Data-Centric Engineering (2025), 6: e3
doi:10.1017/dce.2024.57

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

An environment-adaptive SAC-based HVAC control of single-zone residential and office buildings

Xinlin Wang, Nariman Mahdavi, Subbu Sethuvenkatraman and Sam West

Energy BU, CSIRO, Newcastle, 2304, NSW, Australia
Corresponding author: Xinlin Wang; Email: xinlin.wang@csiro.au

Received: 06 June 2024; **Revised:** 15 October 2024; **Accepted:** 14 November 2024

Keywords: BOPTEST; HVAC; Reinforcement learning; Smart building

Abstract

This study introduces an advanced reinforcement learning (RL)-based control strategy for heating, ventilation, and air conditioning (HVAC) systems, employing a soft actor-critic agent with a customized reward mechanism. This strategy integrates time-varying outdoor temperature-dependent weighting factors to dynamically balance thermal comfort and energy efficiency. Our methodology has undergone rigorous evaluation across two distinct test cases within the building optimization testing (BOPTEST) framework, an open-source virtual simulator equipped with standardized key performance indicators (KPIs) for performance assessment. Each test case is strategically selected to represent distinct building typologies, climatic conditions, and HVAC system complexities, ensuring a thorough evaluation of our method across diverse settings. The first test case is a heating-focused scenario in a residential setting. Here, we directly compare our method against four advanced control strategies: an optimized rule-based controller inherently provided by BOPTEST, two sophisticated RL-based strategies leveraging BOPTEST's KPIs as reward references, and a model predictive control (MPC)-based approach specifically tailored for the test case. Our results indicate that our approach outperforms the rule-based and other RL-based strategies and achieves outcomes comparable to the MPC-based controller. The second scenario, a cooling-dominated environment in an office setting, further validates the versatility of our strategy under varying conditions. The consistent performance of our strategy across both scenarios underscores its potential as a robust tool for smart building management, adaptable to both residential and office environments under different climatic challenges.

Impact Statement

Worldwide, heating, ventilation, and air conditioning (HVAC) systems in buildings account for substantial energy consumption and emissions. They also often contribute to the peak load in buildings causing stress on electricity infrastructure. To meet the demands of HVAC systems while optimizing energy use and efficiency, advanced control strategies are essential. However, traditional control methods, such as rule-based and model-based approaches, often face challenges like extensive model development, slowing their adoption in the industry. In this context, reinforcement learning (RL) has emerged as a promising, model-free solution. Despite its potential, limited research has specifically tailored RL for HVAC control, taking into account the unique characteristics and requirements of these systems while demonstrating its practical application across diverse scenarios. To address these gaps, we have developed an environment-adaptive, single-agent RL control method, showcasing its effectiveness across different climates and building types. This work offers a valuable contribution to the growing body of literature on RL-based control methods for HVAC systems.

This research article was awarded an Open Materials badge for transparent practices. See the Data Availability Statement for details.

1. Introduction

In today's efforts to mitigate global warming, reducing the energy consumption of buildings is crucial, as they are significant contributors to both energy usage and carbon emissions (Wang & Ahn, 2020; Wang et al., 2023). Previous studies have indicated that buildings account for over 30% of total energy consumption and ~33% of global greenhouse gas emissions (Moghaddasi et al., 2021; Wang, Yao, & Papaefthymiou, 2023). A critical focus within this sector is on heating, ventilation, and air conditioning (HVAC) systems, which are typically responsible for around half of a building's energy usage (Jiang et al., 2021; Wang et al., 2023). Moreover, with the intensifying effects of global warming, the demand for energy in HVAC systems is expected to rise further (Wang et al., 2023). Consequently, enhancing the energy efficiency of HVAC systems, while maintaining user comfort, is essential for fostering sustainable development.

In the building industry, conventional control strategies, such as rule-based and model-based algorithms, play a vital role in maintaining thermal comfort and optimizing HVAC operations (Afram & Janabi-Sharifi, 2014; Jiang et al., 2021; Sierla, Ihasalo, & Vyatkin, 2022; Taheri, Hosseini, & Razban, 2022; Wang et al., 2023). These strategies not only ensure the efficient functioning of systems but also contribute to the flexibility of building operations. Rule-based strategies are particularly widely implemented due to their robustness and straightforward operational mechanisms (Jiang et al., 2021; Sierla et al., 2022). Nevertheless, these strategies often fall short when it comes to adapting to the dynamic and unpredictable aspects of modern building operations, which include fluctuating occupancy levels and variable weather conditions (Jiang et al., 2021; Taheri et al., 2022; Yuan et al., 2021). On the other hand, model predictive control (MPC) approaches are becoming increasingly prominent in HVAC control, empowered by advancements in computational capabilities and the growing availability of extensive real-time building data (Taheri et al., 2022; Wang et al., 2023). Previous research has underscored that MPC provides numerous advantages over traditional HVAC control methods, including enhanced efficiency, increased precision in system management, proven robustness, and reductions in both energy usage and operational costs (Afram & Janabi-Sharifi, 2014; Wang et al., 2023). However, the deployment of MPC is not without its challenges. First, implementing an MPC controller requires model development, which is a time-intensive process demanding specialized expertise (Afram & Janabi-Sharifi, 2014). Significant hurdles include the need for a deep understanding of complex system dynamics and a reliance on the accuracy of predictive models (Afram & Janabi-Sharifi, 2014; Taheri et al., 2022). Second, the effectiveness of MPC can be compromised by changes in building usage or physical modifications, which may necessitate frequent updates to maintain accuracy and reliability (Jiang et al., 2021). These challenges underline the need for continued innovation in HVAC control strategies to enhance energy efficiency and adaptability in dynamic environments, aiming to ensure consistent environmental comfort and meet sustainability objectives within the power systems sector (Wang et al., 2023).

In response to the challenges in HVAC control, reinforcement learning (RL) has emerged as a significant tool, attracting considerable attention (Wang et al., 2023; Kumar et al., 2024). Distinct from traditional model-based approaches, RL operates as a model-free method, substantially reducing both the time and resources needed for model development (Sierla et al., 2022; Kumar et al., 2024). Moreover, RL's robust capability to interact dynamically with its environment allows it to adjust control strategies in real time, optimizing based on multiple operational criteria (Jiang et al., 2021). Such adaptability is crucial for HVAC systems, which operate under fluctuating conditions (Jiang et al., 2021). Furthermore, to mitigate the challenges associated with the extensive exploration required in RL, recent advancements have integrated RL with MPC frameworks. For instance, by leveraging MPC calculations to pretrain the RL agent, the agent can begin with a performance level similar to that of MPC, significantly reducing the need for extensive exploration (Hassanpour et al., 2024; Hassanpour, Mhaskar, & Corbett, 2024). This pretraining effectively reduces the need for prolonged exploration and enhances both the safety and efficiency of the learning process, showing considerable promise in improving the practical applicability of RL in control systems.

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

Building on these advancements, our work proposes an RL-based HVAC control strategy that leverages dynamic outdoor temperature fluctuations as key indices to balance control between thermal comfort and energy savings. It leverages dynamic outdoor temperature fluctuations as key indices to balance control between thermal comfort and energy savings. The main contributions of our work are:

1. An RL-based HVAC controller design that utilizes outdoor temperature to dynamically adjust the balance between thermal comfort and energy-saving objectives. Such an environment-adaptive adjustment along with a tailored reward mechanism enables the use of a single soft actor-critic (SAC) agent that operates efficiently without the need for additional refinement to improve the control performance. It features a model-free control strategy that can be easily deployed across different test scenarios and environmental conditions. Simulation results demonstrate the effectiveness of the proposed strategy.
2. Evaluation across two distinct test cases within the building optimization testing (BOPTEST) framework, each catering to unique climatic conditions and building typologies:
   - **Test Case 1: A heating scenario in a residential setting during winter period**. This test case, referred to within BOPTEST as "BESTEST Hydronic Heat Pump" (BOPTEST test case-Bestest hydronic heat pump, 2024), is set in a residential setting during the cold days. It allows us to evaluate our RL-based control strategy against the demands of cold weather, assessing its effectiveness in maintaining thermal comfort and energy efficiency in heating environments.
   - **Test Case 2: A cooling-dominated scenario in an office setting during a cooling period**. Known as "BESTEST Air" (BOPTEST test case-Bestest air, 2024), this scenario assesses our strategy in an office environment under high-temperature conditions. It showcases the method's versatility across different environments and its adaptability to cooling demands during hot climates, demonstrating its performance in cooling-dominated settings.

   Each test case's unique features enable a comprehensive evaluation of our RL-based HVAC control strategy, illustrating its adaptability and effectiveness across a spectrum of conditions—from cold to hot climates and from residential to office settings.

3. Extensive benchmarking to demonstrate the effectiveness of the proposed model-free RL-based approach: Leveraging the open-source nature and standardized key performance indicators (KPIs) of BOPTEST, we benchmark our method against a range of advanced controllers, including:
   - A specially optimized rule-based controller provided by BOPTEST (Blum et al., 2021).
   - Two sophisticated RL-based strategies are utilized, each employing BOPTEST's KPIs as reward references. One strategy uses a model-free RL approach, while the other employs a model-based RL method (Gao & Wang, 2023; Wang et al., 2023)
   - An MPC strategy, specifically tailored for BOPTEST (Wang et al., 2023).

While our method outperforms traditional rule-based and other RL-based strategies in thermal comfort and energy savings, it shows comparable results to MPC. Note this performance is achieved without the extensive modeling requirements of typical MPC systems, and hence suggests our model-free strategy is a scalable and efficient alternative for practical HVAC applications.

The paper is structured as follows: Section 2 reviews the related works to better highlight our contributions by showcasing the limitations of prior works. Section 3 introduces the simulation platform, followed by Section 4, which describes our method. Section 5 presents our findings and benchmarking results, and finally, Section 6 summarizes the study and outlines future directions.

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

2. Related works

Recent studies have shown the effectiveness of RL in enhancing HVAC system control and energy management. Our literature review since 2020 reveals a progressive development in this field, which falls into one or a mix of these four main trends:

- Implementation of additional refinements in actions,
- Employment of multi-agent RL algorithms,
- Incorporation of predicted observation data for improved decision-making, e.g., integration of forecasting engines to facilitate informed decision-making, and
- Comparison with model-based strategies.

2.1. Implementation of additional refinements in actions

To enhance the effectiveness of RL-based HVAC control systems, action masking is commonly employed as a refinement technique. Jiang et al. introduced a controller based on the Deep Q-Network (DQN) algorithm, which is augmented with an action processor (Jiang et al., 2021). This processor utilizes time information to refine the actions suggested by the DQN, aiming to improve decision-making efficiency. Kumar et al. provided a control strategy leveraging proximal policy optimization (PPO) and action masking (Kumar et al., 2024). This refinement mechanism limits the PPO agent's actions based on prior knowledge, leading to cost savings. Han et al. devised a novel deep-forest-based DQN control method (Han et al., 2022). The proposed strategy employs a deep-forest model to refine the original action space of DQN into a more manageable size, facilitating faster convergence.

These refinements significantly boost the control performance of RL-based HVAC control strategies. However, such additional mechanisms often involve a significant reliance on specific prior knowledge, which can constrain their adaptability across diverse scenarios (Zhang et al., 2019). Action masking, for example, theoretically effective in managing invalid actions in large discrete action spaces, presents practical implementation challenges due to its unexplored theoretical and empirical aspects (Gao, Li, & Wen, 2019; Zhang et al., 2019).

