EnergyPlus Weather Data Library. Specifically, weather files in the TMY3 (Typical Meteorological Year) format were imported for multiple U.S. cities to ensure a realistic representation of local climate conditions and to assess the model's generalizability across diverse environments. Simulations were conducted at a fine-grained 5-minute timestep over an entire calendar year, allowing the agent to experience and adapt to dynamic daily weather fluctuations (DOE, 2010).

To implement the proposed control strategy effectively, several core aspects of the simulation environment and algorithm design were addressed:

1. **State Discretization**: For continuous state variables such as temperature and humidity, discretization is typically employed to reduce the state space dimensionality and facilitate tabular representation. The discretization scheme adopted in this study is as follows:

• Temperature: 0.5°C bins for indoor temperature (e.g., 20.0°C, 20.5°C, 21.0°C)

• Outdoor Conditions: 2°C bins for outdoor temperature

• Occupancy: Binary states (occupied/unoccupied) or categorical (low/medium/high)

2. **Action Space Design**: To simplify the control problem and enable efficient tabular learning with SARSA, actions are typically discretized. This reduces computational complexity and facilitates practical implementation in HVAC systems. The discrete action space employed here includes:

• Binary Control: ON/OFF for each HVAC zone

• Multi-level Control: Low/Medium/High operation modes

• Combined Actions: Simultaneous control of multiple zones

3. The tabular SARSA approach requires explicit storage of Q-values for each state-action pair, resulting in memory usage that scales linearly with the size of the state and action spaces, i.e., (|S| × |A|), where |S| denotes the number of discrete states and |A| the number of discrete actions. The memory implications of

this approach are summarized as follows:

• Example: 6-zone building with discretized states might require ~10^4 to 10^6 Q-table entries

• Storage: Each entry requires 4-8 bytes (float), total memory typically <10 MB

• Scalability: Memory requirements grow linearly with problem size, unlike neural network approaches

4. **Convergence Properties:** The learning dynamics of the SARSA-based HVAC controller were examined to characterize how quickly and reliably the algorithm reaches a stable policy. Key observations regarding convergence are summarized as follows:

• Training Episodes: Convergence typically occurs within 50–200 training episodes for building-scale environments.

• Time to Convergence: This corresponds to roughly 20–100 hours of simulated operation, depending on building complexity and discretization granularity.

• Stability: After convergence, the learned policy demonstrates consistent performance with minimal drift, ensuring predictable long-term behavior

5. **Computational Efficiency**: The tabular SARSA approach demonstrates strong computational efficiency, making it particularly suitable for real-time building control applications where rapid decision-making is essential. Its simplicity allows for fast updates and low computational overhead, as summarized below:

• Update Time: O(1) per time step for Q-value updates

• Memory Access: Direct table lookup, no matrix operations

• Hardware Requirements: Standard CPU sufficient, no GPU needed

• Real-time Capability: Suitable for control loops with 1-minute or faster response times

RESULTS AND DISCUSSION

Over 20 simulation episodes, each comprising more than 100,000 timesteps at a 5-second resolution, the SARSA-based HVAC controller was trained and evaluated using TMY3-based weather data to assess policy convergence, learning stability, and progressive performance improvements.

Energy Consumption Analysis

Table 1 presents the comparative analysis of the considered HVAC control strategies regarding the consumption of energy per year, relative energy savings, training period, and the computation requirements, and the respective trends of the energy use are represented in Figure 1.

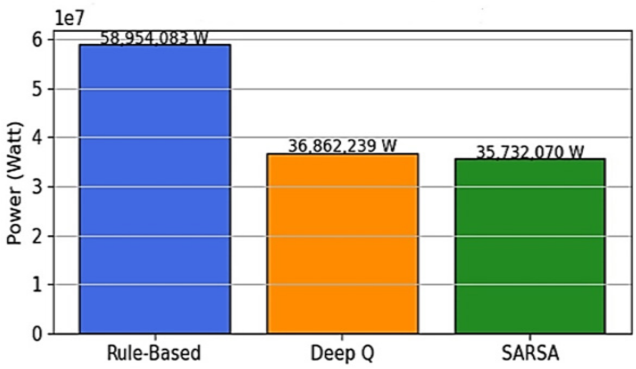

Figure 1: Energy Consumption (Watts)

