

Article

Moving Sustainable Building Operations Toward Carbon Neutrality with Deep Reinforcement Learning: Balancing Energy Savings, Multi-Dimensional Indoor Comfort, and Carbon Permit Revenue Performance

Keonhee Cho ^{1,2}, Hyeonwoo Jang ², Guwon Yoon ² , Younghyun Baek ¹ , Myeong-in Choi ¹ 
and Sehyun Park ^{1,2,*}

¹ Department of Intelligent Energy and Industry, Chung-Ang University, Seoul 06974, Republic of Korea

² School of Electrical and Electronics Engineering, Chung-Ang University, Seoul 06974, Republic of Korea

* Correspondence: shpark@cau.ac.kr; Tel.: +82-2-822-5338

Abstract

The concept of carbon-neutral buildings encompasses not only carbon emission reductions but also sustainability. Building sustainability includes the physical durability of the structure, the health and safety of its tenants, and harmony with the surrounding environment. The achievement of these goals requires alignment among diverse stakeholders associated with buildings; however, such alignment is limited by economic (cost), environmental (global warming), and social (institutions and policies) factors. This study proposes an operation model that integrates buildings, the carbon permit market, and deep reinforcement learning (DRL) to address these limitations. The DRL model reduces energy consumption while maintaining indoor comfort, generates carbon permits equivalent to the amount of energy saved, and creates a new revenue stream by selling them. To achieve more precise comfort management, the model incorporates a policy that combines predicted mean vote (PMV) and Humidex. In the context of a privately owned commercial office building, the DRL model achieved indoor comfort levels of 98.51% for PMV and 97.22% for Humidex, while reducing energy consumption by 34,376 kWh, lowering carbon emissions by 26,607 kgCO₂eq, and generating USD 176 in carbon permit revenue. These results translated into a total reduction in operating costs of 7.5%, amounting to USD 2951. Consequently, the proposed approach provides cost reductions for building owners, comfort for tenants, efficiency for managers, and carbon emission reductions that contribute to carbon neutrality.

Keywords: BEMS; deep reinforce learning; PMV; humidex; carbon permits; sustainable building; carbon neutrality



Academic Editor: Biao Hu

Received: 2 October 2025

Revised: 18 November 2025

Accepted: 24 November 2025

Published: 28 November 2025

Citation: Cho, K.; Jang, H.; Yoon, G.; Baek, Y.; Choi, M.-i.; Park, S. Moving Sustainable Building Operations Toward Carbon Neutrality with Deep Reinforcement Learning: Balancing Energy Savings, Multi-Dimensional Indoor Comfort, and Carbon Permit Revenue Performance. *Buildings* **2025**, *15*, 4334. <https://doi.org/10.3390/buildings15234334>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

1.1. Background

Achieving carbon neutrality is now imperative, not optional, for addressing the climate emergency. Green technology is a critical driver of sustainable growth through innovation, particularly in the industrial sector.

In the building sector, energy demand accounted for 34% of global energy consumption in 2022, which makes it one of the largest energy-consuming sectors. Buildings were also responsible for 37% of global greenhouse gas emissions (both direct and indirect)

in the same year [1]. Sector-wide CO₂ emissions reached an all-time high of 10 GtCO₂ in 2022, largely due to rising electricity consumption during operation, which drove indirect emissions from a life-cycle perspective [1]. This trend is attributed to the growing carbon-neutral initiatives in the building sector [2–5]. While direct emissions decreased to 3 GtCO₂ due to the greening of passive design elements—such as building envelopes, shapes, and layouts—indirect emissions increased to 6.8 GtCO₂ because of higher electricity use by active systems, including heating, ventilation, and air conditioning (HVAC), hot water, lighting, and other services. These findings show that environmentally friendly architectural design and low-carbon construction materials alone are insufficient. They highlight the need for strategies that extend beyond eco-friendly design and low-carbon materials, toward decarbonizing active systems within buildings.

The major energy uses in active systems include heating and cooling, building services (e.g., elevators), and appliances and office equipment. These demands are fundamentally driven by human activity inside buildings, which aims to support daily functions, maintain comfort, and provide indoor environments that remain cool in hot weather and warm in cold conditions [6]. This is a common objective shared by different types of buildings, including residential and commercial buildings, despite their different forms and uses. However, the entity responsible for deciding whether to operate active systems varies between building types. Energy-use behaviors significantly influence the building's environmental conditions, quality of services, and overall energy consumption [7]. A variety of energy-use behaviors with complex drivers exist, which primarily originate from the occupants. In residential buildings, occupants act as active agents who directly interact with and influence the indoor environment [6]. However, in commercial buildings, active decision-making is primarily limited to facility managers, while occupants play a passive role by only providing feedback on indoor environmental conditions. Moreover, residential buildings are centered around individuals, whereas commercial buildings are group-oriented. This highlights that approaches to energy reduction should differ depending on the building's intended use [8]. In residential buildings, efficiency strategies are developed based on occupants' movement and behavioral patterns [9], whereas in commercial buildings, operational strategies are primarily established through scheduling and controlling energy-consuming systems [10]. This study focuses on commercial buildings.

Commercial buildings are complex entities where multiple stakeholders (owners, tenants/occupants, and facility managers) coexist and interact. They are obligated to provide and maintain comfortable indoor environments for occupants, which inevitably requires energy consumption. At the same time, they are expected to optimize energy use, ensure system safety, and deliver both long-term sustainability and economic efficiency for owners.

This study proposes a new approach to building operations by focusing on people as the primary drivers of carbon emissions from active systems. It considers all stakeholders within the building and seeks to develop an inclusive operation and management strategy. The proposed model aims to support a carbon-neutral society by delivering energy savings for owners, generating revenue through integration with the carbon market, and maintaining comfort and livability for occupants without disrupting their daily activities.

1.2. Limitations of the Sustainability Perspective of Buildings and People

Office buildings are shared spaces in which three primary stakeholder groups coexist: owners, who hold property rights; managers, who operate and maintain the building; and tenants, who live and work within it. Building operation and management are primarily planned around tenants, particularly those who reside and work inside the building [11–13]. Building energy management systems (BEMS) focus on minimizing tenant discomfort and

preventing energy-saving measures from leading to complaints [14]. Tenant discomfort refers to psychological and physical stress experienced by occupants [15], and higher levels of discomfort lead to lower satisfaction with the building environment. Long-term discomfort can lead to health issues and psychological distress, ultimately leading to tenants leaving the building through absenteeism, early departure, resignation, or termination of the lease [15–18]. Therefore, maintaining indoor comfort and efficient energy use is essential to preserving building value and ensuring long-term occupancy, thereby enabling the coexistence of people and buildings. However, maintaining this balance poses challenges across economic, environmental, and social dimensions [19].

First, the economic benefits of building operations face clear limits. From the owner's perspective, lower operating and maintenance costs translate into greater economic profit, but for tenants, this often results in increased discomfort. Active systems such as HVAC, electricity, and water usage are key technologies for maintaining a balance between owners and tenants to ensure comfortable and healthy indoor environments [20,21]. Facility managers, who serve as intermediaries between owners and tenants, must balance comfort with operational costs [22]. They are responsible for managing active systems, addressing complaints, and implementing the owner's operational strategies. During operation, minimizing discomfort inevitably increases energy consumption, which in turn raises costs for both owners and managers. These additional costs are often passed on to tenants through higher rent, leading to greater financial burdens for all stakeholders. This highlights the need for a cost-effective connection that maintains balance between the system and its users.

Second, efforts to reduce carbon emissions in buildings are constrained by practical factors. The primary source of these emissions is energy consumption [23,24]. Carbon reduction therefore largely depends on controlling energy use; however, such control is limited by spatial, temporal, and regional conditions. Temporally, energy consumption and potential reduction points are determined by tenant occupancy patterns, as active systems must operate during working hours to maintain an acceptable level of indoor comfort. Spatially, while securing and utilizing low-carbon energy sources such as renewables is highly effective in reducing emissions [25], most urban buildings lack the infrastructure to generate green energy on-site. From the owner's perspective, the motivation to invest in renewable systems for long-term environmental benefits and alignment with national carbon-neutral policies is often weak, as such investments raise short-term operational costs and are difficult to justify without clear economic returns [26]. Economic, environmental, and social conditions vary across regions, leading to significant differences in the challenges faced by the society in which a building is located. While economically stable and peaceful societies can pursue carbon-neutral strategies, others cannot actively reduce carbon emissions due to the external constraints of economic inflation, energy supply–demand imbalances, and rising energy prices [27,28]. These multifaceted constraints highlight the need for technical and operational innovation that goes beyond traditional building management and operation systems.

Third, the current structure of the carbon market, which is primarily driven by government-led, legally mandated carbon reduction quotas for corporations, presents limitations in promoting broad societal engagement with carbon neutrality [29]. Although a decade has passed since the Paris Climate Agreement of 2015, quantifiable progress toward the original medium- and long-term reduction targets remains insufficient [30–32]. While the need for innovation in climate awareness and response has been repeatedly emphasized, the technological, legal, and institutional foundations necessary for implementation are still incomplete [33]. The existing carbon market and emissions trading scheme involve participation exclusively at the national and corporate levels. In this system, cor-

porations are the essential market actors, while individual citizens are largely excluded. However, the path toward carbon neutrality requires expanded citizen participation and a market structure that embraces all stakeholders [34]. Carbon neutrality cannot be achieved through isolated efforts by governments, corporations, or individuals alone. Instead, it requires coordinated and systemic collaboration across all actors [35]. Government must demonstrate inclusive leadership, while corporations and citizens must translate national directives into concrete action. This collaboration should transcend political and economic boundaries, driving behavioral change, technological innovation, and environmentally responsible economic activity through mutual cooperation [36].

1.3. Literature Review

1.3.1. Sustainable Buildings for Carbon Neutrality

In our previous work, we defined various building types related to carbon neutrality and proposed a new approach to sustainable building design [37]. Humans are an essential component of sustainable buildings, alongside architecture, energy, and systems [38]. Existing approaches, however, have focused on energy efficiency, eco-friendly designs such as zero-energy buildings, and smart technologies like Building Information Modeling (BIM) and BEMS. These perspectives are largely system-centric, emphasizing owners, structural elements (e.g., exterior and interior walls), and building operation systems. Buildings are typically maintained for at least 30 years after construction. Over this lifespan, 80–90% of total energy consumption occurs during the use phase, 10–20% during manufacturing, and the smallest share during demolition [39–41]. This demonstrates that the operation stage dominates a building's carbon emissions, which makes its management during this phase critical to emission reduction [42]. While design and construction are important, greater emphasis should be placed on understanding the relationship between building operation mechanisms and carbon emissions during the maintenance phase.