2.2. Employment of multi-agent RL algorithms

Recent advancements in HVAC system optimization increasingly favor multi-agent RL approaches over traditional single-agent methodologies. Blad et al. adopted a multi-agent RL approach for underfloor heating systems, conceptualizing the system as a Markov Game to distribute decision-making among local agents (Blad, Bøgh, & Kallesøe, 2021). This framework not only accelerates convergence but also simultaneously boosts energy efficiency and user comfort. Similarly, Yu et al. implemented a multi-agent-based strategy for commercial building HVAC control, demonstrating its effectiveness and robustness through comprehensive simulations (Yu et al., 2021). Extending these advancements, Bayer et al. developed a multi-agent RL strategy tailored for individual temperature management across different rooms, illustrating the method's precision and adaptability (Bayer & Pruckner, 2022). Building on these implementations, Fu et al., Homod et al., and Hanumaiah et al. have advanced multi-agent techniques that optimize various components of HVAC systems (Fu et al., 2022; Hanumaiah & Genc, 2021; Homod et al., 2023). Their results show significant enhancements over single-agent RL benchmarks, highlighting the advantages of multi-agent systems in managing the intricacies of large-scale power systems.

While advanced multi-agent RL-based methodologies have demonstrated improvements in managing complex and extensive HVAC systems, there remains a lack of discussion regarding the specific requirements and constraints associated with these approaches. First, the increased complexity inherent in multi-agent systems can lead to data scarcity and training inefficiencies (Gao et al., 2019; Wong et al., 2023; Zhang et al., 2019). In addition, such methodologies have faced criticism for relying on unrealistic assumptions and struggling with generalization across diverse settings (Gao et al., 2019; Wong et al., 2023).

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

Furthermore, multi-agent systems are particularly susceptible to dimensionality and nonstationarity problems that grow as the number of agents and the complexity of their interactions increase (Wong et al., 2023).

2.3. Incorporation of predicted observation data for improved decision-making

In RL-based HVAC control, predictive data specific to HVAC systems can equip RL agents with deeper insights, thereby facilitating improved control performances. Fu et al., Gao et al., and Ding et al. have developed various forecasting models to provide more details into the deep deterministic policy gradient (DDPG) algorithm's observation list, thus enhancing the control performance (Ding et al., 2022; Fu & Zhang, 2021; Gao, Li, & Wen, 2020). While these studies highlight the efficacy of DDPG in HVAC control, they also demonstrate how the integration of additional information can further enhance the performance of RL-based control methods. Fu et al. augmented a combination of DDPG and MPC HVAC control (Fu & Zhang, 2021). MPC is used for predicting energy consumption and then facilitates the decision-making process of DDPG. Gao et al. provide a deep feedforward neural network-based predictor for forecasting the occupants' thermal comfort (Gao et al., 2020). Ding et al.'s research employed a hybrid model combining support vector regression and a deep neural network to predict thermal comfort values (Ding et al., 2022). In addition, SAC is another commonly utilized RL algorithm for integrating prediction outputs to optimize HVAC control. Zhuang et al. pioneered an HVAC control system that harnesses the predictive power of 16 different LSTM-based forecasting models (Zhuang et al., 2023). These models predict indoor temperature, relative humidity, and energy consumption, providing a robust dataset for the SAC algorithm to optimize decision-making processes.

While the integration of predicted insights has shown promise in improving RL's control precision, applying these methods in real-world scenarios presents unique challenges. First, such integration necessitates a significant reliance on well-structured training data and advanced hardware, which may not always be feasible or available in many practical settings (Wang & Ahn, 2020; Wang et al., 2023). In addition, uncertainties in forecasting accuracy and variability in prediction horizons can contribute to increased system complexity, posing considerable challenges for real-world applications (Wang et al., 2023; Wang, Yao, & Papaefthymiou, 2023).

2.4. Comparison with model-based strategies

Comparisons between model-based and model-free strategies are conducted to highlight the strengths and weaknesses of each methodology within HVAC systems. Wang et al. explored both model-based approaches like MPC and model-free RL through a comparative study on the BOPTEST platform (Wang et al., 2023). Their findings suggest that both approaches are effective in optimizing HVAC control. Similarly, Gao et al. compared model-based and model-free RL, noting that model-based RL typically yields better results (Gao & Wang, 2023).

These studies provide valuable insights through performance comparisons across different methodologies, highlighting the need for further research in this direction. However, from the perspective of RL, these works mainly focus on comparisons and may not adequately address the specific applicability of each method. As a model-free strategy, RL requires comprehensive observations of the HVAC system dynamics and adjusts its strategies based on feedback. Both studies in (Gao & Wang, 2023; Wang et al., 2023) utilize BOPTEST's well-designed KPIs to design RL reward mechanisms. These KPIs are based on specific test cases within BOPTEST, incorporating pre-known information such as the number and area of test zones. While this combination simplifies the design of RL strategies and reduces model development costs, it has limitations in real-world applications. Real-world scenarios often involve dynamic and unpredictable variables, making it challenging to apply these tailored KPIs across different operational environments without significant adjustments. Currently, there is limited work extending these comparisons. As one of the early efforts, our study continues in this direction by developing a straightforward model-free control approach, emphasizing the necessity for adaptable and widely applicable control strategies in HVAC systems.

Table 1 provides a comprehensive summary of previous studies in RL-based HVAC control. Our approach introduces a straightforward single-agent RL-based control strategy that adeptly balances

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

Table 1. *Recent work in RL-based HVAC control*

	Simulation environment				Features			
	Building types		Climatic conditions		Without add. refine	Without multi-agent	Without pred. data	Compared with model-based
Ref	Residential	Office	Heating	Cooling				
Wang et al. (2023)	✓		✓		✓	✓	✓	✓
Jiang et al. (2021)		✓		✓		✓	✓	
Kumar et al. (2024)		✓		✓		✓	✓	
Zhuang et al. (2023)		✓		✓	✓	✓		
Fu et al. (2022)		✓		✓	✓		✓	
Yu et al. (2021)		✓		✓	✓		✓	
Bayer & Pruckner, (2022)		✓		✓	✓		✓	
Han et al. (2022)				✓		✓	✓	
Hanumaiah & Genc, (2021)		✓		✓	✓		✓	
Blad et al. (2021)		✓		✓	✓		✓	
Fu & Zhang, (2021)				✓	✓	✓		
Homod et al. (2023)	✓			✓	✓		✓	
Gao et al. (2020)				✓	✓	✓		
Ding et al. (2022)	✓			✓	✓	✓		
Gao & Wang, (2023)	✓		✓		✓		✓	✓
Our method	✓	✓	✓	✓	✓	✓	✓	✓

energy efficiency and thermal comfort across diverse environments. It operates without reliance on complex refinements or predictive information, which are commonly used in contemporary studies but can introduce additional challenges and dependencies. Unlike most existing research, which tends to be limited to specific climatic contexts or building setups, our study includes two distinct test cases representing different climatic conditions and building typologies. Furthermore, our benchmarking encompasses rule-based, model-based RL, model-free RL, and MPC-based approaches. This comprehensive evaluation not only confirms the effectiveness of our strategy but also bolsters confidence in its broader applicability, setting a new standard for adaptability in the field.

3. Simulation platform and test cases

The proposed work involves the utilization of an open-source simulation platform to conduct evaluations across two specific test cases. The first test case simulates heating demand conditions, while the other simulates cooling demand periods. Each test case is strategically selected to provide a thorough representation of unique architectural characteristics and distinct climatic conditions.

3.1. BOPTEST

BOPTEST stands as a versatile virtual building simulation tool, equipped with a comprehensive suite of physics-based models that replicate the dynamic behavior of real-world buildings (Blum et al., 2021). It includes 12 distinct test cases, each designed to address a variety of environmental and architectural

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

characteristics found in real-world scenarios. Moreover, the predefined and standardized test scenarios within each test case—such as "peak heat days" and "peak cool days"—are employed to validate control methods under particular environmental conditions. The combination of these uniformly identified test scenarios with BOPTEST's open-source nature enhances transparency, promotes the replication of research findings, and facilitates a thorough evaluation of various methodologies, thereby aiding in their benchmarking. In this study, we utilize two BOPTEST test cases under contrasting scenarios, with the primary objective of maintaining indoor thermal comfort while minimizing energy consumption:

- **Test Case 1—Heating Scenario: BESTEST Hydronic Heat Pump during "peak heat days".** This test case is designed to validate our control method under cold climatic conditions. The simulation models a single-zone residential building in Belgium, designed to meet the heating needs of a family of five, utilizing a heat pump-driven hydronic heating system (BOPTEST test case-Bestest hydronic heat pump, 2024). This building, with a footprint of 12 m by 16 m, faces a winter heating demand of approximately 80 W/m^2 (BOPTEST test case-Bestest hydronic heat pump, 2024). For rigorous evaluation, we leverage the standardized "peak heat days" scenario provided by BOPTEST. This scenario spans two critical weeks (from January 17th to January 31st), centered around the days with the highest annual heating load, challenging the control method's ability to manage peak demand. The scenario is implemented through the BOPTEST API with the following setting:

  Scenario="peak_heat_day"

- **Test Case 2—Cooling Dominated Scenario: BESTEST Air during "peak cool days".** This case is designed to challenge our control methodology in a hot environment. The test case simulates an office room measuring 6 m by 8 m, equipped with an idealized four-pipe fan coil unit (FCU) (BOPTEST test case-Bestest air, 2024). The climate data for this scenario is based on typical conditions near Denver, CO, USA. The "peak cool days" scenario is another standardized test period focusing on the most demanding cooling conditions, covering 2 weeks (from October 9th to October 24th) centered on the days with the highest cooling demand of the year. This scenario is activated in the BOPTEST API using the following setting:

  Scenario"=peak_cool_day"

Figure 1 illustrates the specific layout and components of each test case. To facilitate the training of RL agents, the BOPTEST framework has been integrated with the BOPTEST-Gym interface, an extension of OpenAI Gym, ensuring seamless interaction between the RL agents and the BOPTEST simulators (Blum et al., 2021). As shown in Figure 1, the RL agent interacts with the BOPTEST environment through the BOPTEST-Gym interface. Finally, Table 2 provides an overview of the two test cases used in this study, detailing the control inputs, observations, and other key variables for each scenario. Further discussion on the selection of these variables is presented in Section 4.

3.1.1. Performance assessment from BOPTEST

In pursuit of consistent performance assessment, BOPTEST offers a suite of standard KPIs, encompassing metrics like energy consumption, thermal discomfort, cost efficiency, air quality discomfort, emissions, and computational time ratio (Blum et al., 2021). Our research primarily focuses on the KPIs related to energy consumption and thermal comfort to evaluate and compare the efficacy of strategies. For thermal discomfort KPI, BOPTEST innately sets the indoor temperature boundaries considered comfortable: between 21 °C and 24 °C during occupied hours (from 7 am to 8 pm), and a broader range of 15 °C to 30 °C during unoccupied times (Blum et al., 2021). Then, the thermal discomfort KPI is expressed in K·h

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

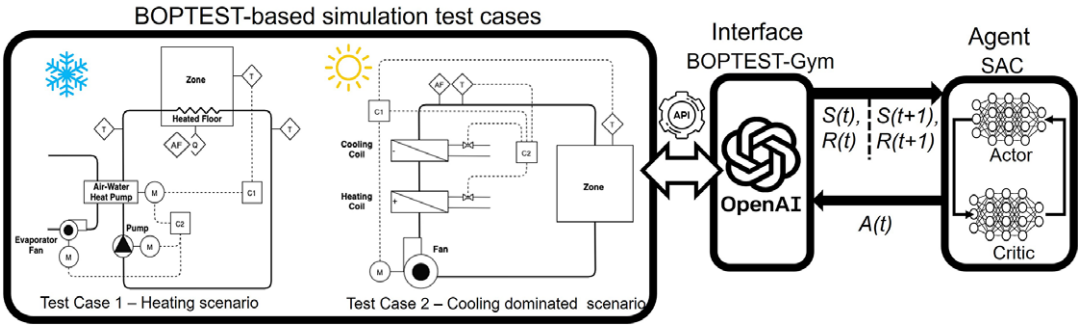

Figure 1. *Framework of the RL agent interface with the BOPTEST framework* via *BOPTEST-Gym.*

Table 2. *Overview of BOPTEST test cases with configurations used in this study*

Test case	No.1 heating scenario ('BESTEST hydronic heat pump')	No. 2 cooling dominated scenario ('BESTEST Air')
Geographic setting	Belgium	Denver, CO, USA
Building type	Single-zone residential	Single-zone office
Area	192 m^2 (12 m × 16 m)	48 m^2 (6 m × 8 m)
Köppen climate	Cfc: Temperate oceanic	BSk: Semi-arid steppe
Model type	Physics-based	
API integration	BOPTEST-Gym	
Time step	15 min (Customizable)	
Data	Synthetic (simulated scenarios)	
Assessment	Standard KPIs: Thermal discomfort KPI and energy use KPI	
Test scenario	"Peak heat days"	"Peak cool days"
HVAC system	Heat pump driven hydronic heating (no cooling)	Four-pipe FCU with heating and cooling coils
Controls	Operative Temp Setpoint	Setpoint for cooling
Observations	Real-time measured indoor air temperature, Surrounding outdoor temperature, Incremental HVAC system energy usage, Thermal comfort zone boundaries.	