The findings show that the proposed SARSA-based controller has the lowest energy consumption (11,030 kWh) and energy savings of 39.39% against the 18,200 kWh rule-based baseline. The MPC method also provides a high degree of savings (31.6) but incurs more computation and more training (~80 min). Despite the DQN controller realising competitive energy savings (37.5%), the long training duration (around 90120 min) and the reliance of the controller on the hardware of a graphical processing unit (GPU) can limit its scope in traditional building management infrastructure. Conversely, the SARSA controller is much faster to converge (in the order of 40 min), and it works well with standard CPU resources, as would be expected in a system with an Intel Core i9 platform and 64 GB memory but no graphics card. These results highlight the feasibility benefits of tabular SARSA in providing high-energy efficiency and low-computational complexities to construct building-scale HVAC control.

Table 1: Comparison of Energy Optimization Using SARSA Versus Alternatives

Controller Type	Annual Energy Use (kWh)	Energy Saving (%)	Training Time	Hardware Requirements
Rule-Based	18,200	--	--	Standard
MPC	12,450	31.6%	~80 min	Medium CPU
DQN	11,370	37.5%	~90-120 min	High (GPU)
SARSA	11,030	39.39%	~40 min	Low (CPU)

Thermal Comfort Analysis

Table 2 provides a comparison of occupant comfort performance between the tested HVAC control strategies based on the comfort violation rate, mean temperature difference, and comfort recovery period. The SARSA controller significantly enhances comfort over the rule-based approach, with a comfort violation rate of 2.85% versus 7.9% and average temperature variations of ±1.2 °F. Whereas the DQN controller has the lowest violation rate (0.64) and minimal deviation in temperature (±0.8 °C), the benefits are obtained at the cost of increased computational complexity and longer training. The MPC controller has a moderate performance level in terms of comfort, with a violation rate of 4.1 % and deviation of ±1.5°F. SARSA also recovers faster when subjected to thermal disturbances, returning the comfort conditions within 15 minutes (similar to rule-based control and comparable to MPC), and it is faster than MPC (18 minutes)

Table 2: Key occupant comfort metrics for the evaluated controllers

Controller	Comfort Violation Rate (%)	Avg. Temperature Deviation (°F)	Comfort Recovery Time (min)
Rule-Based	7.9%	±2.1	25
MPC	4.1%	±1.5	18
DQN	0.64%	±0.8	12
SARSA	2.85%	±1.2	15

Figure 2 shows that SARSA training experiences an initial exploration stage until the convergence stage, which marks confident learning behavior. In general, SARSA has a positive ratio of comfort performance, energy, and computational practicability.

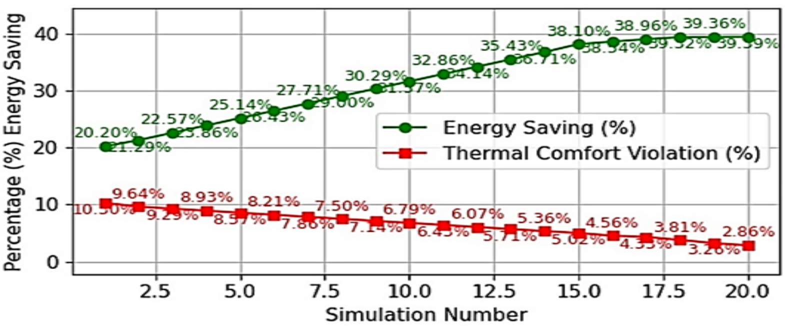

Figure 2: Energy Saving VS Thermal Comfort Violation

Zone Comfort and Stability Analysis

The performance of the SARSA-based controller in terms of spatial comfort in individual zones of the building is presented in Table 3. The Comfort Violation Rates are also higher in the perimeter areas, predominantly south and west-facing areas (4.1 and 3.8, respectively) due to exposure to solar radiation and external thermal loads. Interior and core zones, on the contrary, have significantly lower violation rates (1.2 to 1.8 %), with much higher thermal stability because of this. All zones are well-controlled close to the average temperatures of 72 °C, which is a sign of successful setpoint monitoring. The Stability Index also proves the better thermal stability of interior areas (0.96–0.98), and somewhat lower thermal stability of the perimeter areas under the impact of dynamic outdoor factors.