Efforts to enhance energy efficiency and reduce carbon emissions in the building operation phase have evolved significantly—from early rule-based control and schedule-based operation to optimization-driven approaches such as model predictive control (MPC), and more recently to deep learning- and deep reinforcement learning (DRL)-based self-optimizing control. Technological advancements in building management now extend beyond basic energy reduction to ensuring occupant thermal comfort as a fundamental requirement. Aligned with carbon-neutral green technologies, building operation is undergoing a paradigm shift: moving beyond the scale of a single building toward integrated strategies that simultaneously achieve economic and environmental objectives.

Recent advances in DRL-enabled building operation have focused on optimizing multiple objectives, typically involving energy consumption, indoor comfort, and more recently, carbon emissions. These research trends are shown in Table 1. Multi-objective retrofit frameworks combining building performance simulation, surrogate modeling, and evolutionary optimization have demonstrated significant reductions in energy use and life-cycle carbon emissions. Similarly, DRL-based HVAC control methods—often implemented through EnergyPlus or co-simulation platforms—have achieved 20–40% improvements in energy efficiency while maintaining thermal comfort. Despite these advances, current studies exhibit several limitations. First, carbon is treated only as an emission metric, not as a tradable asset with economic value. None of the existing DRL- or MPC-based frameworks incorporate carbon permit accumulation, price volatility, or revenue optimization within the control objective. Second, comfort modeling remains highly simplified, relying almost exclusively on the PMV index. Humidity-sensitive indices such as Humidex, or dual-index comfort frameworks, have not been integrated into DRL training environments. Third, the vast majority of studies rely on simulation engines, whereas real multi-month

indoor sensor data are rarely used to inform state representation or validate DRL policies. Fourth, while SAC has become a dominant algorithm due to its stability in continuous control, studies do not explicitly address price-sensitive decision-making or carbon-aware operational strategies.

Table 1. Detailed comparison analysis with respect to existing BEMS incorporating DRL technology.

Ref.	Algorithm	Objectives	Reward	Control Level	Comfort Model	Carbon Permits (ETS)	Zone Units	Env.	Year
[43]	SAC	Minimize zone air temperature violations, supply air temperature, and energy cost	Energy, zone temperature violations, zone air temperature violations	AHU	ASHRAE 2006 [44], ASHRAE Guideline 36 [45]	x	5	EnergyPlus (ver.9.6.0)	2025
[46]	SSAC, SAC, DSAC	Minimize energy, maintain comfort	Energy, temp. deviation	AHU supply air temperature setpoint	Temperature-based comfort (not PMV)	x	1	EnergyPlus (accessed on 5 February 2025)	2025
[47]	QMIX (MADRL)	Reduce HVAC energy, maintain comfort	Energy penalty, temperature violation	Multi-zone VAV, ECW transmittance	Temperature band (not PMV)	x	6	EnergyPlus (ver.24.1.0), Python DRL Env. (ver.3.9–3.11)	2025
[48]	NSGA-III, BPNN surrogate (no DRL)	Minimize EDI, LCCO ₂ , Tdh	No RL reward (MOO)	Building envelope, HVAC setpoints	Discomfort hours (PMV)	x	Whole building	EnergyPlus based DesignBuilder (ver.6.1.0), MATLAB (ver.2023b)	2025
[49]	SAC	Minimize objective function (combined zone temperature and energy cost)	Objective function (combined zone temperature and energy cost)	HVAC	PMV (ASHRAE 55 [50])	x	4	Python	2024
[51]	DQN	Minimize zone air temperature violations	Energy, zone air temperature violations, zone humidity violations	HVAC	ANSI/ASHRAE standard 62.1 [52]	x	6	Energy Plus (ver.22.0)	2024
[53]	SAC	Minimize energy consumption, zone temperature violations	Energy, zone temperature violations	HVAC	PPD [50]	x	5	Energy Plus	2022
[53]	PPO	Maintain comfort, minimize electricity consumption	PMV-based comfort, energy cost	Heat pump, ventilation system	PMV [54]	x	1	Custom thermal model, Python PPO	2022

Existing DRL-HVAC studies optimize energy use and thermal comfort, but they do not consider economic returns through carbon permit trading, nor do they incorporate multi-index comfort evaluation tailored to real climatic and cultural conditions. The model proposed in this study addresses this gap by integrating (1) PMV–Humidex dual-index comfort control, (2) a carbon permit accumulation and revenue mechanism driven by real market prices, and (3) a real data-based DRL environment trained on seven months of measured indoor temperature, humidity, PMV, AHU, and MSB data.

1.3.2. Operational Scheduling in Office Buildings

Operational scheduling in buildings should minimize energy consumption and reduce costs by managing energy use efficiently. In commercial buildings, among active systems (mechanical and electrical systems installed within the structure, as opposed to passive elements such as the building envelope and materials), the air-handling unit (AHU) is the largest consumer of energy [55,56]. AHUs are designed for thermal management and air distribution, controlling indoor temperature and humidity through outdoor air intake and regulation of supply air temperature [57]. In most buildings, AHU control is therefore the primary means of ensuring a comfortable indoor environment for tenants. A widely adopted approach to achieve this is the optimal control of heating, ventilation, and air conditioning (HVAC) systems, which account for a significant share of building

energy use. Efficient system control helps reduce both energy consumption and operational expenses [58–60]. The most critical factor in creating operational schedules is maintaining tenant comfort while reducing energy and maintenance costs [61]. Occupant comfort varies according to regional, cultural, racial, gender, age, personality, and emotional differences, which influence thermal perception and preference [62–66]. As a result, HVAC operation strategies and even the definition of indoor environmental problems may differ across contexts. In some regions, occupants experience discomfort due to insufficient heating or cooling, while in others, excessive cooling or heating is the source of discomfort [67]. Although these situations appear to be opposites, both share the same fundamental premise—that a balance must be achieved between comfort and energy use. Understanding these cultural differences is important, as reducing overcooling or overheating alone can lead to substantial energy savings.

A review of prior studies shows that there are two main approaches to building operation scheduling: one centers on tenants as the basis for scheduling decisions [68–73], while the other focuses on temperature and energy consumption patterns within zones [74,75]. In particular, indoor temperature is consistently emphasized as a key factor in operation scheduling, as it is closely linked to thermal comfort [76–78].

However, most previous studies have addressed temperature and energy consumption while giving less attention to humidity when designing building operation schedules. Yet humidity, alongside temperature, is crucial in shaping human comfort. At high humidity, the same temperature feels warmer than it would at optimal levels. At low humidity, the perceived temperature can feel cooler than it actually is. This interaction demonstrates how temperature and humidity jointly influence thermal perception. Humidity is also directly tied to human perspiration: the human body regulates heat through sweat evaporation, but at high humidity, evaporation becomes less effective, often resulting in a feeling of stuffiness. Despite its importance, relatively few studies have addressed comfort by considering both temperature and humidity together. This gap in the literature motivated the development of our building operation model, which aims to use energy efficiently while ensuring comfort based on both temperature and humidity.

1.3.3. Carbon Allowances and the Carbon Market

The carbon pricing system is a policy instrument that applies market mechanisms to assign costs to carbon emitters [79]. The two most common approaches are the carbon tax and the emissions trading system (ETS). Both mechanisms share the same fundamental feature: the government sets an upper limit on allowable emissions. However, under a carbon tax, emitters pay a fixed charge per unit of excess emissions, whereas in an ETS, the emission volume is determined by the market price of emission allowances. Therefore, in a carbon tax system, the government acts as the price setter, while in an ETS, the market sets the price [80].

The ETS is a policy mechanism in which the government determines a national cap on total greenhouse gas emissions and allocates allowances to emitters, who are limited to companies as mandatory participants. Companies with excess allowances or shortages can trade them with other companies [81]. Carbon permits include both allowances and credits. In this paper, we use the term collectively without distinguishing the two. The price of carbon permits is determined by supply and demand rather than by government [82]. Carbon trading markets can be classified in various ways depending on the types of permits and trading methods [83]. However, this study does not classify trading markets; instead, it focuses on the general concept of a market in which carbon permits are exchanged. Carbon crediting, which involves the issuance of tradable certificates that verify reduced carbon emissions, is also excluded.

The carbon market is broadly divided into the compliance carbon market and the voluntary carbon market. In the compliance market, trading takes place on government-designated platforms and involves companies with assigned emission allowances. In contrast, the voluntary market is operated by private entities and allows participation by non-covered companies and individuals [84]. Applying a carbon permit trading mechanism to building operation models has the potential to offer greater economic efficiency than conventional BEMS [85]. The aim of this study is not to dismiss the compliance market, but to explore the integration between the compliance and voluntary markets. At present, only covered companies can manage their emissions within assigned allowances and trade carbon permits exclusively among themselves. This study seeks to expand that structure to enable carbon trading between covered and non-covered entities, including companies and individuals.

This study proposes a method to reduce building operational costs by integrating carbon permit trading mechanisms, with the broader aim of contributing to building sustainability. An office building was selected as a demonstration site, and environmental, energy, and facility operation data were collected. Based on these data, we developed a simulation environment and proposed an operational model that enables cost-efficient building management. The model not only reduces energy consumption but also incorporates carbon permit trading. The proposed approach demonstrates that, beyond traditional methods of reducing building energy costs through energy savings and renewable energy infrastructure, total operating costs can be further reduced by linking these efforts with the voluntary carbon market, thereby generating additional revenue from carbon permits. In addition, to improve indoor comfort management, we extend the conventional approach that relies on a single comfort index by introducing a method that employs two comfort indices. This dual-index approach allows for more refined and accurate indoor environmental control compared to using only one index.

As a result, this study advances previous DRL-based building operation frameworks by (1) integrating carbon permit trading into system control, (2) redefining operational profit to include both energy cost savings and carbon permit revenue, and (3) combining PMV and Humidex to better capture thermal comfort in tropical climates.

The remainder of the paper is structured as follows. Section 2 describes the demonstration building and outlines its key characteristics. Section 3 introduces the reinforcement learning-based sustainable building operation model, the study scope, and the fundamental equations for model validation. Section 4 explains the rationale for adopting the Soft Actor–Critic algorithm and details the reinforcement learning framework design. Section 5 discusses experimental results and analyzes the model’s performance. Finally, Section 6 concludes the paper by summarizing findings, discussing limitations and the potential for further advancement and expansion.