(Kelvin-hours), measuring the cumulative deviation of indoor air temperature from the comfort zone (Blum et al., 2021).

The energy consumption KPI is straightforward. It accounts for the cumulative energy consumption throughout the test period, normalized to the area of the building in kWh/m^2. Further details regarding the calculation of each KPI are detailed in the related work (Blum et al., 2021). A detailed discussion of the simulation results and benchmark comparisons, based on the proposed KPIs, will be presented in Section 5.

4. RL implementation

In our study, the management of HVAC systems is formulated as a Markov Decision Process (MDP), focusing on three pivotal elements

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

$$M = \langle S(t), A(t), R(t) \rangle, \tag{4.1}$$

where

- $S(t)$: State space. This set of dynamic states for the RL agent to observe.
- $A(t)$: Action space. This set represents the decisions available to the agent.
- $R(t)$: Reward function. This function provides a value for the reward corresponding to a particular state-action pair.

The objective of the RL agent in this work is to maintain thermal comfort and optimise energy consumption. While traditional MDP formulations include transition probabilities and discount factors, our model implicitly incorporates these elements within the RL framework. The transition probabilities are embedded in the environment's dynamics, which are learned by the agent through interaction, and the reward function is designed to implicitly capture the long-term impact of actions, thus obviating the need for an explicit discount factor in our formulation.

Figure 1 illustrates the RL process: the agent continuously interacts with the environment, observing the current state $S(t)$, executing an action $A(t)$, receiving a resultant reward $R(t)$, and moving to the next state $S(t+1)$. This process is iteratively refined to enhance decision-making over time.

4.1. Design of state space: $S(t)$

It is imperative to offer a comprehensive and representative set of data for the learning agents to observe and process. In the proposed work, we emphasize that if the state space is too limited, agents may fail to achieve the desired learning outcomes; conversely, an excessively large state space can escalate computational costs, and propose obstacles for real-world applications. Consequently, to provide a flexible and unified control work, the state space $S(t)$ in this work is decided to include critical observations: real-time measured indoor air temperature $t_{\text{in}}(t)$, surrounding outdoor temperature $t_{\text{out}}(t)$, incremental HVAC system energy usage $e(t)$, which represents heating energy consumption in winter and cooling energy consumption in summer, and the observed thermal comfort zone boundaries $b_{\text{high}}(t)$ and $b_{\text{low}}(t)$.

4.2. Design of action space: $A(t)$

The action space $A(t)$ enables the agent to dynamically interact with its environment, aiming to optimise thermal comfort and energy efficiency effectively. As detailed in Table 2, each HVAC system is characterized by distinct control inputs reflecting its unique operational characteristics. Consequently, we have tailored the action spaces for each test case to suit these differences.

In Test Case 1, thermal regulation is managed by modulating the zone operative temperature setpoint, which is initially provided by BOPTEST with an adjustable range from 5 to 30 °C.

Test Case 2 presents a more complex scenario, involving both heating and cooling demands within the system. Our empirical testing has demonstrated that controlling individual heating or cooling zone temperature setpoints—whether separately for hot or cold conditions, or simultaneously—is more efficient than adjusting a singular zone supply temperature setpoint. This enhanced efficiency likely arises from the office environment's need to rapidly adapt to fluctuating internal heat gains from electronic equipment and occupancy changes, as well as external environmental variations such as sunlight exposure and outdoor temperatures. Therefore, to maintain consistency with the single action spaces used in Test Case 1, and given that this test is conducted in hot weather conditions, the selected action space for this scenario is limited to adjusting the cooling zone temperature setpoint, which is initially designed to range from 15 to 30 °C. These settings ensure precise management of thermal conditions across varied operational contexts, facilitating an effective balance between comfort and energy consumption.

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

4.3. Reward function: $R(t)$

RL operates on a reward mechanism that provides feedback to the agent concerning its actions. An effectively designed reward function is crucial, as it not only measures the progress of learning outcomes but also promotes faster convergence of the algorithm (Wang et al., 2023). The primary motivation of this work is to develop a control strategy that ensures thermal comfort while optimizing energy consumption. The most flexible and effective method to achieve these dual objectives is by utilizing environmental variations to dynamically balance control priorities. For example, when the outdoor temperature is favorable, the system prioritizes energy savings by reducing the effort expended on controlling indoor temperature. Conversely, during extreme environmental changes, our strategy adjusts its reward mechanism to ensure thermal comfort is maintained.

To this end, we have carefully designed the reward function, $R(t)$, which consists of three key components:

- The reward for thermal comfort: $R_{\text{th}}(t)$,
- The reward for energy consumption: $R_e(t)$,
- The thermal weight $\alpha(t)$

$R_{\text{th}}(t)$ and $R_e(t)$ are integrated using the time-varying, outdoor temperature-dependent weighting factor $\alpha(t)$, which ranges from 0 to 1. This allows for a dynamic trade-off that adapts to the fluctuating outdoor environment. The reward function is mathematically expressed as:

$$R(t) = \alpha(t) \times R_{th}(t) + (1 - \alpha(t)) \times R_e(t), \tag{4.2}$$

Accordingly, Section 4.3.1 details the thermal comfort reward, $R_{\text{th}}(t)$; Section 4.3.2 explains the reward for energy consumption, $R_e(t)$; and Section 4.3.3 discusses the design of the thermal weighting factor, $\alpha(t)$.

4.3.1. Thermal comfort reward:$R_{\text{th}}(t)$

The thermal comfort reward $R_{\text{th}}(t)$ assesses how closely the indoor air temperature $t_{\text{in}}(t)$ aligns with the defined thermal comfort zone, bounded by $b_{\text{high}}(t)$ and $b_{\text{low}}(t)$. To quantify this, we use a dynamic reference temperature $t_{\text{ref}}(t)$ to gauge how closely $t_{\text{in}}(t)$ matches the comfort zone. The closer $t_{\text{in}}(t)$ is to $t_{\text{ref}}(t)$, the more comfortable the environment is considered, and hence, a larger reward is given. Conversely, the further away $t_{\text{in}}(t)$ is from $t_{\text{ref}}(t)$, particularly if it falls outside the comfort zone, a smaller reward or even a penalty is applied.

As depicted in Figure 2a, $t_{\text{ref}}(t)$, is defined based on the time of day and the season to align with the occupancy and comfort requirements. Specifically, it is defined as follows:

- During occupied hours (7 am to 8 pm), $t_{\text{ref}}(t)$ is set at the median of the thermal comfort zone:

$$t_{\text{ref}}(t) = \frac{b_{\text{high}}(t) + b_{\text{low}}(t)}{2}, \quad \text{for } 7\text{am} \le t < 8\text{pm} \tag{4.3}$$

- During unoccupied periods in summer (8 pm to 7 am the next day), $t_{\text{ref}}(t)$ is set to the higher bound of the comfort zone:

$$t_{\text{ref}}(t) = b_{\text{high}}(t), \quad \text{for } 8\text{pm} \le t < 7\text{am}(\text{summer}) \tag{4.4}$$

- During unoccupied hours in winter (8 pm to 7 am the next day), $t_{\text{ref}}(t)$ is set to the lower bound of the comfort zone:

$$t_{\text{ref}}(t) = b_{\text{low}}(t), \quad \text{for } 8\text{pm} \le t < 7\text{am}(\text{winter}) \tag{4.5}$$

where $b_{\text{high}}(t)$ and $b_{\text{low}}(t)$ represent the upper and lower bounds of the thermal comfort zone, respectively.

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

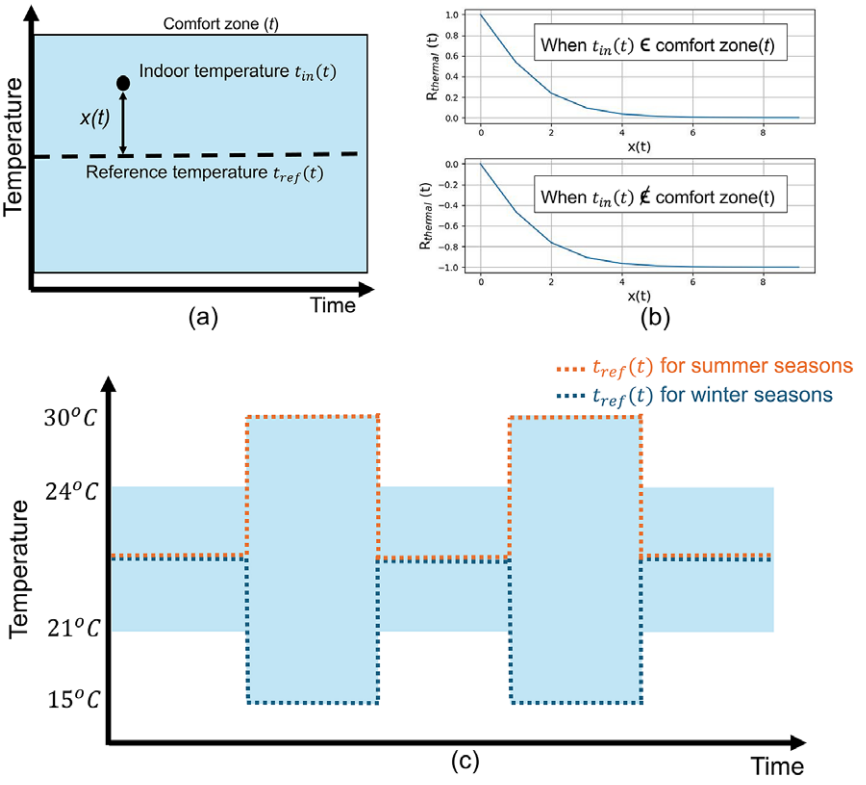

Figure 2. *The proposed reward mechanism.*

Consequently, $R_{\text{th}}(t)$ is calculated as

$$R_{\text{th}}(t) = \begin{cases} 2 - \dfrac{2}{1 + e^{-|t_{\text{in}}(t) - t_{\text{ref}}(t)|}} & \text{if } t_{\text{in}}(t) \in \left[b_{\text{low}}(t), b_{\text{high}}(t)\right] \\ 1 - \dfrac{2}{1 + e^{-|t_{\text{in}}(t) - t_{\text{ref}}(t)|}} & \text{otherwise,} \end{cases} \tag{4.6}$$

Figure 2b illustrates $R_{\text{th}}(t)$ values for both cases of t_{in} being inside or outside of the comfort zone. The exponential form of $R_{\text{th}}(t)$ ensures a smooth and continuous transition in reward values as the indoor temperature deviates from $t_{\text{ref}}(t)$. This design allows for nuanced reward or penalization, ensuring the control system remains sensitive even to slight changes in temperature, while continuously striving to maintain thermal comfort. Figure 2c shows our choices of $t_{\text{ref}}(t)$ during occupied/unoccupied period of winter/summer seasons. It is important to note that $t_{\text{ref}}(t)$ serves solely as a reference for quantifying the deviation of indoor temperature; it does not represent an ideal target for temperature tracking. For instance, if the indoor temperature is 24 °C and the comfort zone is [21 °C, 24 °C], the controller will still assign a positive reward, as the temperature remains within the range. Even if not optimal, a positive thermal reward is applied as long as the temperature stays within the comfort boundaries. In our approach, penalties are only applied when the temperature falls outside these boundaries.