Table 3: zone-level comfort metrics under the SARSA-based control policy

Zone Location	CVR (%)	Avg. Temp (°F)	Stability Index*
Zone 1 (North)	3.2%	72.1	0.92
Zone 2 (South)	4.1%	72.8	0.88
Zone 3 (East)	2.1%	72.3	0.95
Zone 4 (West)	3.8%	72.6	0.89
Zone 5 (Center)	1.2%	72.2	0.98
Zone 6 (Core)	1.8%	72.0	0.96

Mechanical Switching Stability Analysis

Table 4 is a comparative table of controllers in terms of switching frequency, equipment wear, and maintenance cost. The rule-based controller was the fastest in terms of the switching rate (8.2 events/day per zone), which was more than what the manufacturers recommended,

Table 4: The controllers in terms of average daily switching frequency

Controller	Daily ON/OFF Switches (Avg per Zone)	Equipment Wear Index**	Maintenance Cost Factor
Rule-Based	8.2	1.00	1.00
MPC	6.0	0.73	0.81
DQN	6.7	0.82	0.87
SARSA	4.1	0.50	0.64

and this created the highest wear index (1.00) and maintenance factor (1.00). MPC and DQN also decreased cycling with wear indices of 0.73 and 0.82. The SARSA-based controller had the fewest switching frequency (4.1 events/day), reduced the wear index (0.50), and cost maintenance (0.64). This fact suggests that SARSA is the best compromise between operational efficiency and component lifecycle, and reduces equipment cycling and possible long-term maintenance needs.

Performance Analysis of SARSA-Based Controller in Multiple Climate Zones

Table 5 presents the results of the SARSA-based controller in 6 cities of the U.S. with different climates. The energy savings were up to 32.6 % in Los Angeles and 40.7 % in Phoenix, and this represents the predictability of the thermal loads to the controller. The Comfort Violation Rates were all less than 5% meaning that occupant comfort was reliable. The number of switches per zone day was kept within acceptable ranges (3.946), and this minimized wear. SARSA was effective for managing both sensible and latent loads and did well in a humid climate, such as Miami, and a dry climate, such as Phoenix. Generally, the findings indicate that SARSA is a robust product in terms of maximization of energy efficiency and comfort and minimization of equipment wear in different climatic conditions.

Table 5: The performance of the proposed SARSA-based controller across six representative U.S. cities

City	Climate Zone	Energy Saving (%)	CVR (%)	Avg. Switches/Day	Annual kWh
Greenville, SC	Mixed Humid	39.39	2.85	4.1	11,030
New York, NY	Cold	37.1	3.2	4.3	11,445
Los Angeles, CA	Mild	32.6	2.6	3.9	12,264
Miami, FL	Hot Humid	34.7	3.7	4.5	11,884
Chicago, IL	Cold	38.2	3.4	4.4	11,236
Phoenix, AZ	Hot Dry	40.7	4.8	4.6	10,783

Discussion

The study results suggest that an on-policy tabular SARSA controller can achieve substantial energy consumption savings of HVAC and predictable thermal comfort in a complex, multi-zone office environment (Tariq *et al.*, 2025). It may be related to the accumulating body of knowledge that reinforcement learning (RL) methods may be efficient in providing adaptive control strategies. However, the degree of execution and the character of the performance are contingent on the design and type of building, weather conditions, and the place of application (Tariq *et al.*, 2025). Several simulation studies have compared different RL techniques to the traditional baseline controllers and have reported varied results in saving energy and comfort depending on the complexity of the model and objectives. Indicatively, even simple architectures (Deep Q-Network (DQN) and Deep Deterministic Policy Gradient (DDPG)) can be used by RL agents to capture subtle tradeoffs between operating cost and comfort in residential HVAC simulations, with state-of-the-art RL controllers achieving 512-12 and 1530-30 of energy cost savings and comfort improvements. Similarly, DDPG-directed approaches are more effective than DQN and lessen comfort breakage by as much as 98 % in simulation, with better energy cost savings, owing to the significance of more effective function representation to handle uncertainty and nonlinear dynamics (Lazaridis *et al.*, 2024).