2. Methodology

This study expands the conventional focus on energy efficiency and indoor comfort by incorporating a carbon-neutral perspective. The objective is to develop a reinforcement learning model that not only reduces energy consumption while maintaining indoor comfort but also enables business actions driven by carbon reduction. This study adopts a reinforcement learning-based strategy, which has demonstrated strong performance in solving optimization problems in complex environments [86]. Based on previous studies that applied various reinforcement learning models to building systems, we selected the Soft Actor–Critic (SAC) algorithm, which has performed effectively in this domain [87–89]. In reinforcement learning, the agent must extensively explore the environment to discover an optimal control strategy. An agent’s exploration ability determines the quality of its

learning ability. Unlike existing reinforcement learning models (Trust Region Policy Optimization (TRPO) and Proximal Policy Optimization (PPO)), SAC can improve quality by using the concept of entropy. By incorporating the maximum entropy principle into the reward function, SAC promotes broader exploration, expands the search space for scheduling strategies, and enhances system robustness. The superior performance of SAC has been demonstrated by previous studies conducted by other researchers [89–92]. The primary focus of this study is to examine whether a reinforcement learning-based model can (i) reduce energy consumption, (ii) maintain indoor comfort under climatic constraints, and (iii) achieve carbon revenue through the proposed mechanism. The comparison of multiple reinforcement learning algorithms to identify the most suitable model for this environment is beyond the scope of this paper and will be pursued in future research.

The first objective of the proposed model is to minimize building operational costs while also minimizing tenant discomfort. Reference [93] identified a strong correlation between building operation and maintenance costs and energy scheduling strategies. Reference [56] demonstrated that efficient control of Air Handling Units (AHU) can achieve both energy savings and thermal comfort for tenants. Reference [58] analyzed how reinforcement learning models can optimize decision-making based on environmental feedback signals and effectively control various building actuators to achieve energy efficiency while maintaining comfort.

The second objective is to promote carbon reduction and to establish economic policies based on the results. Reference [94] confirmed that carbon emissions trading is both economically and environmentally effective.

This study does not employ simulation tools such as EnergyPlus, which are commonly used in research on building HVAC systems and thermodynamics. Instead, we constructed a Python-based simulation environment using data collected through Energy IoT (Internet of Things) systems. Although this approach relies on a smaller dataset compared to EnergyPlus, it has the advantage of replicating conditions that closely resemble those of an actual building.

The research process consists of four main stages. First, data from the target building were collected and analyzed. Second, a simulation platform of the building was developed based on the data analysis. Third, a reinforcement learning-based model was designed and trained to minimize operational costs. Finally, the proposed model was validated by comparing performance before and after applying the deep reinforcement learning algorithm. Python is used for two main purposes: first, to build a Soft Actor–Critic model, a type of deep reinforcement learning algorithm, and second, to construct a virtual environment similar to a building. The Python-based model and environment receive a training dataset and temperature/humidity change rate data as input. Based on the input data, the agent interacts with the built environment. As a result of this interaction, the agent receives a reward and gradually converges to output an optimal policy.

2.1. Data Collection

The building selected for the empirical study is an office tower located in Malaysia with a typical administrative work environment. It has 28 floors, of which 20, including the lobby, are used as office spaces, and 8 are allocated for parking, as shown in Figure 1.

The building is equipped with 60 AHUs to maintain indoor comfort. One AHU is installed on each side (left and right) of the 1st, 2nd, and 3rd floors. From the 4th floor upward, each floor has two AHUs on the left side and one on the right. For our case study, data were collected on the building's energy consumption, indoor and outdoor temperature and humidity, fine dust concentration, solar radiation, and AHU operational status.

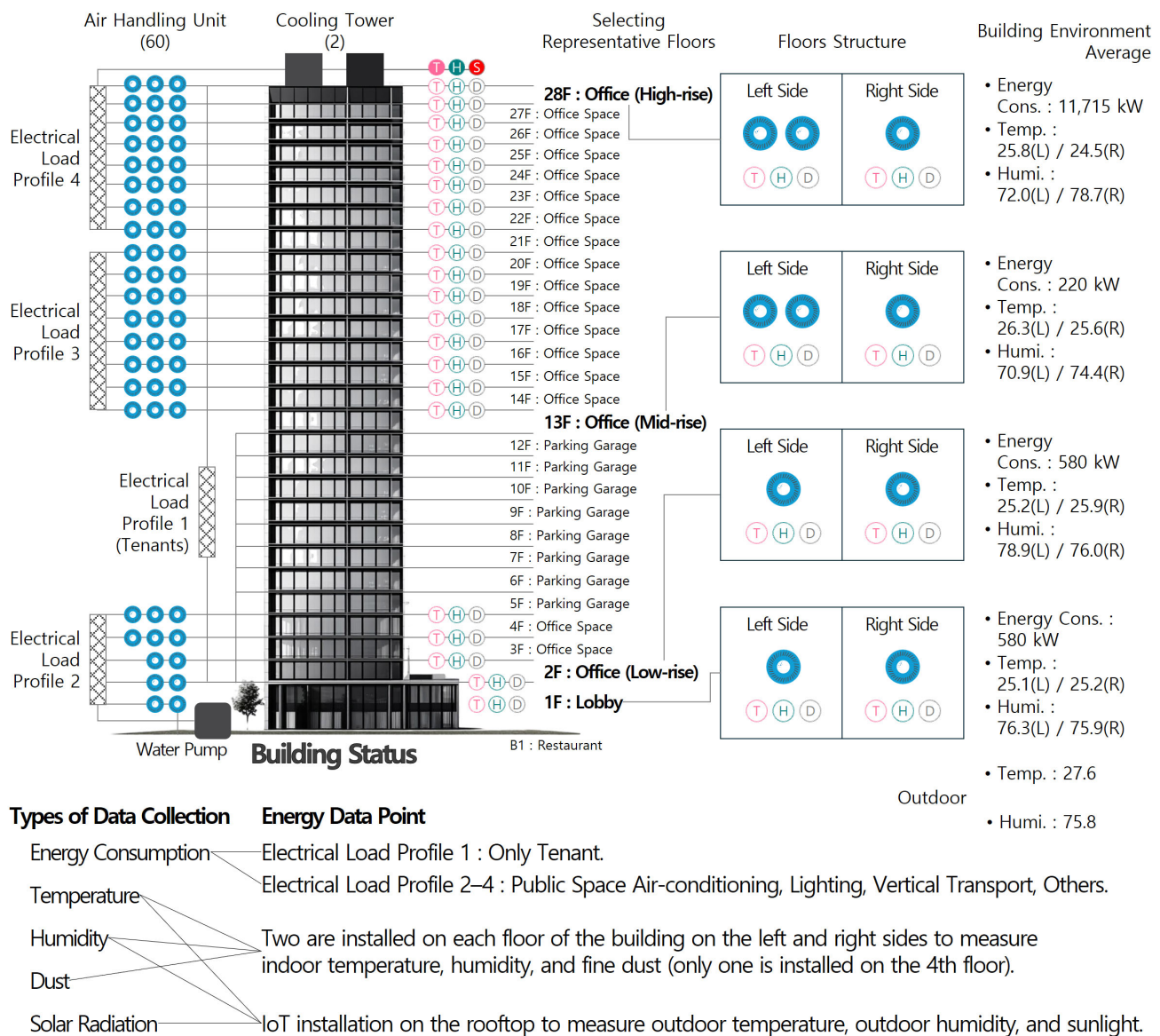


Figure 1. Building structure and data collection points of the case study.

In Figure 1, Electrical Load Profile 1 represents the tenant-exclusive load corresponding to the total electrical consumption by tenants. Electrical Load Profile 2 measures combined energy consumption from AHUs on Basement Level 1 to the 4th floor and water pumps. Electrical Load Profile 3 records energy consumption from AHUs on floors 14 to 20, while Electrical Load Profile 4 measures the combined energy consumption from AHUs on floors 21 to 28, along with two cooling towers and one cooling pump.

For deep reinforcement learning, four representative floors were selected. The 1st floor serves as the lobby with the main entrance and primary amenities. The 2nd floor represents the lower section, the 13th floor the middle section, and the 28th floor the upper section of the building. Energy consumption patterns for the lower, middle, and upper sections were analyzed using Electrical Load Profiles 2, 3, and 4, respectively. Since the building is used for office purposes, all four floors follow a common schedule from Monday to Friday, aligned with standard working hours.

Data were collected from the testbed over a nine-month period, from 1 July 2024 to 31 March 2025, for use in the deep reinforcement learning model. The details are summarized in Table 2.

Table 2. Data collected from the office building between 1 July 2024 and 31 March 2025.

Data Type	Data Collection Cycle	Unit	Data Utilization
Energy consumption	5 min	kW	Electrical energy data were used to analyze patterns and demand in building operation and management. These patterns are key to identifying the building's characteristics and daily usage profiles. In addition, the energy consumption attributable solely to AHU operation from the aggregate energy dataset.
Energy cost	Month	MYR/USD	Energy cost was calculated by applying the Malaysian electricity tariff per kWh and converting it to USD using the exchange rate. It was then used as a component of the reward function during the training of the DRL model.
AHU ON/OFF	5 min	-	AHU ON/OFF data were used to analyze changes in indoor temperature and humidity according to the AHU's operational status and duration. Based on this analysis, the results were used as environmental factors for monthly and hourly variations when the agent controls the AHU (one AHU 24.9 kW).
Indoor temp.	5 min	°C	Indoor environmental data served as a reference for assessing tenant discomfort. They determined whether indoor temperature and humidity become excessive when reducing energy consumption to obtain carbon permits, or whether further energy savings could be achieved. Zone-specific characteristics were also used to define comfort ranges.
Indoor humidity	5 min	%	
Outdoor temp.	5 min	°C	Outdoor temperature and humidity did not directly affect the agent's actions but were included as observation variables.
Outdoor humidity	5 min	%	
Carbon permit price	Day	KRW/USD	Carbon permit price data were collected in KRW and converted to USD using the applicable exchange rate, which also serves as a reference factor for CP sales.
Carbon emission factor	year	CO ₂ eq	The carbon emission factor was applied to calculate actual emissions based on measured electrical energy and to estimate virtual emissions generated by the DRL model (factor: 0.774).