4.3.2. Energy consumption reward:$R_e(t)$

The energy reward function evaluates the dynamic energy consumption by the HVAC system at each timestep. Align with the thermal reward, the energy reward is divided into positive and negative scenarios:

- Reward ($R_e(t)$>0): When the indoor temperature $t_{\text{in}}(t)$ remains within the dynamic comfort zone $[b_{\text{low}}(t), b_{\text{high}}(t)]$, the HVAC system is rewarded for efficient energy use. The reward value scales from 0 to 1, depending on the level of energy efficiency achieved.

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

- Penalty ($R_e(t)$<0): When $t_{\text{in}}(t)$ deviates from the comfort zone, the system incurs a penalty. The magnitude of the penalty is determined by the amount of energy used, ranging from 0 to −1.

Specifically, $R_e(t)$, is determined as

$$R_e(t)=\begin{cases}1-E(t) & \text{if } t_{\text{in}}(t)\in\left[b_{\text{low}}(t),b_{\text{high}}(t)\right]\\ -E(t) & \text{if } t_{\text{in}}(t)>b_{\text{high}}(t)\,(\text{winter})\text{ or } t_{\text{in}}(t)<b_{\text{low}}(t)\,(\text{summer})\\ E(t)-1 & \text{if } t_{\text{in}}(t)<b_{\text{low}}(t)\,(\text{winter})\text{ or } t_{\text{in}}(t)>b_{\text{high}}(t)\,(\text{summer})\end{cases} \tag{4.7}$$

Where:

$$E(t)=\frac{e(t)-e_{\min}}{e_{\max}-e_{\min}},t>0 \tag{4.8}$$

Here, $E(t)$ represents the normalized energy consumption of the HVAC system at time t, with $e(t)$ indicating the actual energy consumption. The constants $e_{\max}$ and $e_{\min}$ denote the maximum and minimum energy consumption observed during the training period. Specifically, $e_{\min}$ is set at 0 W, and $e_{\max}$ is fixed at 4500 W, based on empirical observations.

This dynamic approach ensures that the system's response is contextually appropriate—penalizing excessive heating or cooling that leads to discomfort and inefficiency. Such adaptability is crucial for handling varying conditions throughout the day and across seasons. We use the winter season as an example to further explain:

- If $t_{\text{in}}(t)<b_{\text{low}}(t)$—meaning the indoor temperature is below the lower boundary, indicating insufficient heating—then the more heating used, the smaller the penalty received. This policy encourages the system to use more energy to achieve a comfortable temperature by penalizing less as the energy use approaches what is necessary for comfort.
- If $t_{\text{in}}(t)>b_{\text{high}}(t)$—meaning the indoor temperature exceeds the upper boundary, indicating overheating—then the more heating used, the greater the penalty incurred. This discourages excessive heating that leads not only to discomfort but also to wasteful energy expenditure.

During summer, the focus shifts to cooling, with a similar penalty logic applied.

4.3.3. Thermal weight:$\alpha(t)$

The proposed strategy utilizes time-varying, outdoor temperature-dependent weighting factors, $\alpha(t)$, to quantitatively represent the trade-off between $R_{\text{th}}(t)$ and $R_e(t)$, which is mathematically defined as follows:

$$\alpha(t)=\frac{1}{1+e^{\beta\times(|t_{\text{out}}(t)-t_{\text{ref}}(t)|-\delta)}} \tag{4.9}$$

Where:

- β is a season-dependent parameter, set to −1 in winter and 1 in summer.
- $|t_{\text{out}}(t)-t_{\text{ref}}(t)|$ represents the absolute deviation of the outdoor temperature $t_{\text{out}}(t)$ from the reference temperature $t_{\text{ref}}(t)$. $t_{\text{ref}}(t)$ varies according to occupancy and seasonal changes, as explained in Section 4.3.1.
- δ is a threshold that fine-tunes the balance between thermal comfort and energy efficiency, adjusting the control strategy based on the outdoor temperature.

Figure 3 visualizes the adaptive behavior of the weighting factor $\alpha(t)$ in response to seasonal changes, as described by Equation 4.9. The exponential function's sigmoid shape ensures a smooth transition between

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

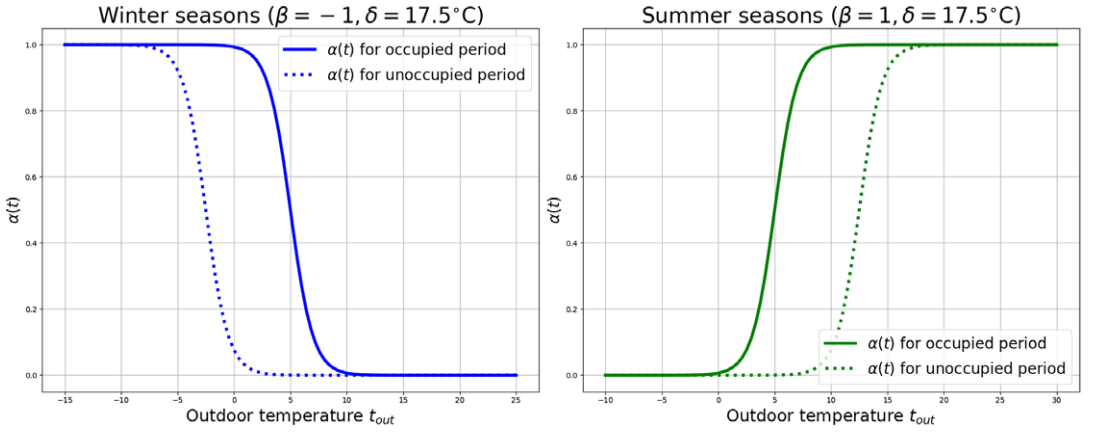

Figure 3. *Variation of thermal weight α(t) in response to seasonal outdoor temperature changes.*

prioritizing thermal comfort and energy efficiency, avoiding abrupt control shifts that could destabilize system performance. In winter, $\beta = -1$ causes $\alpha(t)$ to approach 1 during low outdoor temperatures, prioritizing thermal comfort. As outdoor temperatures become milder, $\alpha(t)$ decreases, enhancing energy efficiency. In summer, $\beta = 1$ makes $\alpha(t)$ increase as outdoor temperatures rise, emphasizing thermal control. Under milder conditions, $\alpha(t)$ decreases to focus on energy savings.

The parameter δ plays a crucial role in determining the environmental conditions under which the reward mechanism shifts focus between thermal comfort and energy usage. Based on empirical data analysis, $\delta = 17.5\,°\mathrm{C}$ to ensure consistency across test cases. The sensitivity of δ will be further discussed in Section 5.4.

4.4. Agent

In the context of HVAC control, advanced RL algorithms, such as SAC, DDPG, and PPO have been widely utilized due to their capabilities in handling complex, dynamic systems (Gao et al., 2020; Zhuang et al., 2023).

DDPG is an off-policy actor-critic algorithm that excels in high-dimensional, continuous action spaces (Gao et al., 2020). However, it often requires sensitive hyperparameter tuning and may converge to suboptimal policies, leading to significant performance variability (Gao & Wang, 2023; Zhuang et al., 2023). PPO, designed for optimizing stochastic policies, is noted for its sample efficiency and relatively easier tuning process. Despite its popularity across various applications, PPO's performance is contingent on larger batch sizes and sensitive to the choice of clipping parameters, which manage the exploration-exploitation balance (Farsang & Szegletes, 2021; Kumar et al., 2024).

In contrast, SAC distinguishes itself with some key features that address the challenges of HVAC control in real-world settings:

- Entropy-augmented exploration: SAC's entropy-augmented reward structure strikes a balance between exploration (entropy) and exploitation (reward maximization) (Haarnoja et al., 2018a,b). By encouraging diverse action exploration without relying solely on extensive real-time interactions, this mechanism enhances sample efficiency, which is crucial in real-world HVAC scenarios where interactions can be costly or potentially dangerous (Gao & Wang, 2023; Haarnoja et al., 2018; Zhuang et al., 2023).

In addition, the entropy-augmented approach helps the agent remain adaptable and robust by preventing premature convergence to suboptimal strategies (Haarnoja et al., 2018). This is particularly beneficial in

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

the presence of sensor noise or model uncertainties, allowing the agent to effectively handle fluctuating conditions and unreliable data.

- Off-policy learning: SAC's off-policy mechanism reduces the need for extensive real-time exploration, particularly in hazardous environments, by allowing the agent to refine its policy using historical data stored in a replay buffer (Haarnoja et al., 2018a,b). This approach helps mitigate concerns about the long-term random interactions often associated with RL in real-world applications.

Furthermore, this replay buffer can include noisy or uncertain data, making SAC particularly advantageous in scenarios where sensor noise or model inaccuracies are present. By learning from a diverse set of past experiences, SAC is capable of handling a wide variety of situations without relying solely on new, potentially unreliable data (Haarnoja et al., 2018; Zhuang et al., 2023).

Given these advantages of SAC, it has been selected as the preferred algorithm for our study. A detailed examination of its effectiveness and a comparative analysis with other algorithms will be provided in Section 5.

5. Simulation results

This section presents the simulation results from two test cases on BOPTEST. The simulations in this study are configured using Python 3, leveraging the Baseline3 library for implementing the SAC RL algorithm. The BOPTEST-Gym interface is used to seamlessly connect the SAC agent with the BOPTEST simulation environment, allowing for the integration and testing of our control strategies under various environmental conditions. Both two test cases undergo a training period of 100 days excluding the test period. Testing employs predefined and standardized scenarios to evaluate performance under specific conditions. The timestep for the simulations is set to 15 min, aligning with benchmark comparisons.

The SAC algorithm employed in our study is configured and optimized for performance, with consistent hyperparameters applied across both Test Case 1 and Test Case 2. These parameters are summarized in Table 3. While the table details the hyperparameters for Test Case 1, it is important to note that the same configuration is applied uniformly across all test cases for consistency and comparison. The learning rate is set at 0.001 to balance the trade-off between efficient convergence and stability during training. This value is selected after evaluating a range of learning rates, from 0.0001 to 0.01, with 0.001 consistently delivering the most stable learning outcomes across our environments. The convergence plots of our algorithm are presented and discussed in the following sections. A discount factor of 0.99 ensures the SAC agent prioritizes long-term rewards, essential for optimizing energy efficiency and thermal comfort over extended periods. The neural network architecture (400 × 300) is selected for its ability to strike an optimal balance between model complexity and computational efficiency, effectively capturing the environmental intricacies needed for control without imposing excessive computational demands (Wang et al., 2023; Zhuang et al., 2023). Furthermore, the batch size is determined to be 96, a value that correlates directly with the 15-min time resolution, spanning a full 24-h period. This configuration not only facilitates efficient learning by ensuring that each batch represents a complete daily cycle but also enhances the diversity of mini-batch samples, thereby promoting stability in gradient updates. The performance of our method is discussed in the following sections.

5.1. Results of Test Case 1—Heating scenario and comparative analysis of KPIs

Figure 4 illustrates the test results from Test Case 1—heating scenario. It should be noted that outdoor temperatures range from generally cool conditions, around 10 °C, to relatively cold weather, reaching as low as −2 °C. Despite these fluctuations, our method demonstrates a robust ability to dynamically adjust

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

Table 3. *Configuration of hyperparameters in our method and benchmark approaches*

Approach	Our method	Benchmark 1	Benchmark 2	Benchmark 3	Benchmark 4
Environment	Test Case 1: BOPTEST-"Best Hydronic Heat Pump"				
Time resolution	15 min	15 min	15 min	Not specified	15 min
Control algorithm	SAC	PI logic	DDPG	DDQN	MPC
Training period	100 days	N/A	1 year	Not specified	1 year
Test period	"Peak heat days"—14 days				
Learning rate	0.001	N/A	0.003	0.0001	N/A
Discount factor	0.99	N/A	0.95	0.99	N/A
Model architecture	400 × 300	N/A	400 × 300	200 × 200 × 200 × 200	N/A
Batch size	96	N/A	1024	512	N/A

the zone-operative temperature setpoint, effectively keeping the indoor environment within the dynamic comfort zone.