Other studies have made comparisons of Proximal Policy Optimization (PPO) and Advantage Actor-Critic (A2C); they have discovered that the choice of algorithm may significantly influence performance. A2C in cooling mode in different climates consumed a smaller amount of thermal energy than conventional rule-based methods, with the objective of maintaining thermal comfort by 4-22 % but these savings varied according to the environmental condition and network structure. These findings suggest that it is a significant design variable, the choice of algorithms and the reward structure, that determines the success of RL agents in the trade-off of competing HVAC control objectives (Purwanto & Kang, 2024). In addition, according to a literature review on the numerous commercial and residential building applications, RL-based structures could reduce the total energy consumption and peak demand loads by an average of -27.3%, and by half entirely, compared to the typical building management systems, and could achieve agreeable comfort levels of nearly 97 % of the total building occupied time. This form of aggregated field data points to the potential usefulness of RL structures not only to the situation when individual simulations are used, but to other situations where agents change in response to weather and occupancy changes (Michailidis *et al.*, 2025).

Interestingly, the trade-offs are practicalized in comparative field deployment implementation employing Model Predictive Control (MPC). In one of the residential deployments, RL had 22 % of the energy conservation when compared to the incumbent controller, a little bigger than the 20 % of MPC, but with some instances of greater occupant discomfort. According to energy saving and comfort performance, MPC tended to have high scores in the normalized space, which emphasizes the concept that robustness of control and the performance of comfort are still issues that are alive in RL. Despite these positive results, some studies indicate that the provided changes using RL are also susceptible to the gap between the simulation and the actual performance, as well as the training paradigm (Afram & Janabi-Sharifi, 2014). Online learning and imitation learning RL has shown dramatic simulated performance improvements - e.g., 65 % energy savings, 25 % comfort improvement with a particular form of model - but these factors are commonly linked to ideal white-box models and short test times, which may not always be representative of the actual conditions of buildings. Moreover, the black-box reinforcement learning techniques can hardly be interpreted and checked in those systems when the safety and health of people take priority (Zare *et al.*, 2024). Recent initiatives have suggested interpolable and quantifiable techniques of getting RL policy, and propositions of deterministic decision tree-like techniques can ensure a performance decrease without radically raising the computational cost - in one case, a 68.4 % reduction in energy consumption and 14.8 % increase in comfort over other policy-extraction techniques based on model-guided RL.

Another key insight of the literature considered is that optimizing multi-objective HVAC is often not confined to the energy efficiency and comfort, but it is also dynamic to weather variability and operational cost. The 8.3-minute average time of adaptation in some of the most sophisticated systems indicates that the RL agents can adjust to the change in the environment quickly to make the accuracy of the load prediction over 94% and save costs by 24%, which is estimated. These dynamic adaptation attributes complement the conventional performance measures that tend to be centralized within a simulation work and support the importance of performance dimensions like response time and adaptability (Wei *et al.*, 2015). Overall, these results justify the usefulness of RL-based HVAC control, as well as revealing that quantitative performance is strictly reliant on the structure of algorithms, the accuracy of the model that is used to represent the environment, climatic environment, and implementation strategy. The SARSA controller of the current work is near the border of practical simplicity and competitive performance, as well as demonstrating that properly tuned, computationally efficient RL architectures can produce an outcome that is similar or even better than larger RL architectures, at least in a structured multi-zone office setting. The field validations, hybrid modeling and techniques, transfer learning, and safety-conscious control strategy will need further research in order to bridge the gap between the promising simulations and the scalable real-world implementation.