The dataset used in this study was collected through the Internet of Things (IoT) and consists of heterogeneous data sources. Data from 10 AHUs, 4 MSB (Main Switch Board) meters, 8 indoor temperature sensors, 8 indoor humidity sensors, 1 outdoor temperature sensor, 1 outdoor humidity sensor, and 4 carbon permit price indices were integrated into a unified dataset through preprocessing. Afterward, at every 5 min interval, predicted PMV and Humidex values for the eight indoor zones were calculated and included, and energy consumption values measured in kW were converted to kWh and appended to the dataset. These additional variables were included to accelerate training and to support preliminary indoor environmental analysis. As a result, the final dataset for the RL environment contained 61 variables and 78,907 steps sampled at a 5 min resolution.

This resolution was preserved to retain the temporal characteristics of the real building environment and to prevent information loss; thus, the agent directly receives 5 min data as state inputs. However, the dataset used to reflect indoor environmental changes according to the agent's actions was aligned with the control interval. The change in environmental conditions that occurs every 5 min is different from the change that occurs over one control interval, because both continuous (natural thermal drift) and discontinuous changes (e.g., AHU switching, ventilation) coexist. In contrast, the control interval for the agent's actions was set to 15 min. This design reflects the operational characteristics of real HVAC systems, where indoor thermal responses are not immediate and AHU operations are typically scheduled in 10–30 min cycles [95,96]. This setting balances responsiveness and system stability while remaining consistent with building operation practices. Moreover, to reduce discrepancies between simulation and real operation, sub-hourly control intervals are recommended for model calibration [97].

Data collected through IoT systems are not always complete or of perfect quality. In our 9-month dataset, indoor and outdoor temperature and humidity measurements contained several missing intervals. Missing values can be imputed using statistical, machine learning, or deep learning approaches. In this study, we adopted a machine learning-based method, which is one of the most widely used techniques for handling missing data. Gradient boosting was selected due to its strong predictive capability and ability to model nonlinear relationships.

To impute missing indoor and outdoor temperature and humidity data, we employed a Light Gradient Boosting Machine (LightGBM) regression model. LightGBM offers up to 20 times faster training speed than traditional gradient boosting algorithms while efficiently handling large datasets within a short time [98]. Although gradient boosting models may be prone to overfitting when trained on small datasets, the integrated reinforcement learning dataset used in this study consists of 78,907 rows, which is considered sufficient for reliable model training.

The LightGBM model was implemented using the LGBMRegressor module from the LightGBM library in Python, following the official documentation and guidelines [99]. LightGBM supports several internal mechanisms, including histogram-based decision tree splitting, Gradient-based One-Side Sampling (GOSS), and Exclusive Feature Bundling (EFB). For model training, the input features included indoor and outdoor temperature and humidity, the operational status of 10 AHUs, four MSB power load measurements, and temporal features such as hour, month, minute, weekday, work/non-work hours, and time encoded with sine and cosine transformations. When a specific indoor or outdoor variable was the target of prediction, it was excluded from the feature set; otherwise, it was used as an input variable. The model was trained using rows without missing values and then used to predict and impute missing entries.

Since the primary goal of the model was to preserve temporal continuity and restore realistic patterns—rather than maximizing predictive accuracy—complex regularization or ensemble tuning methods were not applied. Instead, a relatively simple configuration with 100 boosting trees was adopted for computational efficiency and stable imputation performance. Model accuracy was evaluated using RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error), and the resulting error levels for each zone were within acceptable range, as shown in Table 3.

Table 3. LightGBM imputation performance for indoor/outdoor temp and humidity.

Category	Spaces	RMSE	MAE
Indoor Temperature	1F Left	0.257	0.175
	1F Right	0.15	0.106
	2F Left	0.209	0.146
	2F Right	0.122	0.089
	13F Left	0.165	0.123
	13F Right	0.182	0.134
	28F Left	0.265	0.197
	28F Right	0.231	0.166
Indoor Humidity	1F Left	1.265	0.919
	1F Right	1.208	0.84
	2F Left	1.443	1.002
	2F Right	1.326	0.904
	13F Left	1.292	0.936
	13F Right	1.707	1.143
	28F Left	1.629	1.186
	28F Right	1.226	0.903
Outdoor	Temperature	0.426	0.324
	Humidity	2.16	1.653

2.2. Carbon Emissions Trading Market

The price of carbon permits is determined by market mechanisms. Volatility arises from multiple factors, including the supply and demand for CPs, economic conditions, policy changes, industrial trends, fluctuations in the energy market, and regulatory standards [100–102]. The volatility of carbon permit prices, which is influenced by various interrelated factors, is difficult to predict. Therefore, this study designed the model to sell CPs based on historical prices rather than forecasting future values.

Carbon permit prices serve as a benchmark for improving energy efficiency and reducing building operation and management costs. Here, the benchmark denotes the timing (time or day) when the agent decides to sell CPs and the level of energy reduction required to obtain them. In some cases, stabilization mechanisms are used to adjust prices or secure the allocation of supply [23], but these mechanisms are outside the scope of this study.

Malaysia does not currently operate a carbon emissions trading market. Therefore, in this study, the proposed building operation and management model is applied to the Korean carbon market to perform a comparative analysis of operational cost reduction. The Korea Emissions Trading Scheme (K-ETS), implemented by South Korea in January 2015, is a market-based mitigation system in which the government allocates greenhouse gas emission allowances (carbon permits) to companies. It is the first nationwide emissions trading system in East Asia. The K-ETS covers the highest share of jurisdictional emissions among ETSs worldwide, including 89% of total GHG emissions. It covers the power, industry, buildings, transport, maritime, domestic aviation, and waste sectors, with forestry excluded [103]. Carbon permits data, including closing, opening, high, and low prices, were collected from the KRX ETS (Korea Exchange Emissions Trading System) market information platform.

On the K-ETS, if a company's actual emissions exceed its allocated amount, it must purchase additional permits from other companies, whereas if its emissions are below the allowance, it may sell the surplus. Emission allowances are allocated in two ways: through free allocation, in which the government provides allowances to companies at no cost, and through paid allocation, in which companies purchase allowances via auctions. In the auction system for paid allocation, a minimum bid volume of 1000 tons is required. In the K-ETS market, the carbon price is approximately USD 6–10 per ton, which is significantly lower than the European market, where prices are around USD 70–80 per ton.

2.3. Operational Status of the Target Building

We investigated the operational status of the target building through interviews with the building's manager and owner. In the target building, the operation of each floor's AHU was manually controlled by facility managers rather than automated by a BEMS or BAS system. The AHUs are part of a legacy system, requiring the operator to switch each unit ON and OFF manually according to the tenants' working hours. Starting from 07:00 until 18:30, the AHUs are sequentially activated floor by floor to match occupant schedules. Since the building uses a Water-Cooled Package Unit to supply air conditioning to all areas, there is no separate chiller controller. Indoor comfort management is maintained by keeping the AHUs ON throughout working hours. If tenants complain during weekdays, the set temperature is temporarily lowered to resolve the problem.

Scheduling-based AHU control in a demonstration building is defined as follows:

- The 60 AHUs on each floor are sequentially turned ON by the manager every day at 07:30 and OFF at 18:30.
- The AHUs are not controlled on weekends and public holidays, as they are closed.

2.3.1. Zone-Specific Comfort Range

PMV is the most widely used model for assessing thermal comfort. However, under certain environmental conditions, it may overestimate or underestimate actual comfort [104]. In such cases, adjustments are made to reflect the prevailing environmental conditions [105–107]. Although a standard comfort range is recommended for PMV, it can vary depending on environmental and regional characteristics. In this study, the range was adapted to reflect the hot and humid climatic conditions of Malaysia. To complement PMV, which is highly sensitive to temperature, the Humidex index was also applied to better account for humidity-related discomfort.

In the first step, the average PMV and Humidex during working hours were analyzed, which led to adjustments of the comfort range from the recommended boundaries to newly defined ones. Figure 2a shows the average PMV index of the target building from 1 July 2024 to 31 March 2025, with observed values between 0.8 and 0.2. Although the recommended range for Humidex is 20 to 29, Figure 2b shows that the average values during the same period ranged from 28 to 29 at the minimum to 33 to 34 at the maximum. Using the observed PMV and Humidex ranges as references, the comfort range applied in this study was defined as -1.0 to 0.2 for PMV and 27 to 33 for Humidex.