While the real-time cumulative thermal discomfort KPIs show slight degradation, the overall performance remains satisfactory. Energy usage KPI displays a continuous upward trend throughout the test period. A comparison between our approach and several optimized and advanced benchmarks is provided in Section 5.1.1, offering further insights into our method's performance.

To illustrate the learning process, we include a cumulative reward plot, which provides deeper insight into the agent's performance across training episodes. The smooth upward trajectory of cumulative rewards, which eventually stabilizes, indicates that our control strategy is progressively optimizing over time. The trajectory of cumulative reward, aligned with both satisfactory thermal discomfort and energy use KPIs, underscores the effectiveness of our approach in achieving optimal control under cold environmental conditions.

5.1.1. Benchmarking based on BOPTEST KPIs

An in-depth comparative analysis of KPIs offers a nuanced assessment of our strategy relative to established benchmarks. This analysis highlights our method's strengths and identifies opportunities for further enhancements. Since all benchmarks, including our approach, are implemented on the open-sourced BOPTEST platform and evaluated under standardized testing scenarios, the KPIs enable direct and reliable comparisons. The benchmarks involved in this work are introduced below:

Benchmark 1 is the embedded rule-based HVAC control method inherently implemented within BOPTEST (Blum et al., 2021). As the baseline controller, it employs optimized PI logic to maintain the operative zone temperature, providing adequate indoor comfort without excessive energy use.

Benchmark 2 features an RL-based controller developed by Wang et al. (Wang et al., 2023). This controller employs a reward function intricately designed around the BOPTEST KPIs. The reward function for the next step is formalized as follows (Wang et al., 2023):

$$r_{t+1} = -\left(J_{t+1} - J_t\right), \tag{5.1}$$

where the cost function J_t at the current step t is defined by:

$$J_t = \text{Cost} + 10 \times t\text{dis}. \tag{5.2}$$

In this model, Cost indicates the operational cost KPI, and $t$dis is the thermal discomfort KPI. The coefficient ω serves as a balancing factor between operational cost and thermal discomfort, representing a critical hyper-parameter requiring careful tuning.

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

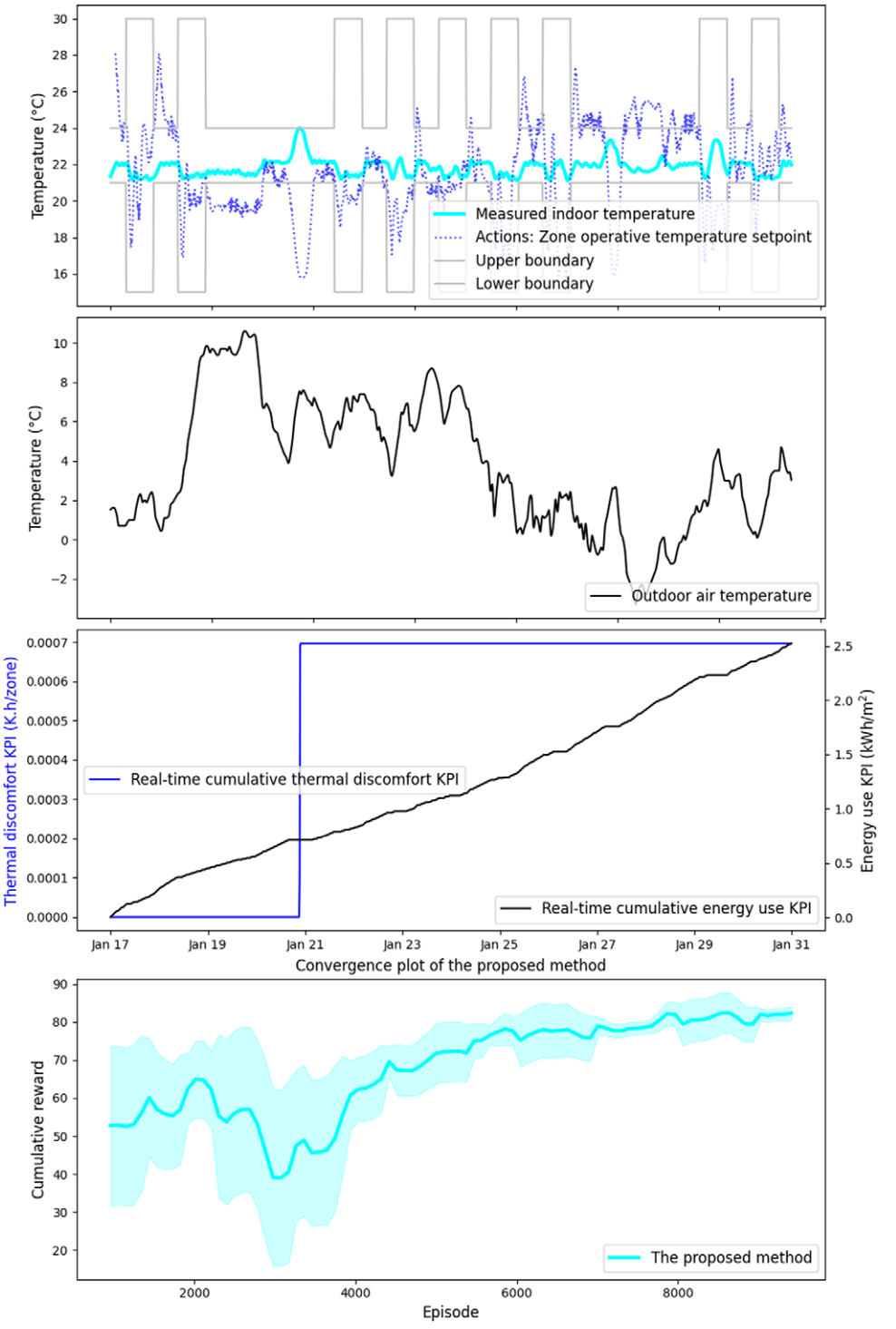

Figure 4. *Simulation results for Test Case 1: Heating scenario during cold days.*

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

The DDPG algorithm serves as the control agent. Through rigorous testing and comparisons with other RL algorithms, such as SAC and Double Deep Q-Network (DDQN), Wang et al. (Wang et al., 2023) demonstrated that the DDPG-based strategy, when paired with a customized reward function, outperforms other approaches. Consequently, we have selected the best performer from Wang et al.'s work, namely the DDPG-based controller, as Benchmark 2. This allows for a thorough evaluation of our proposed method, ensuring that our comparisons reflect the most effective version of their approach.

Benchmark 3 is another RL-based strategy developed by Gao et al. (Gao & Wang, 2023). Operating within the same test environment and employing a unified reward mechanism, this strategy provides a detailed comparative analysis of model-based versus model-free RL algorithms by cycling through various RL agents.

Similar to Benchmark 2, Benchmark 3 utilizes a reward function intricately designed around BOPTEST KPIs. Its current cost J_t is the same as Benchmark 2, function 5.2. However, its final reward for step t is calculated as:

$$r_t = 0.05 \times (J_{t-1} - J_t), \tag{5.3}$$

According to their critical evaluation, Gao et al. demonstrate that model-based RL agents, particularly the model-based DDQN, excel when integrated with the proposed reward function. This approach achieves the lowest thermal discomfort KPI while still maintaining competitive energy usage. Consequently, we select this specific model-based approach as Benchmark 3 for our study.

Benchmark 4 is an MPC-based control strategy developed by Wang et al. (Wang et al., 2023), specifically tailored for the "BESTEST Hydronic Heat Pump" test case utilized in this study. This model leverages a data-driven gray-box approach, utilizing a thermal resistance-capacitance (RC) network to simulate building thermal dynamics effectively. The RC model chosen for this work is a first-order system, preferred for its ability to transform the control problem into a linear programming problem, where global optima are easily attainable. The following formula details the model; it includes a lumped parameter and RC representations focusing on a single system state T_z, representing the zone operative temperature. This state parameter evolves according to the differential equation (Wang et al., 2023):

$$C\frac{dT_z}{dt} = \frac{T_{\text{out}} - T_z}{R} + q_{\text{HVAC}} + q_{\text{inter}} + A \times I_{\text{solar}}$$

where:

- C and R denote the thermal capacitance and resistance of the zone and envelops, respectively.
- T_{out} indicates the outdoor temperature, as mentioned in Section 4.
- I_{solar} represents the solar irradiation.
- q_{HVAC} indicates the heat influx from the HVAC system.
- q_{inter} indicates the internal heat load from occupants, lighting, and equipment.
- A refers to the effective area of the windows.

This model-based control approach allows for an in-depth comparison with other advanced methodologies within both model-based and model-free paradigms, highlighting the sophisticated handling of dynamic thermal responses in building environments.

The hyperparameters utilized in the (Blum et al., 2021; Gao & Wang, 2023; Wang et al., 2023) are also outlined in Table 3.

Figure 5 displays the cumulative KPIs for thermal discomfort and energy consumption for each control strategy evaluated. Our method shows a significant improvement in managing discomfort levels,

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

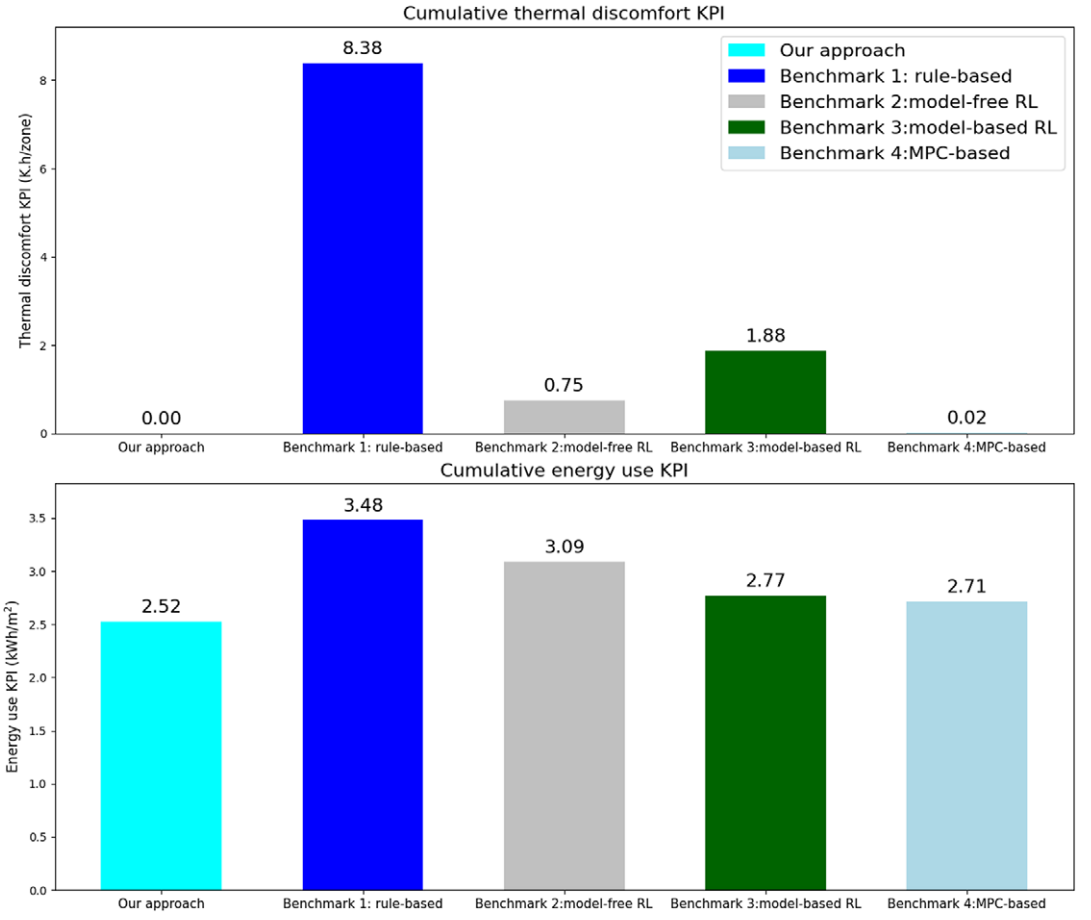

Figure 5. *Benchmark cumulative KPIs across the 2 weeks heating scenario in test case 1.*

achieving the lowest thermal discomfort KPIs among all tested methods. Similarly, our approach also reports the lowest energy usage KPI, indicating that it not only excels in thermal comfort but also in energy efficiency.