CONCLUSION

This study showed that a SARSA-based reinforcement learning controller could effectively be used to manage HVAC energy usage in open-plan, multi-zone office buildings that are highly thermally coupled and undergo changing occupancy patterns. The proposed controller was implemented in an Energy Plus simulation setting and was found to save up to 39.39 % of energy relative to traditional rule-based techniques, and at the same time, it kept thermal comfort violations below 3 % and mechanical switching frequency at significantly lower levels. In contrast to deep reinforcement learning algorithms, the tabular SARSA algorithm had a fast convergence, minimal computational requirements, and stable control dynamics, and thus was suitable to be deployed in real time on typical building management hardware. It was also robust and adaptable, as the controller also exhibited a high level of generalization with different climate zones without the need to undergo retraining. Generally, all the results indicate that on-policy learning in reinforcement can be a viable and scalable alternative to sophisticated model-based and deep learning techniques used in the smart building context. Future directions include field validation, extensions to continuous-state, and additional integration with renewable energy sources and demand response programs to improve building-level energy intelligence further.

REFERENCES

Afram, A., & Janabi-Sharifi, F. (2014). Theory and applications of HVAC control systems–a review of model predictive control (MPC). *Building and Environment, 72,* 343-355.

Afroz, Z., Shafiullah, G., Urmee, T., & Higgins, G. (2018). Modeling techniques used in building HVAC control systems: a review. *Renewable and sustainable energy reviews, 83*, 64-84.

Ahmad, M. W., Mourshed, M., Yuce, B., & Rezgui, Y. (2016). *Computational intelligence techniques for HVAC systems: a review.* Building Simulation,

Asim, N., Badiei, M., Mohammad, M., Razali, H., Rajabi, A., Chin Haw, L., & Jameelah Ghazali, M. (2022). Sustainability of heating, ventilation and air-conditioning (HVAC) systems in buildings—an overview. *International journal of environmental research* *and public health, 19*(2), 1016.

Azuatalam, D., Lee, W.-L., De Nijs, F., & Liebman, A. (2020). Reinforcement learning for whole-building HVAC control and demand response. *Energy and AI, 2,* 100020.

Balaji, B., Xu, J., Nwokafor, A., Gupta, R., & Agarwal, Y. (2013). Sentinel: occupancy based HVAC actuation using existing WiFi infrastructure within commercial buildings. *Proceedings of the 11th ACM conference on embedded networked sensor systems,*

Bengea, S. C., Kelman, A. D., Borrelli, F., Taylor, R., & Narayanan, S. (2014). Implementation of model predictive control for an HVAC system in a mid-size commercial building. *HVAC&R Research, 20*(1), 121-135.

Bianchini, G., Casini, M., Pepe, D., Vicino, A., & Zanvettor, G. G. (2019). An integrated model predictive control approach for optimal HVAC and energy storage operation in large-scale buildings. *Applied Energy, 240,* 327-340.

Cao, Y., Wei, W., Wang, J., Mei, S., Shafie-khah, M., & Catalao, J. P. (2018). Capacity planning of energy hub in multi-carrier energy networks: a data-driven robust stochastic programming approach. *IEEE Transactions on Sustainable Energy, 11*(1), 3-14.

Ding, X., Du, W., & Cerpa, A. E. (2020). Mb2c: Model-based deep reinforcement learning for multi-zone building control. *Proceedings of the 7th ACM international conference on systems for energy-efficient buildings, cities, and transportation*

DOE, U. (2010). EnergyPlus Documentation. *US Department of Energy.*

Freire, R. Z., Oliveira, G. H., & Mendes, N. (2008). Predictive controllers for thermal comfort optimization and energy savings. *Energy and Buildings, 40*(7), 1353-1365.

Fu, Q., Hu, L., Wu, H., Hu, F., Hu, W., & Chen, J. (2018). A Sarsa-based adaptive controller for building energy conservation. *Journal of Computational Methods in Science and Engineering, 18*(2), 329-338.

Gao, G., Li, J., & Wen, Y. (2020). DeepComfort: energy efficient thermal comfort control in buildings via reinforcement learning. *IEEE Internet of Things Journal, 7*(9), 8472-8484.

Ghahramani, A., Tang, C., & Becerik-Gerber, B. (2015). An online learning approach for quantifying personalized thermal comfort via adaptive stochastic modeling. *Building and Environment, 92*, 86-96.