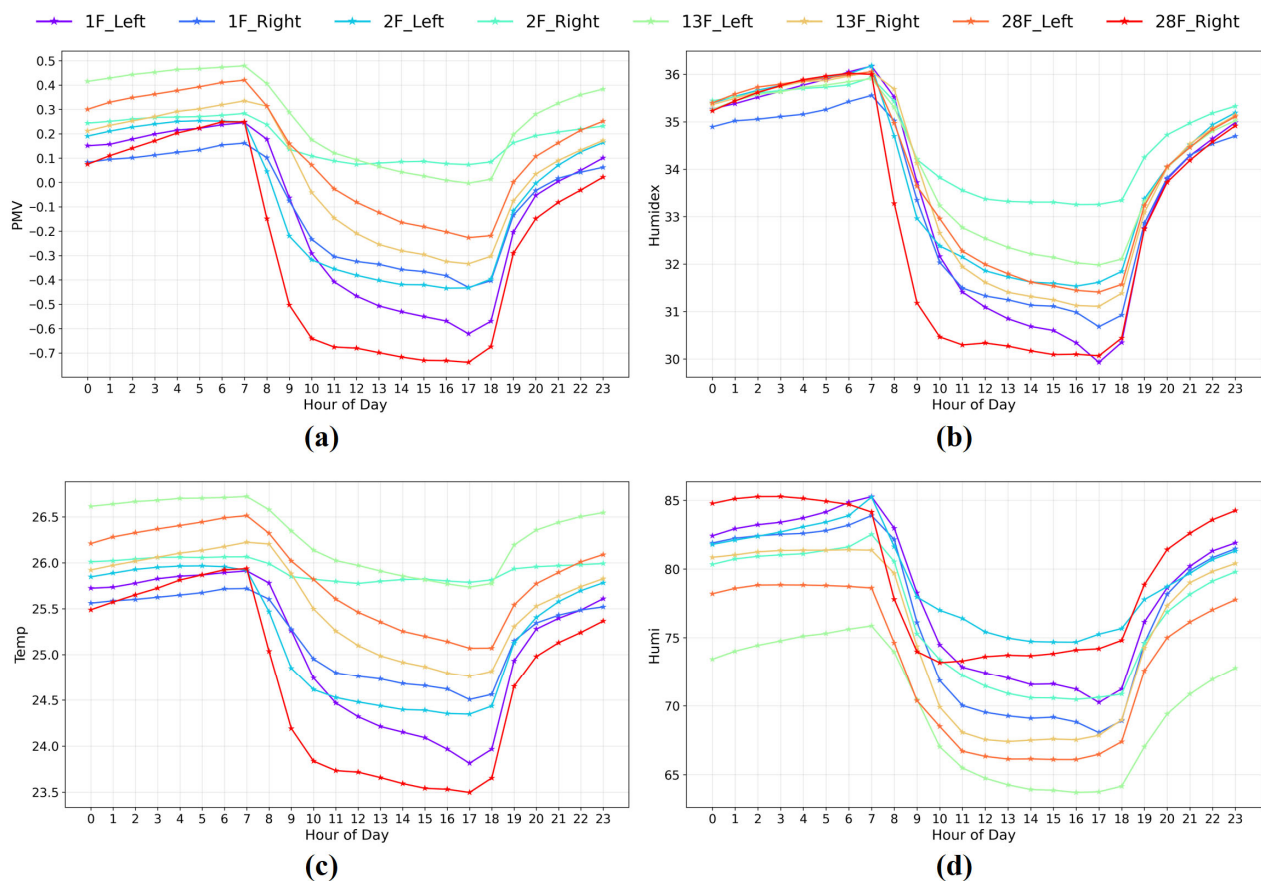


Figure 2. (a) Average indoor PMV of eight spaces by period from July 2024 to March 2025. (b) Average indoor Humidex of eight spaces by period from July 2024 to March 2025. (c) Average indoor temperature of eight spaces by period from July 2024 to March 2025. (d) Average indoor humidity of eight spaces by period from July 2024 to March 2025.

Figure 2 illustrates the cultural differences in indoor environment control in the target building. The PMV is maintained at levels lower than the comfort range recommended by ISO 7730 [108] and ASHRAE Standard 55 [50], while the Humidex is managed at

levels higher than the commonly recommended range. This indicates a deviation from international standards due to climate-specific office environments and cultural differences in clothing. Accordingly, we defined a customized comfort range that reflects these regional and cultural characteristics.

These values were extracted based on the 80th percentile range of observed data. This approach was adopted because mean values included extreme outliers, and applying the percentile range reduced distortions in the defined comfort boundaries. Details on setting the comfort threshold are described in Figure A1.

2.3.2. AHU Energy Consumption

In 2023, the total energy consumption of the target building was approximately 80,000 kWh per month. Analysis of energy usage showed that the building owner accounted for 82% of the consumption, while tenants accounted for 18%. Of the owner's share, approximately 90% was used by AHU systems in common areas, which confirms them as the primary source of energy consumption. Air-conditioning was found to be the most urgent area to require energy efficiency improvements in the testbed building. Equation (1) illustrates how AHU energy consumption was calculated from the collected energy data. Because the data contained mixed loads, the portion attributable to AHUs had to be inferred. Here, E_t^m denotes total energy consumption measured in the owner-managed areas, E_t^{load} is the baseline load observed on weekends when all AHUs were OFF, and AHU_{rate} is the ratio of energy attributed to AHU operation within E_t^m . Based on Equation (1), the total AHU energy consumption E_T^{AHU} over the period T was calculated and used to derive energy savings and carbon emission reductions.

$$E_T^{\text{AHU}} = \sum_{t=1}^T (E_t^m - E_t^{\text{load}}) \times \text{AHU}_{\text{rate}}. \quad (1)$$

2.4. Thermal and Moisture Control Capabilities of AHU Operation

The target building exhibited different temperature and humidity maintenance ranges across zones, with variation patterns also differing by zone. Accordingly, two approaches can be considered: applying individualized temperature and humidity variation rates for each zone, or applying monthly, time-of-day average variations uniformly across all zones. In this study, all AHUs were assumed to have identical mechanical specifications and therefore to provide the same cooling and dehumidification performance. Here, identical performance refers only to their mechanical capabilities. However, the actual thermal response of each zone differs depending on factors such as floor level, cardinal orientation (east, west, south, north), floor area, and occupant density. All these effects are implicitly reflected in the measured indoor temperature and humidity data, which were collected every 5 min (Section 2.3 (Operational Status of the Target Building) and Interpretation through Attachment Figure A1). However, it should be noted that these data do not include a sophisticated thermodynamic model. The applicability of a more precise exchange model, reflecting spatial characteristics, should be recognized and supplemented with additional thermodynamics.

Variations in temperature and humidity were derived from this measured data and calculated at 15 min intervals by month and time of day, to match the agent's control cycle. Equations (2) and (3) present the calculation formulas. Variations in temperature and humidity were calculated at 15 min intervals by month and time of day, aligned with the agent's control cycle. Equations (2) and (3) present the formulas.

$$\Delta T_{\text{avg}}(y, m, h, mn) = \frac{1}{Z} \sum_{Z=1}^Z (T_Z(t + 15mn) - T_Z(t)), \quad (2)$$

$$\Delta H_{\text{avg}}(y, m, h, mn) = \frac{1}{Z} \sum_{Z=1}^Z (H_Z(t + 15mn) - H_Z(t)). \quad (3)$$

The average variations in temperature T_{avg} and humidity H_{avg} at a given time (y, m, h, mn) are obtained from the differences between the temperature $T_Z(t)$ and humidity $H_Z(t)$ of zone Z at time t , and the corresponding values $T_Z(t + 15mn)$ and humidity $H_Z(t + 15mn)$ after 15 min. This rate of change is discussed in detail in Section 4.5.2.

2.5. Interrelationships Among Indoor Comfort Indices

Figure 3 illustrates the correlations among the AHUs controlled by the agent, indoor and outdoor temperature and humidity, the PMV index representing thermal comfort, and the enthalpy and Humidex indices related to humidity. The indices in Figure 3 were derived from AHU ON/OFF status, indoor temperature, and indoor humidity data collected from the target building. Enthalpy refers to the amount of heat contained in air at a given temperature, with higher values indicating greater heat content in moist air. Based on previous findings that it is sensitive to humidity, enthalpy was included as a comparative indicator [109].

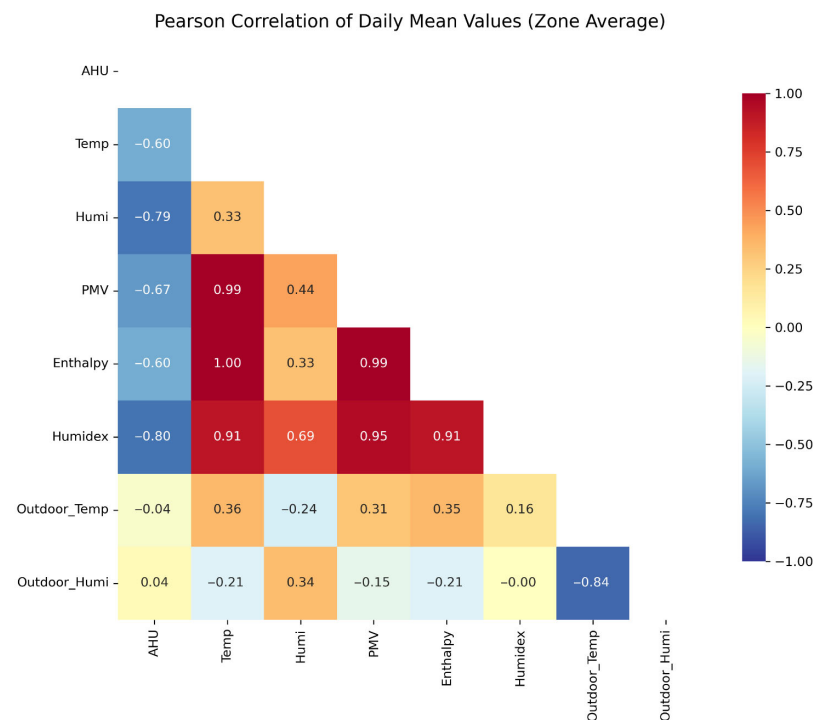


Figure 3. Correlation between temperature and humidity according to PMV and Humidex.

It was observed that indoor temperature and the related indices decreased concurrently with AHU operation. In terms of temperature, strong correlations were found with both the enthalpy and PMV indices, whereas humidity exhibited the strongest correlation with the Humidex index. While [109] reported that humidity showed the strongest relationship with enthalpy, in this study, humidity and Humidex demonstrated the highest correlation, with a coefficient of -0.8 . Therefore, PMV and Humidex were employed to predict thermal comfort in relation to temperature and humidity.

The formula for calculating PMV is detailed in Appendix C. As seen in Equations (A2)–(A12) in Appendix C, temperature was used directly as an input variable. However, relative humidity was not used directly but was first converted to water vapor pressure. The indirect humidity input shows a strong correlation of 0.99 between PMV and temperature, but a slightly less strong correlation of 0.44 with humidity. These

results indicate that PMV can represent comfort well in terms of indoor temperature, but the results for humidity may be somewhat disappointing. Humidex uses temperature and dew point temperature, as shown in Equations (A13)–(A15) in Appendix C. This is used to express the discomfort caused by excessive humidity [110]. Humidex also uses water vapor pressure but takes a humidity correction via dew point temperature. The relationship between relative humidity and Humidex is higher than the correlation with PMV, because the Humidex index can be calculated based on temperature and relative humidity. PMV and Humidex show a high correlation of 0.91, indicating that the rise and fall of the two indices are the same in expressing indoor comfort. An operating strategy that satisfies both indices suggests that it can satisfy both hot and humid discomfort.

3. Main Concept and Environment Modeling

3.1. The Proposed Framework for Sustainable Buildings

In this study, we propose a building operation strategy that ensures mutual sustainability among buildings, tenants, and society, while simultaneously achieving both economic and environmental benefits. Figure 4 illustrates the framework for achieving sustainability among stakeholders within a building. The framework addresses carbon neutrality in sustainable buildings from three key perspectives, as previously discussed. The owner establishes an energy-saving operational policy that enables environmentally friendly and cost-effective building management. By reducing energy consumption, the owner also secures carbon permits and develops a new profit-oriented policy for building operations. The building manager achieves both environmentally friendly and cost-efficient operations through deep reinforcement learning-based building management. Tenants can experience the same level of comfort as before the application of reinforcement learning, which reduces the likelihood of relocating due to discomfort and minimizes associated costs. As a result, both direct and indirect carbon emissions from building operations (as well as additional emissions linked to tenant relocation) are reduced, which contributes to a carbon-neutral society. The harmony among these three perspectives represents a meta-value that serves as a higher-order principle for addressing the fundamental challenges of building a sustainable carbon-neutral society.