Despite Benchmark 1 being optimized specifically for the test scenario, it still records the highest KPIs in terms of both thermal comfort and energy efficiency.

Among the RL-based methodologies, which include our approach (model-free RL), Benchmark 2 (model-free RL), and Benchmark 3 (model-based RL), our strategy stands out by providing superior thermal comfort. Although all three RL-based strategies exhibit similar levels of energy consumption, our method shows the most efficient energy use. This suggests that while our method achieves the best thermal comfort control, it also reaches comparable energy usage levels to the other RL-based strategies.

Benchmark 4, employing an MPC-based control strategy, demonstrates performance closely aligned with our method in terms of both energy use and thermal discomfort KPIs. Considering the performances of both Benchmark 3 and 4, it becomes evident that model-based strategies, such as Benchmark 3 (model-based RL) and Benchmark 4 (MPC-based), potentially offer better energy management performance than model-free approaches, though they are limited by the significant time required for model development and refinement.

Our proposed model-free strategy not only matches the thermal comfort provided by leading model-based benchmarks but also competes strongly in energy efficiency, offering a compelling alternative that

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

bypasses the extensive modeling requirements of typical MPC systems. This approach makes it a practical solution for HVAC applications.

5.2. Results of Test Case 2: Cooling-dominated scenario

Figure 6 presents the simulation outputs for Test Case 2, highlighting our method's capability to adapt to environmental changes during hot seasons. This test case features more dramatic fluctuations in outdoor temperatures, ranging from near-freezing conditions around 0 °C to relatively hot conditions exceeding 25 °C. Consistent with the findings from Test Case 1, our method demonstrates a strong ability to dynamically adjust the zone temperature setpoint for cooling, ensuring that the indoor environment remains within the dynamic comfort zone despite these extreme variations.

The results for the thermal discomfort KPI and energy use KPI further confirm the efficiency of our method. It achieves commendable control over thermal discomfort while maintaining a similar level of energy consumption as observed in Test Case 1. Additional insights into the robustness and efficiency of the proposed method across these two distinct scenarios will be discussed in Section 5.3.

The cumulative reward plot mirrors the results observed in Test Case 1, characterized by a smooth upward trend in rewards that eventually stabilizes. This pattern suggests that our control strategy is successfully optimizing over time, adapting to varying environmental conditions.

5.3. Test Case Scenario Analysis: Adaptability and Robustness

While Sections 5.1 and 5.2 demonstrate the effectiveness of the proposed method in separate heating and cooling scenarios, this section extends the analysis by taking a holistic perspective to compare how our strategy adapts and maintains robust control across different environmental conditions, building typologies, and HVAC system variations. We first discuss the distinct characteristics of each test case.

5.3.1. Fluctuating environmental conditions

Figure 7 provides a comparative analysis of the outdoor temperatures observed in the two test cases. The box plot in Figure 7a illustrates the mean outdoor temperatures and their variability, revealing that while Test Case 1 presents challenging cold climatic conditions, Test Case 2 exhibits significantly greater variability. To quantify the challenges posed by these variations, we provide the first-order differences of the outdoor temperatures for both test cases and analyze their distributions. This statistical measure captures changes between consecutive data points, reflecting both the magnitude and direction of temperature shifts within each 15-min interval in this study. This analysis is crucial for understanding the dynamics of the outdoor environment, as it quantifies the rapidity and extent of temperature shifts from one time step to the next.

Figure 7b clearly demonstrates that while both cases exhibit significant step-by-step fluctuations, the first-order differences of outdoor temperatures in Test Case 2 show a broader distribution, indicating more rapid and frequent temperature changes compared with Test Case 1. These findings align with the climatic characteristics detailed in Table 2, where Test Case 2, classified as BSk (Cold Semi-Arid Steppe), is identified as more dynamic than the Cfc (Temperate Oceanic) climate of Test Case 1 (Semi-arid climate, 2024; Semi-arid climate, 2024). The heightened dynamism in BSk climates, characterized by extreme temperature fluctuations and variable precipitation patterns, imposes significant demands on any control method, requiring it to dynamically adjust its actions to consistently meet thermal control objectives.

These test cases encompass a range of environmental conditions—from cold to hot, and from relatively stable to highly dynamic climates—proving the ability of the RL controller to perform under variable climatic conditions.

5.3.2. Varying building typologies and HVAC system complexities

Beyond climatic influences, differences in building typology and HVAC system complexity among the test environments further challenge the adaptability and robustness of our approach. The building

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

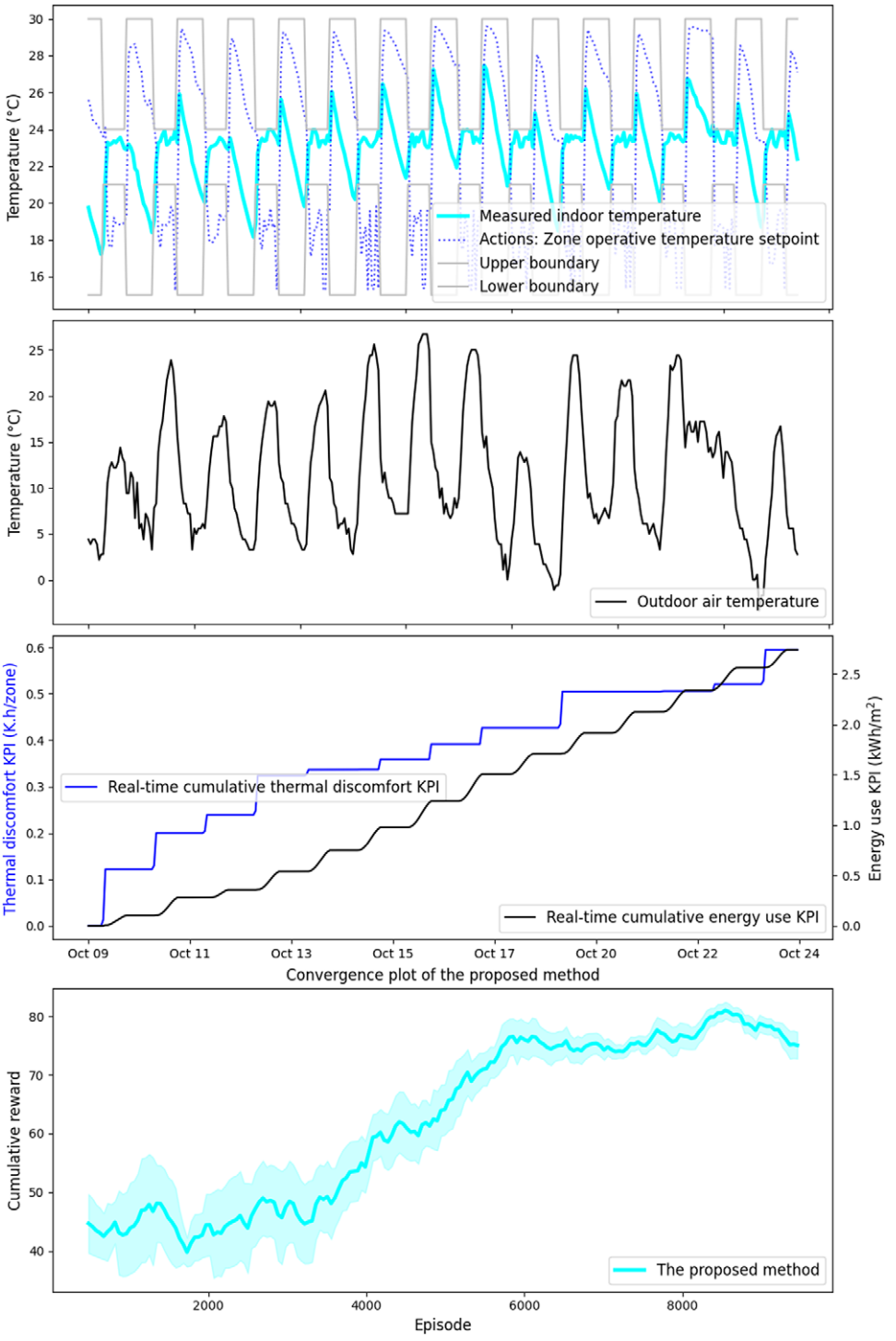

Figure 6. *Simulation results for Test Case 2: Cooling dominated scenario.*

envelope materials for both test cases are based on the BESTEST Case 900 building (Blum et al., 2021; BOPTEST test case-Bestest hydronic heat pump, 2024; BOPTEST test case-Bestest air, 2024; Judkoff & Neymark, 1995). However, as detailed in Table 2, Test Case 1 is designed to evaluate control methods in a

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

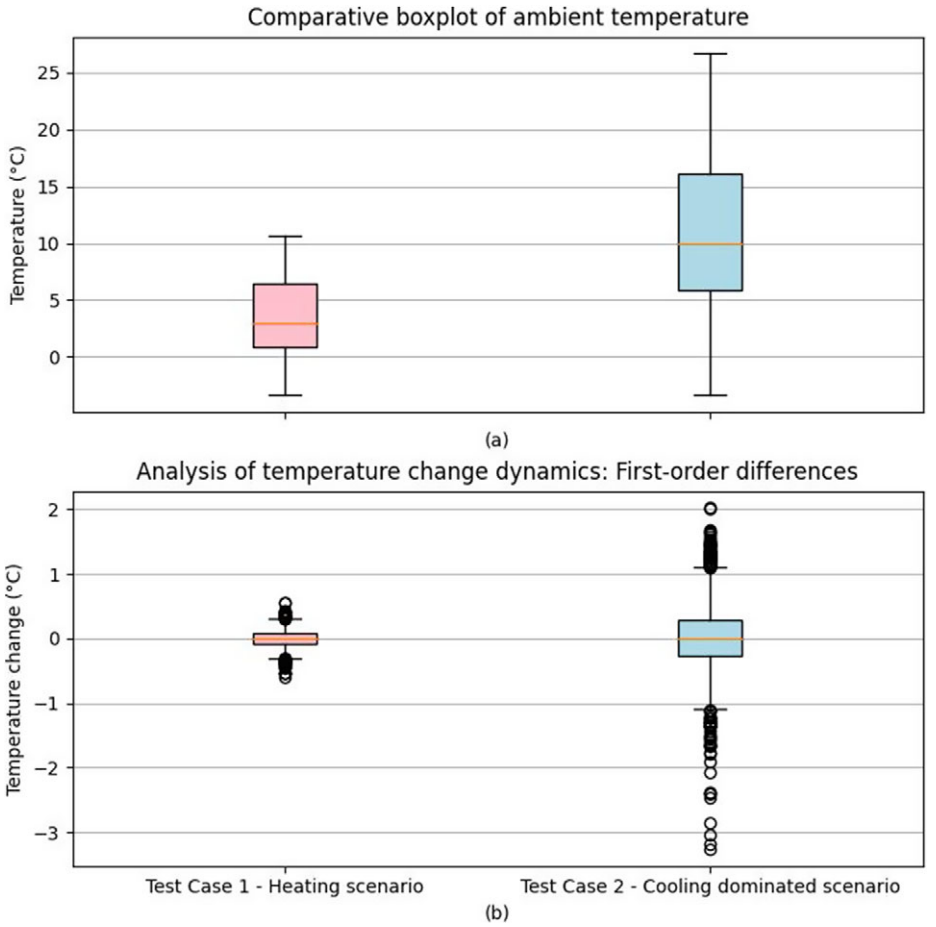

Figure 7. *Temperature Profiles and Variability: Comparative Analysis of two test cases.*

residential setting under cold environments. For this, the building's envelope and internal wall mass are scaled to four times the area of the original design, resulting in a larger single-zone residential setting with substantial thermal mass. Conversely, Test Case 2 adheres to the original BESTEST design dimensions and simulates a smaller office environment with both heating and cooling. As an office setting, it features large south-facing windows and a higher window-to-wall ratio, which significantly increases solar gains. These design elements expose the interior more directly to external temperature variations, necessitating more frequent and precise adjustments by the HVAC system to maintain comfort.