Guo, S., Zheng, S., Hu, Y., Hong, J., Wu, X., & Tang, M. (2019). Embodied energy use in the global construction industry. *Applied Energy, 256*, 113838.

Homod, R. Z., Mahlia, T., & Mohamed, H. A. (2009). Pid-cascade for hvac system control. *International Conference on Control, Instrumentation and Mechatronic Engineering (CIM09)*

Jiang, Z., Risbeck, M. J., Ramamurti, V., Murugesan, S., Amores, J., Zhang, C., Lee, Y. M., & Drees, K. H. (2021). Building HVAC control with reinforcement learning for reduction of energy cost and demand charge. *Energy and Buildings, 239*, 110833.

Kouvaritakis, B., & Cannon, M. (2016). Model predictive control. *Switzerland: Springer International Publishing, 38*(13-56), 7.

Lazaridis, C. R., Michailidis, I., Karatzinis, G., Michailidis, P., & Kosmatopoulos, E. (2024). Evaluating reinforcement learning algorithms in residential energy saving and comfort management. *Energies, 17*(3), 581.

Li, F., & Du, Y. (2023). Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning. In Deep Learning for Power System Applications: *Case Studies Linking Artificial Intelligence and Power Systems* (pp. 71-96). Springer.

Michailidis, P., Michailidis, I., & Kosmatopoulos, E. (2025). Reinforcement learning for optimizing renewable energy utilization in buildings: a review on applications and innovations. *Energies, 18*(7), 1724.

Nguyen, D. H. (2024). *The role of reinforcement learning control for optimizing building energy management systems.*

Niemann, P., & Schmitz, G. (2020). Impacts of occupancy on energy demand and thermal comfort for a large-sized administration building. *Building and Environment, 182*, 107027.

Ojadi, J. O., Odionu, C. S., Onukwulu, E. C., & Owulade, O. A. (2024). AI-enabled smart grid systems for energy efficiency and carbon footprint reduction in urban energy networks. *International Journal of Multidisciplinary Research and Growth Evaluation, 5*(1), 1549-1566.

Pokharel, B. (2025). A comprehensive study of modern control strategies: From classical PID to optimal, robust, adaptive and intelligent methods. *International Research Journal of Modernization in Engineering Technology and Science, 7*(9), 3035–3049. https://doi.org/10.56726/IRJMETS83160

Purwanto, Y. K., & Kang, D.-K. (2024). Multi-agent deep reinforcement learning for fighting game: a comparative study of PPO and A2C. *International Journal of Internet, Broadcasting and Communication, 16*(3), 192-198.

Sierla, S., Ihasalo, H., & Vyatkin, V. (2022). A review of reinforcement learning applications to control of heating, ventilation and air conditioning systems. *Energies, 15*(10), 3526.

Sivamayil, K., Rajasekar, E., Aljafari, B., Nikolovski, S., Vairavasundaram, S., & Vairavasundaram, I. (2023). A systematic study on reinforcement learning based applications. *Energies, 16*(3), 1512.

Standard, A. (1992). Thermal environmental conditions for human occupancy. *ANSI/ASHRAE, 55, 5.*

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1). MIT press Cambridge.

Tariq, S., Ali, U., Kim, S., & Yoo, C. (2025). Multi-agent distributed reinforcement learning for energy-efficient thermal comfort control in multi-zone buildings with diverse occupancy patterns. *Energy*, 137082.

Watkins, C. J., & Dayan, P. (1992). *Q-learning*. *Machine learning, 8(*3), 279-292.

Wei, T., Wang, Y., & Zhu, Q. (2017). Deep reinforcement learning for hvac control in smart buildings. *Design Automation Conference (DAC)*

Wei, X., Kusiak, A., Li, M., Tang, F., & Zeng, Y. (2015). Multi-objective optimization of the HVAC (heating, ventilation, and air conditioning) system performance. *Energy, 83,* 294-306.

Zare, M., Kebria, P. M., Khosravi, A., & Nahavandi, S. (2024). A survey of imitation learning: algorithms recent developments, and challenges. *IEEE Transactions on Cybernetics.*