3.2. Definition and Scope of Building Operating Costs

In this study, building operating costs refer to the energy expenses incurred during the use phase of a building, from completion through to reconstruction or decommissioning. In general, operating costs include both energy costs and maintenance costs, such as repairs, replacements, and inspections necessary to ensure comfort and safety. However, this study focuses exclusively on energy costs and excludes other management-related expenses. Building management costs vary significantly depending on factors such as the owner's priorities, the building's size, and its structural characteristics. Because of this high variability and limited generalizability, the analysis is restricted to energy costs, which exhibit relatively lower volatility. Energy costs in this context refer to charges arising from electricity consumption, including power used for electrical loads, lighting, elevators, ventilation, and air conditioning.

The energy billing structure of the case study building is divided between the building owner and the tenants. The costs charged to the owner correspond to energy consumption for common areas and shared services, such as AHUs, lighting, and elevators. In contrast, the costs charged to tenants are associated with energy use within leased spaces for business operations or personal use (e.g., PCs, fans, and other electrical appliances). Accordingly, both the owner and tenants pay their respective portions of the energy expenses. In this

study, only the energy costs borne by the building owner are considered, and the objective is to reduce energy consumption related to AHU operation.

Sustainable Initiatives for Achieving Carbon Neutrality : Human-Building-Society Meta-Values

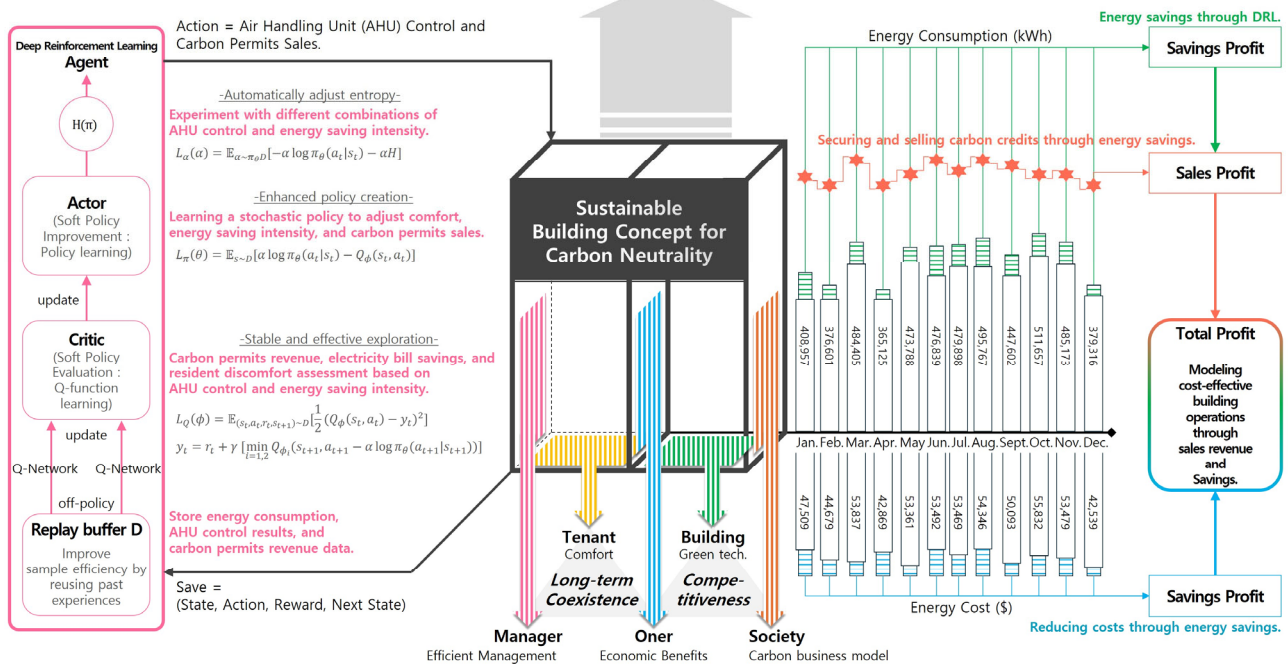


Figure 4. Sustainable building intelligent framework for a carbon-neutral society.

The building selected for this study is a general office high rise building located in Malaysia. The Malaysian electricity tariff system consists of several components: base electricity usage charges, maximum demand charges, additional surcharges related to fuel and power generation costs, and contributions to the Renewable Energy Fund. The energy cost associated with electricity usage is calculated from overall consumption and peak demand charges. In this study, country-specific taxes and additional fees are excluded, and the focus is placed solely on charges directly related to electricity consumption. The energy cost based on electricity use is calculated using the following equations:

$$E_{\text{peak}} = \max(E_{\text{elec},t-n}, \dots, E_{\text{elec},t}), \quad (4)$$

$$C_{\text{total}} = \sum_{n=1}^{\min(n, \text{month})} (E_{\text{elec}} \times P_{\text{kWh}}) + E_{\text{peak}} \times P_{\text{peak}}, \quad (5)$$

where C_{total} denotes the total energy cost resulting from electricity consumption in the building over n days, E_{elec} represents the general electricity consumption during the n -day period, P_{kWh} is the contracted unit rate per kilowatt-hour (kWh), E_{peak} indicates the peak demand power during the n -day period, and P_{peak} refers to the contracted rate for peak demand power. The peak demand power E_{peak} is defined as the highest recorded power demand within a 30-day window (kW). In this study, the energy cost associated with peak demand power is excluded. The deep reinforcement learning model is not designed to predict monthly peak demand but instead focuses on maximizing the reward within each day.

The energy consumption E_{elec} refers to the electricity used for basic building loads, as well as for HVAC, cooling and heating, lighting, and elevator operations. This consumption directly represents the total energy cost required to operate the building, which can be

calculated using Equation (5). Since peak demand charges are excluded in this study, the total operating cost is estimated solely from general electricity consumption.

3.2.1. Energy Savings and Cost Reduction in Building Operations

The building energy savings E_{save} are calculated based on the actual measured electricity consumption E_{elec} and the power consumption of AHUs when operating p_{AHU} . The energy use is simulated through reinforcement learning-based control. The total building energy consumption, E_{elec} , refers to the overall electricity consumed by the building. Energy savings resulting from AHU control, E_{save} , can be evaluated by comparing the actual energy consumption calculated from AHU status information at 5 min intervals during working hours ($E_{\text{calc}}^{\text{AHU}}$) with the simulated energy consumption under the reinforcement learning model ($E_{\text{DRL}}^{\text{AHU}}$).

In addition, comparison with E_{elec} allows the assessment of energy savings at the whole-building level. These relationships are expressed in the following equations:

$$E_{\text{save}}(t) = E_{\text{calc}}^{\text{AHU}}(t) - E_{\text{DRL}}^{\text{AHU}}(t), \quad (6)$$

$$E_{\text{calc}}^{\text{AHU}}(t) = \text{on}_{\text{AHU}}(t) \times p_{\text{AHU}} \times h, \quad (7)$$

$$E_{\text{DRL}}^{\text{AHU}}(t) = \text{on}_{\text{DRL}}(t) \times p_{\text{AHU}} \times h. \quad (8)$$

3.2.2. Building Comfort Cost

The building comfort cost, $\text{Comf}_{\text{cost}}$, refers to the energy cost incurred by operating AHUs to maintain appropriate indoor conditions and provide comfort for tenants. This cost is one of the components used in the reward system, and the reinforcement learning agent is designed to minimize it. The comfort cost can be calculated by multiplying the simulated AHU energy consumption, $E_{\text{DRL}}^{\text{AHU}}$, by the electricity unit price, P_{kWh} , as follows:

$$\text{Comf}_{\text{cost}} = E_{\text{DRL}}^{\text{AHU}} \times P_{\text{kWh}}. \quad (9)$$

3.3. Definition and Scope of Carbon Emissions

In this study, carbon emissions are calculated from energy consumption. More specifically, emissions were estimated based on the amount of electricity used within the building. The amount of greenhouse gas (GHG) emissions resulting from electricity consumption can be calculated using Equation (10), while carbon dioxide (CO_2) emissions are determined using Equation (11). The emission factor for CO_2 equivalent (CO_2eq) applies the global warming potential (GWP) to three major greenhouse gases: CO_2 , CH_4 , and N_2O .

$$\text{GHG}_{\text{building}} = \sum (E_{\text{elec}} \times \text{CEF}_{\text{contury}} \times \text{GWP}), \quad (10)$$

$$\text{CE}_{\text{building}} = E_{\text{elec}} \times \text{CEF}_{\text{contury}} \times \text{GWP}, \quad (11)$$

where $\text{GHG}_{\text{building}}$ denotes the greenhouse gas emissions resulting from electricity consumption in the building (tCO_2eq), $\text{CE}_{\text{building}}$ denotes carbon emissions from electricity use (tCO_2), E_{elec} is the total electricity consumption, $\text{CEF}_{\text{contury}}$ represents the country-specific electricity emission factor, and GWP is the global warming potential. Electricity consumption E_{elec} is measured and collected through Energy IoT devices installed in the building. $\text{CEF}_{\text{contury}}$ is calculated using data officially published by the national authority. GWP is determined according to standards defined by the Intergovernmental Panel on Climate Change (IPCC). In this study, active elements are selected as control targets to ensure efficient energy consumption. Therefore, the emission factor for the consumption stage, based on indirect emissions, is applied.

3.4. Estimation of Building Carbon Emissions

In the building sector, the primary source of carbon emissions is energy consumption. Therefore, reducing energy use is an effective way to acquire carbon permits. However, to secure such permits, a clear baseline or criterion must be established. Under the current carbon permit trading system, entities are allocated emission allowances by the government. These designated companies must not exceed their assigned limits, and if their actual emissions are lower than the allowance, they can obtain carbon permits.

By contrast, general citizens or non-designated entities who are not part of the allowance system lack a predefined baseline. Thus, a new criterion is required for them to acquire carbon permits. In this study, two simulation scenarios are established to propose baseline criteria that enable non-designated users to earn and trade carbon permits, thereby generating revenue.