These varying building settings—ranging from large to small, residential to office, and from heating-only to both heating and cooling—highlight the adaptability challenges inherent in applying control strategies across different environments.

5.3.3. Integrated performance assessment

Figures 4 and 6 present the real-time cumulative KPIs for the two distinct test cases under uniform experimental settings, see Tables 2 and 3. In Test Case 1, which simulates a cold climatic environment under a residential setting, our method effectively maintains thermal comfort and energy efficiency, as evidenced by an almost zero thermal discomfort KPI and stable energy usage. In contrast, Test Case 2 introduces hot and dynamic environmental conditions along with a different building typology and HVAC setting. Despite these increased challenges, our method continues to perform well, with only a

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

slight increase in the thermal discomfort KPI to 0.60 K·h per zone over the 2-week evaluation period. This minor deviation underscores the method's adaptability and robustness to challenging conditions without compromising overall performance. The consistency in energy usage KPIs across both test cases is particularly noteworthy, highlighting the robustness of our strategy in maintaining energy efficiency across different building typologies and climatic conditions.

5.4. Parameter study

The core of this study lies in the design of an effective reward mechanism specifically tailored to the operational characteristics of HVAC systems. The mechanism consists of two components: a thermal reward and an energy reward, linked by a dynamic, time-varying outdoor temperature-dependent thermal weight $\alpha(t)$. This weight adapts to changing climatic conditions, ensuring a balanced consideration of both thermal comfort and energy efficiency. To demonstrate the effectiveness of the proposed reward mechanism, this section first compares the efficiency of the dynamic thermal weight mechanism with static weighting, followed by a sensitivity analysis of the parameter δ.

5.4.1. Comparative analysis of dynamic versus static thermal weight

To demonstrate the effectiveness of the proposed $\alpha(t)$, we conducted simulations comparing it with a constant thermal weight α, set at 0.5, under various scenarios. These simulations help assess the impact of dynamic versus static weighting on HVAC performance, particularly in terms of thermal comfort and energy efficiency.

Figure 8 illustrates the results of these simulations. The outcomes indicate that our method, with its well-designed reward mechanism, provides satisfactory results for both dynamic $\alpha(t)$ and fixed α. However, the KPIs provide deeper insights:

- Thermal comfort: Across various climatic and architectural conditions, our adaptive method consistently delivers high levels of thermal comfort. In contrast, the fixed weight α performs well

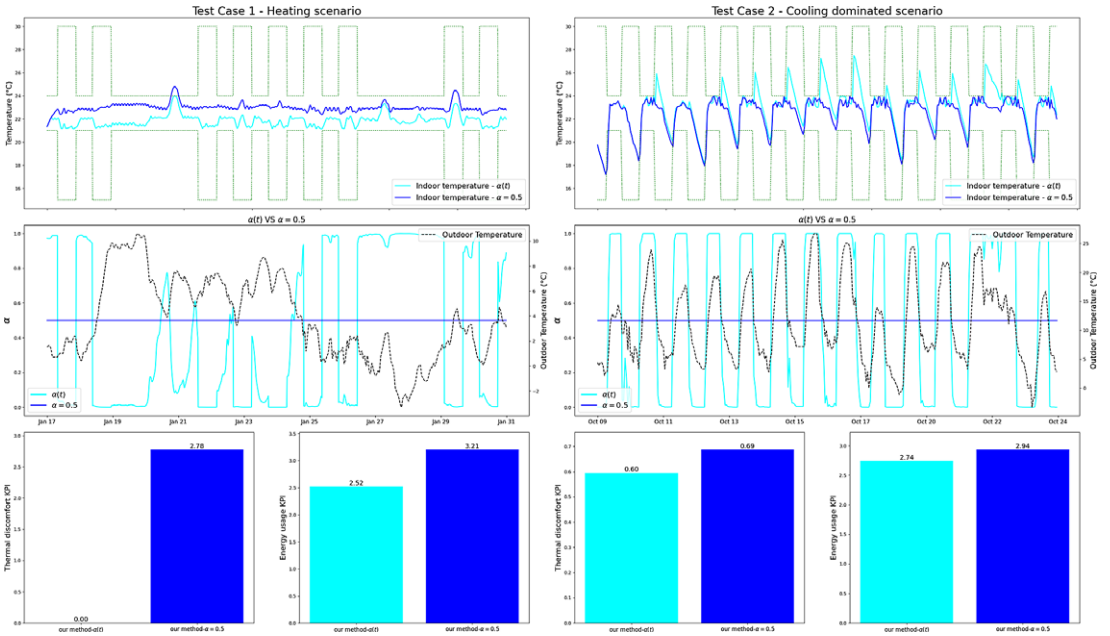

Figure 8. *Comparison of dynamic thermal weight* $\alpha(t)$ versus *fixed thermal weight* $\alpha = 0.5$.

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

in Test Case 2 but significantly underperforms in Test Case 1. This discrepancy highlights the importance of the adaptive, outdoor temperature-dependent $\alpha(t)$, which dynamically optimises HVAC operations to maintain comfort efficiently.

- Energy efficiency: The dynamic $\alpha(t)$ also excels in energy conservation, outperforming the fixed α across scenarios. By leveraging real-time outdoor temperature data, the strategy activates heating or cooling only when deemed necessary by the agent. This approach not only ensures optimal thermal comfort but also enhances energy efficiency.

These findings underscore the dual advantages of the proposed $\alpha(t)$ in both heating and cooling seasons. It enables adaptive management of HVAC systems, adjusting to fluctuating external conditions to optimise both comfort and energy use. This robustness and efficiency affirm the potential of $\alpha(t)$ as a critical component in future HVAC control strategies.

5.4.2. δ Sensitivity analysis

Figure 9 illustrates how δ influences the controller's balance between thermal comfort and energy efficiency, based on deviations between outdoor temperature $t_{\text{out}}(t)$ and reference temperature $t_{\text{ref}}(t)$, as defined in Formula 4.9. As noted in Section 4.3.1, $t_{\text{ref}}(t)$ varies between occupied and unoccupied hours. However, to avoid redundancy and enhance clarity, the x-axis in Figure 9 represents the absolute deviation between $t_{\text{out}}(t)$ and $t_{\text{ref}}(t)$. Because of this, both occupied and unoccupied hours are represented by the same line. To evaluate the sensitivity of δ, we conduct simulations with three distinct values:

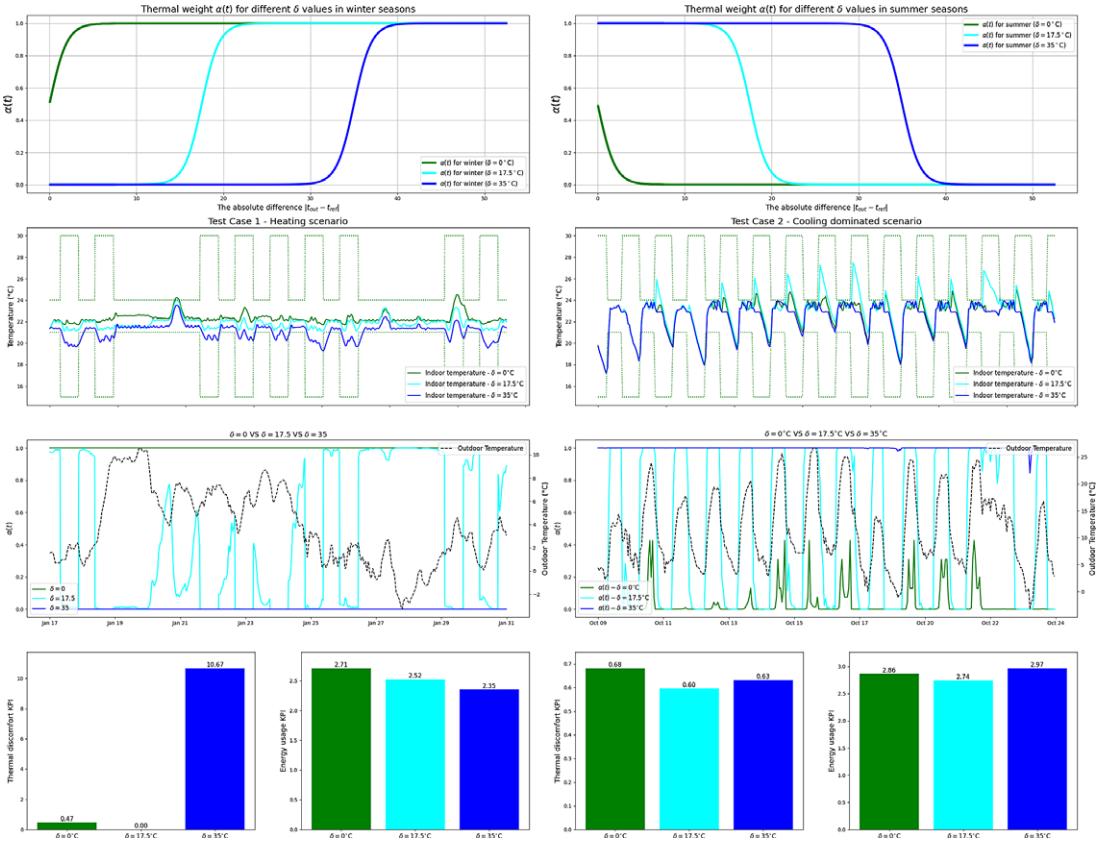

Figure 9. *δ sensitivity analysis.*

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

- $\delta = 0\,^{\circ}\mathrm{C}$: Serves as a baseline with no threshold, causing the balance to shift even with minimal deviations between $t_{\mathrm{out}}(t)$ and $t_{\mathrm{ref}}(t)$.
- $\delta = 17.5\,^{\circ}\mathrm{C}$: The empirically determined value, where the balance adjusts when the deviation is around 17.5, providing a moderate and balanced response.
- $\delta = 35\,^{\circ}\mathrm{C}$: Tests the method under more extreme conditions, with adjustments occurring when the deviation reaches around 35.

The simulation results, as illustrated in Figure 9, demonstrate the effectiveness of the proposed reward mechanism across different test cases. In our reward function for HVAC system control, energy consumption is optimized only when thermal comfort is maintained. This design inherently ensures the dual objectives are met simultaneously, making sure that energy savings are not prioritized at the expense of thermal comfort. However, in the dynamic HVAC control process, the degree to which each objective is prioritized depends on the value of δ, which plays a critical role in fine-tuning this balance. In extreme cases, where δ is set to favor one objective heavily, the system may still deliver adequate results, but there could be compromises in either energy efficiency or thermal comfort. The detailed KPIs reveal important nuances in the performance of varying δ values.

- **Test Case 1—Heating scenario:** In winter, $\delta = 0\,^{\circ}\mathrm{C}$ causes $\alpha(t)$ to remain close to 1 for most of the test period, emphasizing thermal comfort. While this results in a thermal discomfort KPI comparable to that of the empirically set $\delta = 17.5\,^{\circ}\mathrm{C}$, it neglects energy efficiency, leading to a higher energy usage KPI. Conversely, setting $\delta = 35\,^{\circ}\mathrm{C}$ prioritizes energy efficiency, as reflected by the lower energy usage KPI. However, this setting under severe cold causes the controller to consistently maintain the indoor temperature near the lower boundary of the thermal comfort zone. While this approach is effective in conserving energy, it increases the risk of the indoor temperature falling outside the comfort zone. As a result, this trade-off leads to a higher thermal discomfort KPI, indicating a compromise in maintaining thermal comfort. The empirical setting of $\delta = 17.5\,^{\circ}\mathrm{C}$ offers the best balance, providing satisfactory thermal comfort and reasonable energy efficiency.
- **Test Case 2—Cooling-dominated scenario:** A similar pattern is observed in Test Case 2. The setting $\delta = 35\,^{\circ}\mathrm{C}$ prioritizes thermal comfort, with $\alpha(t)$ close to 1, resulting in a low thermal discomfort KPI. However, this comes with increased energy usage, indicated by a slightly higher energy usage KPI. On the other hand, with $\delta = 0\,^{\circ}\mathrm{C}$, $\alpha(t)$ stays close to 0, focusing on energy efficiency once thermal comfort is achieved. Although this reduces energy consumption, it results in the highest thermal discomfort KPI, suggesting a slight compromise in comfort to achieve greater energy savings. Once again, $\delta = 17.5\,^{\circ}\mathrm{C}$ achieves the optimal balance, effectively managing both thermal comfort and energy efficiency.

In summary, among the three settings, $\delta = 17.5$ provides the optimal balance, enabling the controller to respond appropriately to environmental changes while avoiding overreactions to minor deviations and underreactions to significant ones. This balance ensures consistent performance across diverse operational conditions.

6. Conclusion

This study introduces and evaluates an advanced, environment-adaptive RL-based control strategy for HVAC systems, demonstrating its universality across different operating conditions. Employing a SAC agent, it features a reward mechanism tailored to HVAC system characteristics. This mechanism dynamically balances thermal comfort and energy efficiency by incorporating outdoor temperature-dependent weighting factors, thereby optimizing performance and reducing the need for further refinements.

We evaluated our approach across different scenarios using the open-source simulation platform BOPTEST, to assess its effectiveness under a variety of architectural and climatic conditions. The

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

evaluations cover a spectrum of environments, ranging from residential settings during cold weather to office settings in warmer climates. This diversity allows for a comprehensive assessment of the proposed strategy's efficiency. Simulation results demonstrate that our method consistently delivers satisfactory outcomes across these scenarios, underscoring the strategy's broad applicability and robustness.

Furthermore, we compare our method with several advanced control strategies in the heating scenario in BOPTEST. These include a customized rule-based controller, two sophisticated RL-based strategies using BOPTEST's KPIs as reward references, and a MPC approach specifically developed for BOPTEST. The results reveal that our method surpasses traditional rule-based and other RL-based strategies in maintaining thermal comfort and optimizing energy consumption. Importantly, it achieves outcomes comparable to the MPC-based controller but with reduced complexity and no reliance on precise modeling, demonstrating its flexibility for real-world applications.

It should be noted that while BOPTEST provides a robust framework for simulating HVAC control strategies, it is relatively new and lacks comprehensive benchmarking across all its test scenarios. Previous studies have predominantly focused on the residential heating test case. This focus has resulted in a scarcity of extensive comparative KPIs for newer scenarios like the office setting cooling scenario, which have not been as thoroughly explored. Despite these limitations, our research serves as a pioneering effort, establishing foundational benchmarks for future investigations. In our ongoing efforts to advance the field, we aim to evolve our method into a more adaptable framework that reduces reliance on predefined parameters, enabling greater flexibility across varied environments. In addition, to address the limitations of the physics-based BOPTEST platform in assessing stability against sensor noise and model uncertainties, and to further demonstrate our method's robustness and efficiency across various building types, future work will focus on adapting the strategy to a broader range of simulation platforms, including those leveraging real-world data and machine learning-based simulators. Furthermore, to enhance the implementation of RL in real-world settings, particularly where extensive random interactions are typically required, we plan to explore transfer learning techniques to adapt learned policies to real-world conditions, ensuring safer and more effective deployment.

Data availability statement. The data and resources used in this study are fully accessible through the open-source BOPTEST platform, which provides standardized test cases and scenarios utilized to simulate and validate the proposed control methods. All models, scripts, and configuration files employed in this work are based on the standard offerings within the BOPTEST framework, publicly available at the official repository: https://github.com/ibpsa/project1-boptest. Additionally, the control approach developed in this study is open-source. The full implementation, including all code and relevant resources, is available on the GitHub repository: https://github.com/xinlin-CSIRO/RL_based_HVAC_control_Boptest. An archived version of this repository with a permanent DOI is accessible on Zenodo: https://doi.org/10.5281/zenodo.14178227. No proprietary or custom data were created for this study.

Author contribution. Xinlin Wang: Conceptualization (equal); Software (lead); Validation (lead); Formal Analysis (lead); Data Curation (lead); Writing—Original Draft (lead); Writing—Review and Editing (equal); Visualization (lead). Nariman Mahdavi: Conceptualization (lead); Software (supporting); Validation (supporting); Formal Analysis (supporting); Data Curation (supporting); Writing—Original Draft (supporting); Supervision (supporting); Writing—Review and Editing (equal); Visualization (supporting); Project Administration (equal). Subbu Sethuvenkatraman: Conceptualization (lead); Supervision (lead); Project Administration (equal); Writing—Review and Editing (equal). Sam West: Conceptualization (supporting); Supervision (supporting); Writing—Review and Editing (equal).

Competing interest. The authors declare no competing interests.

Funding statement. This research is supported by the AI for Missions program of the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

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.

Bayer D and Pruckner M. "Enhancing the performance of multi-agent reinforcement learning for controlling HVAC systems." In: *2022 IEEE Conference on Technologies for Sustainability (SusTech)*. 2022, pp. 187–194.

Blad C, **Bøgh S**, & **Kallesøe C** (2021). A multi-agent reinforcement learning approach to price and comfort optimization in hvac-systems. *Energies*, *14*(22), 7491.

Blum D, *et al.* (2021). Building optimization testing framework (BOPTEST) for simulation-based benchmarking of control strategies in buildings. *Journal of Building Performance Simulation*, *14*(5), 586–610.

BOPTEST test case-Bestest air. (2024) url: https://ibpsa.github.io/project1-boptest/testcases/ibpsa/testcases_ibpsa_bestest_air/.

BOPTEST test case-Bestest hydronic heat pump. (2024) url: https://ibpsa.github.io/project1-boptest/testcases/ibpsa/testcases_ibpsa_bestest_hydronic_heat_pump/.

Ding Z-K, *et al.* (2022). Energy-efficient control of thermal comfort in multi-zone residential HVAC via reinforcement learning. *Connection Science*, *34*(1), 2364–2394.

Farsang M and Szegletes L. "Decaying clipping range in proximal policy optimization." In: 2021 *IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. IEEE. *2021*, pp. 000521–000526.

Fu C, & **Zhang Y** (2021). Research and application of predictive control method based on deep reinforcement learning for HVAC systems. *IEEE Access*, *9*, 130845–130852. doi:10.1109/ACCESS.2021.3114161.

Fu Q, *et al.* (2022). Optimal control method of HVAC based on multi-agent deep reinforcement learning. *Energy and Buildings*, *270*, 112284.

Gao C, & **Wang D** (2023). Comparative study of model-based and model-free reinforcement learning control performance in HVAC systems. *Journal of Building Engineering*, *74*, 106852.

Gao G, **Li J, and Wen Y**. "Energy-efficient thermal comfort control in smart buildings via deep reinforcement learning." *In: arXiv preprint arXiv:1901.04693* (2019).

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.

Haarnoja T *et al.* "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor." In: *International conference on machine learning*. PMLR. 2018a, pp. 1861–1870.

Haarnoja T *et al.* (2018b). "Soft actor-critic algorithms and applications." In: *arXiv preprint arXiv:1812.05905*.

Han Z, *et al.* (2022). Deep Forest-Based DQN for CoolingWater System Energy Saving Control in HVAC. *Buildings*, *12*(11).

Hanumaiah V and Genc S (2021). "Distributed multi-agent deep reinforcement learning framework for whole-building HVAC control." *In: arXiv preprint arXiv:2110.13450.*

Hassanpour H, **Mhaskar P**, & **Corbett B** (2024). A practically implementable reinforcement learning control approach by leveraging offset-free model predictive control. *Computers & Chemical Engineering*, *181*, 108511.

Hassanpour H, *et al.* (2024). A practically implementable reinforcement learning-based process controller design. *AIChE Journal*, *70*(1), e18245.

Homod RZ, *et al.* (2023). Deep clustering of cooperative multi-agent reinforcement learning to optimize multi chiller HVAC systems for smart buildings energy management. *Journal of Building Engineering*, *65*, 105689.

Jiang Z, *et al.* (2021). Building HVAC control with reinforcement learning for reduction of energy cost and demand charge. *Energy and Buildings*, *239*, 110833.

Judkoff R and Neymark J. (1995) *International Energy Agency building energy simulation test (BESTEST) and diagnostic method (No. NREL/TP-472-6231). National Renewable Energy Lab.(NREL), Golden, CO (United States)*.

Kumar SR *et al.* (2024) "Towards safe model-free building energy management using masked reinforcement learning." In: *IEEE ISGT conference.*

Moghaddasi H, *et al.* (2021). Net zero energy buildings: variations, clarifications, and requirements in response to the Paris Agreement. *Energies*, *14*(13), 3760.

Semi-arid climate. (2024) URL: https://en.wikipedia.org/wiki/Semi-arid_climate.

Semi-arid climate. (2024) url: https://en.wikipedia.org/wiki/Oceanic_climate.

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.

Taheri S, **Hosseini P**, & **Razban A** (2022). Model predictive control of heating, ventilation, and air conditioning (HVAC) systems: A state-of-the-art review. *Journal of Building Engineering*, 105067.

Wang D, *et al.* (2023). Comparison of reinforcement learning and model predictive control for building energy system optimization. *Applied Thermal Engineering*, *228*, 120430.

Wang X, & **Ahn S-H** (2020). Real-time prediction and anomaly detection of electrical load in a residential community. *Applied Energy*, *259*, 114145.

Wang X, **Yao Z**, & **Papaefthymiou M** (2023). A real-time electrical load forecasting and unsupervised anomaly detection framework. *Applied Energy*, *330*, 120279. doi:10.1016/j.apenergy.2022.120279.

Wang X, *et al.* (2023). AI-empowered methods for smart energy consumption: A review of load forecasting, anomaly detection and demand response. *International Journal of Precision Engineering and Manufacturing-Green Technology*, 1–31.

Wong A, *et al.* (2023). Deep multiagent reinforcement learning: Challenges and directions. *Artificial Intelligence Review*, *56*(6), 5023–5056.

Yu L, *et al.* (2021). Multi-agent deep reinforcement learning for HVAC control in commercial buildings. *IEEE Transactions on Smart Grid*, *12*(1), 407–419.

https://doi.org/10.1017/dce.2024.57 Published online by Cambridge University Press

Yuan X *et al.* "Study on the application of reinforcement learning in the operation optimization of HVAC system." In: *Building Simulation*. Vol. *14*. Springer. 2021, pp. 75–87.

Zhang Z, *et al.* (2019). Asynchronous episodic deep deterministic policy gradient: Toward continuous control in computationally complex environments. *IEEE transactions on cybernetics*, *51*(2), 604–613.

Zhuang D, *et al.* (2023). Data-driven predictive control for smart HVAC system in IoT-integrated buildings with time-series forecasting and reinforcement learning. *Applied Energy*, *338*, 120936.

Cite this article: Wang X, Mahdavi N, Sethuvenkatraman S and West S (2025). An environment-adaptive SAC-based HVAC control of single-zone residential and office buildings. *Data-Centric Engineering*, 6, e3. doi:10.1017/dce.2024.57