It should be noted that the deep reinforcement learning agent developed in this research does not control all energy-consuming systems within a building. Its scope is limited to the Air Handling Units (AHUs). Accordingly, the calculation of carbon emissions is confined to the portion of energy consumption attributable to AHU operations.

3.4.1. Building Carbon Emissions

According to the Energy Commission of Malaysia, the carbon emission factor for indirect emissions at the point of electricity consumption (Scope 2) is 0.774 tCO₂eq/MWh. The carbon emissions resulting from AHU operation can be calculated using Equation (12). In addition, total building-level emissions and the emissions under reinforcement learning control can be calculated by multiplying Equations (7) and (8) by the emission factor, respectively.

$$emission_{building}^{AHU}(t) = E_{save}(t) \times CEF_{country}. \quad (12)$$

In national greenhouse gas inventories, the carbon emission factor $CEF_{country}$ is calculated based on historical energy production and consumption data through the following steps:

1. Identification of emission sources;
2. Selection of a calculation methodology (Tier 1–3);
3. Collection of activity data and selection of corresponding emission factors;
4. Calculation and reporting of emissions.

This approach typically assumes a fixed energy mix when determining CEF values.

Recently, however, increasing attention has been given to dynamic carbon emission factors that vary across time scales (e.g., hourly, daily, seasonally) according to real-time grid conditions and the share of energy sources. Dynamic CEFs allow more accurate estimation and monitoring of actual CO₂ emissions resulting from electricity consumption. In this study, we focused on developing an operational management framework for buildings; therefore, a fixed national carbon emission factor reported by the Malaysian Energy Commission is adopted. The corresponding national electricity generation mix used for the CEF calculation consists of coal (731,154 TJ), natural gas (394,844 TJ), diesel (1381 TJ), and fuel oil (2636 TJ).

3.4.2. Citizen-Driven Voluntary Carbon Reduction Scenario

Carbon emission reduction in buildings can be estimated under two scenarios. The first scenario calculates emission reduction relative to the previous year. In this case, the reduction amount, denoted as $Carbon_{save}^{building}$, can be computed using Equation (13). This method calculates carbon emissions based on the previous year's energy consumption, $E_{y-1}^{building}$, on a monthly or daily basis, and then estimates the reduction, $\Delta E_y^{building}$, by

comparing it with the current year's energy consumption for the same period, as shown in Equation (14). To calculate energy savings under this scenario, annual energy data must be collected through utility bills or an energy management system (EMS). This scenario is suitable from a mid- to long-term perspective. To maximize revenue from carbon permits, greater energy savings than the previous year are required. If energy use decreases each year, the total consumption gradually decreases over time. As a result, energy is eventually used efficiently without unnecessary waste, which causes the building's initial energy consumption, E_0^{building} , to converge toward the minimum achievable level, $E_{\min}^{\text{building}}$, as shown in Equations (15) and (16). As energy consumption approaches $E_{\min}^{\text{building}}$, the building achieves maximum reduction in both operational costs and carbon emissions over its lifecycle.

$$\text{Carbon}_{\text{save}}^{\text{building}} = \Delta E_y^{\text{building}} * CEF_{\text{country}}, \quad (13)$$

$$\Delta E_y^{\text{building}} = E_{y-1}^{\text{building}} - E_y^{\text{building}}, \quad (14)$$

$$E_y^{\text{building}} = E_{\min}^{\text{building}} + \frac{(E_0^{\text{building}} - E_{\min}^{\text{building}})}{1 + y}, \quad (15)$$

$$\lim_{y \rightarrow \infty} E_y^{\text{building}} = E_{\min}^{\text{building}}. \quad (16)$$

The second scenario estimates the reduction based on the energy consumption prior to the application of reinforcement learning. This is a short-term scenario that compares carbon emissions from energy consumption before and after applying the reinforcement learning-based model, using a monthly or weekly basis. This approach enables a rapid evaluation of the model's performance and the analysis of its effectiveness. In this study, carbon reduction is calculated based on the difference between the carbon emissions after applying reinforcement learning, $\text{DRL}_{\text{emission}}^{\text{AHU}}$, and the baseline defined in Equation (17), as follows:

$$\text{Carbon}_{\text{save}}^{\text{AHU}}(t) = \text{control}_{\text{emission}}^{\text{AHU}}(t) - \text{DRL}_{\text{emission}}^{\text{AHU}}(t). \quad (17)$$

In accordance with current emissions trading scheme (ETS) policies, emission allowances in the compliance market are allocated based on the average emissions of the previous three years when a facility is designated as a covered entity. In the voluntary carbon market, carbon credits are issued for verified emission reductions achieved during the implementation period of a carbon reduction project.

Aligned with these policy principles, the baseline definitions used in this study are interpreted as follows:

- Scenario 1 (comparison with the previous year) corresponds to the compliance ETS approach, in which historical emissions—typically the reference year or the past three-year average—serve as the baseline for certifying emission reductions;
- Scenario 2 (comparison with pre-DRL operation) reflects the voluntary market approach, where the baseline is defined as the operational state prior to the application of a carbon reduction measure such as the DRL-based control strategy.

In this study, Scenario 2 was adopted. Prior to implementing the DRL-based control strategy, we collected baseline data on energy consumption and electricity costs under manually scheduled AHU operations by the building manager. These data serve as the reference baseline for evaluating the DRL model's effectiveness in reducing carbon emissions, energy consumption, and operating costs.

3.4.3. Hypothetical Carbon Trading Market Reflecting Real-World Policy Limits

Currently, most carbon permit markets operate in a top-down manner, where participation is limited to companies subject to government-mandated emission caps. In such systems, individual citizens typically cannot participate. For example, in South Korea, only companies emitting more than 125,000 tCO₂-eq per year, facilities emitting over 25,000 tCO₂-eq, and voluntarily registered entities are eligible. For individual citizens who own commercial buildings, government-imposed emission caps do not apply, and thus, no emission allowances are allocated. Therefore, this study acknowledges several limitations and defines a hypothetical carbon trading market accordingly:

- Ordinary citizens who are not subject to regulation and who are not participating in carbon reduction projects can participate in the carbon permit market;
- When energy consumption is reduced relative to a specified baseline, the corresponding reduction in carbon emissions grants the right to sell an equivalent amount of carbon permits. A certified platform exists to officially verify carbon emissions and accredit verified reductions;
- Citizens can freely participate in the carbon permit market, buying or selling permits without restrictions between the compliance and voluntary markets;
- These assumptions establish the premise that citizens who are not subject to regulatory obligations can acquire and trade carbon permits.

The price of carbon emissions is determined through either carbon taxation or emissions trading schemes (ETS) [111]. The fundamental difference between the two systems lies in how prices and emission levels are determined. Under an ETS, the government sets a cap on total emissions, and the price of emission allowances is determined by the market. In contrast, under a carbon tax system, the government sets a fixed tax rate on carbon, while the actual emission levels are determined by the economic activities of market participants. Therefore, the key distinction between these mechanisms is the trade-off between certainty and uncertainty in terms of carbon prices and emission quantities.

In Europe, these two policy instruments are applied to different sectors: the power and industrial sectors primarily operate under the ETS, while the building and transport sectors are regulated through carbon taxes [112]. However, beginning in 2027, the European Union will introduce a separate emissions trading system for the building and transport sectors—sectors that have traditionally not been covered by the existing ETS—replacing the carbon tax approach [113]. This policy shift reflects the fact that the economic value of carbon reduction differs across industries [112–114]. Similarly, in the voluntary carbon market, carbon credit prices vary significantly depending on project type, location, market conditions, and the quality of emission reductions. As shown in Table 4, the value of carbon reduction differs depending on the type of project [115].

Table 4. VCM (Voluntary Carbon Market) transaction volumes, values, and prices, by project category, 2023.

Category	Volume (MtCO ₂ e)	Value (USD)	Price (USD)
Forestry and Land Use	36.2	353.3 M	9.72
Renewable Energy	28.6	111.1 M	3.88
Chemical Processes/ Industrial Manufacturing	12.2	50.2 M	4.10
Household/ Community Devices	9.9	76.6 M	7.70

Table 4. *Cont.*

Category	Volume (MtCO ₂ e)	Value (USD)	Price (USD)
Energy Efficiency/ Fuel Switching	9.4	34.4 M	3.65
Agriculture	4.7	30.6 M	6.51
Waste Disposal	1.5	10.9 M	7.48
Transportation	-	-	-

It should be noted that the carbon trading mechanism applied in this study is modeled as a hypothetical scenario. In reality, individual building owners are not authorized to freely trade emission permits in compliance carbon markets, and voluntary markets are subject to strict certification and participation criteria. Furthermore, because the building sector is not included in the emissions trading sector, measuring the value of carbon emissions requires new research. The carbon emissions value of the buildings used in this study is limited to be equal to the value of carbon emissions currently traded in the emissions trading system market.

Therefore, the financial reward from carbon permit sales presented in this study should be interpreted as a theoretical estimation, rather than a directly realizable profit under current policy conditions.

4. Modeling of Building Operating Costs Minimization Based on Deep Reinforcement Learning Using Soft Actor–Critic

4.1. Design of a Reinforcement Learning Model for Optimizing Building Operational Profit

Reinforcement learning (RL) is a paradigm in which an agent interacts with an environment to maximize a reward $R(t)$ that corresponds to every state $s_t \in S$. In other words, everything is based on the reward $R(t)$. The agent interacts with the environment by optimizing its policy π to select the optimal action $a_t \in A$. Specifically, the agent takes an action s_t in state a_t , receives a reward r_t , and transitions to a new state s_{t+1} , continuously learning to maximize expected returns. Deep reinforcement learning (DRL) combines reinforcement learning with deep neural networks (DNNs) to identify optimal actions in complex environments such as real-world systems. The following section describes how the proposed building energy management system is modeled using the Actor–Critic framework.

The DRL model must continuously manage indoor comfort while also determining when to sell carbon permits. This constitutes a continuous action space problem rather than a discrete one. Therefore, managing the intensity of comfort control and adjusting carbon permit sales in a continuous action space requires a policy-based approach [116]. However, policy-only methods can result in unstable learning, which makes it necessary to evaluate rewards using a value function. Therefore, the Actor–Critic algorithm, which combines both policy-based and value-based methods, enables more stable learning [88]. In the Actor–Critic algorithm, the actor is implemented using a policy gradient method to determine actions, while the critic evaluates those actions using a Q-learning approach and provides feedback for policy updates. However, this combined method may suffer from limited exploration; if the policy converges too quickly during early learning, the agent may stop exploring before identifying the optimal policy. To address this limitation, we adopt the Soft Actor–Critic (SAC) algorithm, which enhances exploration performance.

SAC extends the Actor–Critic framework by incorporating maximum entropy reinforcement learning. Unlike conventional Actor–Critic methods that learn a deterministic policy, SAC optimizes a stochastic policy, which encourages exploration. It also employs a double Q-network architecture to mitigate overestimation bias and improve stability.

The SAC algorithm introduces an entropy reward term $H(\pi)$, to encourage diverse action selection and enhance exploration [117]. The entropy $H(\pi)$ measures policy uncertainty: preventing the reinforcement learning model from converging prematurely to suboptimal actions and guiding it toward discovering an optimal policy [117]. A high entropy value indicates that the policy exhibits uncertainty, which suggests that the agent is in an exploration-dominant learning phase where various strategies are tested. Conversely, a low entropy value implies that the policy is confident in specific actions and signifies a shift toward an exploitation-focused phase with reduced exploration and repeated actions [117]. As the entropy $H(\pi)$ increases, it indicates that the action probabilities are more evenly distributed, which suggests that the policy explores a broad range of actions. The SAC algorithm, based on maximum entropy reinforcement learning, learns the optimal policy as follows [90]:

$$J(\pi) = \sum_t \mathbb{E}_{(s_t, a_t) \sim p_\pi} [r(s_t, a_t) + aH(\pi(\cdot|s_t))]. \quad (18)$$

Here, $J(\pi)$ denotes the objective function for policy optimization, $\mathbb{E}$ represents the expected value used to calculate the average reward experienced by the agent, p_π is the state-action distribution sampled under the policy, $r(s_t, a_t)$ denotes the reward function at time, $aH(\pi(\cdot|s_t))$ represents the entropy bonus to encourage exploration, and a is the entropy coefficient that balances exploration and exploitation. Unlike on-policy learning (where data collected through interaction with the environment is discarded immediately after use), SAC adopts an off-policy learning approach. In off-policy learning, the behavior policy used to collect data differs from the target policy being optimized. This approach allows SAC to reuse collected data multiple times and to explore a wider range of actions. Consequently, SAC leverages both off-policy learning and entropy-based exploration within the maximum entropy reinforcement learning framework, thereby improving the stability of learning in complex environments.

The deep reinforcement learning problem can be formulated as a Markov Decision Process (MDP):

- **Agent:** The agent is the learner that explores various policies to discover the optimal strategy for the environment. In this study, the agent learns policies that both satisfy the building's basic indoor environmental requirements and reduce energy consumption. Furthermore, it learns the appropriate timing for selling carbon permits based on the energy saved. Through this process, the agent aims to maximize building operational profit while minimizing tenant discomfort.
- **State S :** The state represents variables that characterize the dynamic environment, including energy-related metrics, indoor and outdoor environmental conditions, carbon permit information, and time. The state space consists of 55 dimensions, including electricity consumption, AHU operational status, zone-specific indoor temperature, humidity, PMV, and Humidex across eight zones on four floors, outdoor temperature and humidity, energy savings, carbon emissions, carbon permit variables (price, average price, inventory, accumulation period, expected revenue), and time. In this study, the state space is defined as

$$S = \{E_{elec}, E_{info}, C_{info}, AHU_{state}, T_{in}, H_{in}, PMV_{zone}, Humidex_{zone}, T_{out}, H_{out}, t\}.$$

The potential redundancy among indoor temperature, humidity, PMV, and Humidex was addressed in Section 2.5 (Interrelationships Among Indoor Comfort Indices), where their mutual correlations were analyzed. Based on these results, all four variables were retained as state inputs because they capture different aspects of thermal comfort—temperature and humidity as physical measurements, and PMV and Humidex as perceptual comfort indices [118,119]. Additionally, all state variables were normalized to $[0, 1]$ using min-max scaling to ensure numerical stability for SAC.

- **Action A:** The agent's actions consist of a set of adjustable decisions that influence building operation. Specifically, the agent determines the ON/OFF status of 10 AHUs deployed across eight zones on four floors, evaluates whether to sell carbon permits acquired through energy savings, selects the quantity to sell, decides how many permits to retain, and controls the intensity of temperature and humidity adjustments between 08:00 and 10:00, a period deemed critical for tenant comfort. As a result, the action space consists of 21 dimensions and is defined as
 $A = \{\text{AHU control, carbon permit sale, sale quantity, permit retention, intensity of indoor environmental change}\}.$
- **Reward R:** The reward serves as a performance metric for the policy, and evaluates the outcome of each action in a given state. The agent aims to maximize building operational profit by reducing energy consumption, maintaining indoor comfort, and generating revenue through carbon permits, while minimizing tenant discomfort. Thus, the reward is defined as
 $R = \{\text{Comfort Cost, Indoor Discomfort, Carbon Permit Revenue}\}.$

4.2. Objective Function

In the previous subsection, the SAC algorithm was selected to maximize building operational profit while minimizing tenant discomfort, and its concept and training procedure were outlined. Here, we describe the model that applies SAC to cost-effectively optimize building operation strategies. This framework is structured as a multi-objective model, where J_{DRL1} minimizes tenant discomfort and J_{DRL2} maximizes economic efficiency through cost reduction and revenue generation.

Equation (19) represents the first objective, which is to minimize tenant discomfort in the indoor environment. Tenant discomfort is calculated and predicted using the predicted mean vote (PMV) and Humidex indicators. PMV is a thermal comfort index based on the human body's heat balance. It is calculated using six environmental and personal factors: air temperature, humidity, air velocity, mean radiant temperature, metabolic rate, and clothing insulation. The amount of water vapor in the air can be expressed in various ways. The Humidex is an index developed by Canadian meteorologists to assess thermal comfort by combining temperature and humidity [110] to represent the degree of hot and humid weather as perceived by an average person. The formula to calculate the humidity index generally includes dew point information, which distinguishes it from the heat index as it relates to the dew point rather than relative humidity [120]. Although Humidex is typically used to assess outdoor conditions, this study confirms that it can also be applied indoors to evaluate comfort [121]. As a result, DRL_1 is designed to ensure stable regulation within the ranges of temperature and relative humidity where tenants do not experience discomfort, based on both PMV and Humidex indicators.

$$J_{DRL1} = \min \left(\sum_{i=1}^N \left(Disc_i^{PMV} + Disc_i^{Humidex} \right) \right). \quad (19)$$

PMV is a thermal comfort model standardized by ISO 7730 that is widely used to evaluate thermal comfort in buildings [122]. It quantifies comfort by considering the body's thermoregulation mechanisms along with environmental factors such as temperature, humidity, and indoor air velocity. PMV is expressed on a seven-point scale from -3 to $+3$, where -3 indicates "cold", 0 indicates "neutral", and $+3$ indicates "hot". The closer the PMV is to 0 , the more comfortable the indoor environment is perceived to be. Since maintaining a constant PMV of 0 in all zones is impractical, a range between -0.5 and $+0.5$ is generally accepted as thermally comfortable. If the indoor temperature deviates significantly, the agent incurs a penalty due to increased discomfort. Through training,

the agent learns to minimize discomfort by adjusting its actions so that PMV converges as closely as possible to zero.

The comfort scale for the Humidex index defines values of 20–29 as comfortable, 30–39 as indicating varying degrees of discomfort, 40–45 as conditions where almost everyone feels uncomfortable, and values above 46 as conditions where most labor must be restricted [110]. When indoor humidity is excessively high or low, the agent receives a penalty. As a result, the agent learns to regulate the Humidex within the comfortable range defined by the scale.

The second objective of the agent is to reduce energy consumption by operating AHUs more efficiently, since they account for a large portion of total building energy use. The agent reduces energy consumption by appropriately controlling the AHUs, which account for a significant portion of the building's energy use. Based on reduced energy consumption, carbon permits can be acquired and sold to generate new revenue. The DRL model is designed to minimize the total building operating cost by reducing energy expenses and selling the carbon permits obtained from the corresponding energy savings, as represented in Equation (20).

$$J_{DRL2} = \max(-comf_{cost} + Rev_{carbon}). \quad (20)$$

Here, Rev_{carbon} denotes the revenue generated from carbon permit sales at time t , and $Cost_{save, t}$ denotes the electricity cost saved through energy reduction at time t . Rev_{carbon} can be maximized by selling carbon permits when the market price is high. $Cost_{save}$ increases as the amount of energy savings grows, which leads to a greater reduction in electricity expenses. As building energy consumption decreases, operational profit increases correspondingly.

4.3. Reward System

Reinforcement learning is based on rewards obtained through experience, with the goal of maximizing long-term returns. Therefore, the reward system is fundamentally oriented toward maximizing the total reward [123]. The reward R is designed to minimize tenant discomfort, maximize energy savings, and generate revenue from selling carbon permits at favorable market prices, thereby reducing overall operating costs. The reward function consists of three components: cost savings from energy reduction, revenue from carbon permit sales, and tenant discomfort. These components are integrated into a single reward function, as shown in Equation (21). All components of the reward function are normalized to the range $[0, 1]$, as shown in Equation (22):

$$r_t = -\lambda_1 \tilde{Comf}_{cost, t} - \lambda_3 \tilde{Disc}_{PMV, t} - \lambda_4 \tilde{Disc}_{Humidex, t} + \lambda_2 \tilde{Rev}_{carbon, t}. \quad (21)$$

Here, $Comf_{cost}$ can be obtained from Equation (9), $Rev_{carbon, t}$ can be calculated through Equations (22) and (23), and $Disc_{PMV, t}$ and $Disc_{Humidex, t}$ are calculated through Equations (24)–(27). Weighting parameters λ_1 , λ_3 , and λ_4 are fixed at 3×10^{-1} , and λ_2 is fixed at 1×10^{-1} . The proportion of weighting parameters is considering the trade-off structure between energy cost, discomfort, and carbon revenue.

To ensure numerical stability and balanced feature scales, all reward variables with different units (USD, CO₂eq, and comfort indices) were normalized to the same $[0, 1]$ range using min–max scaling.

$$\text{norm}(x) = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \quad (22)$$

$Comf_{cost, t}$ is negative because excessive energy consumption for comfort is considered a penalty, which helps the agent to minimize comfort-related costs. Figure 5 illustrates the reward system used in this study.

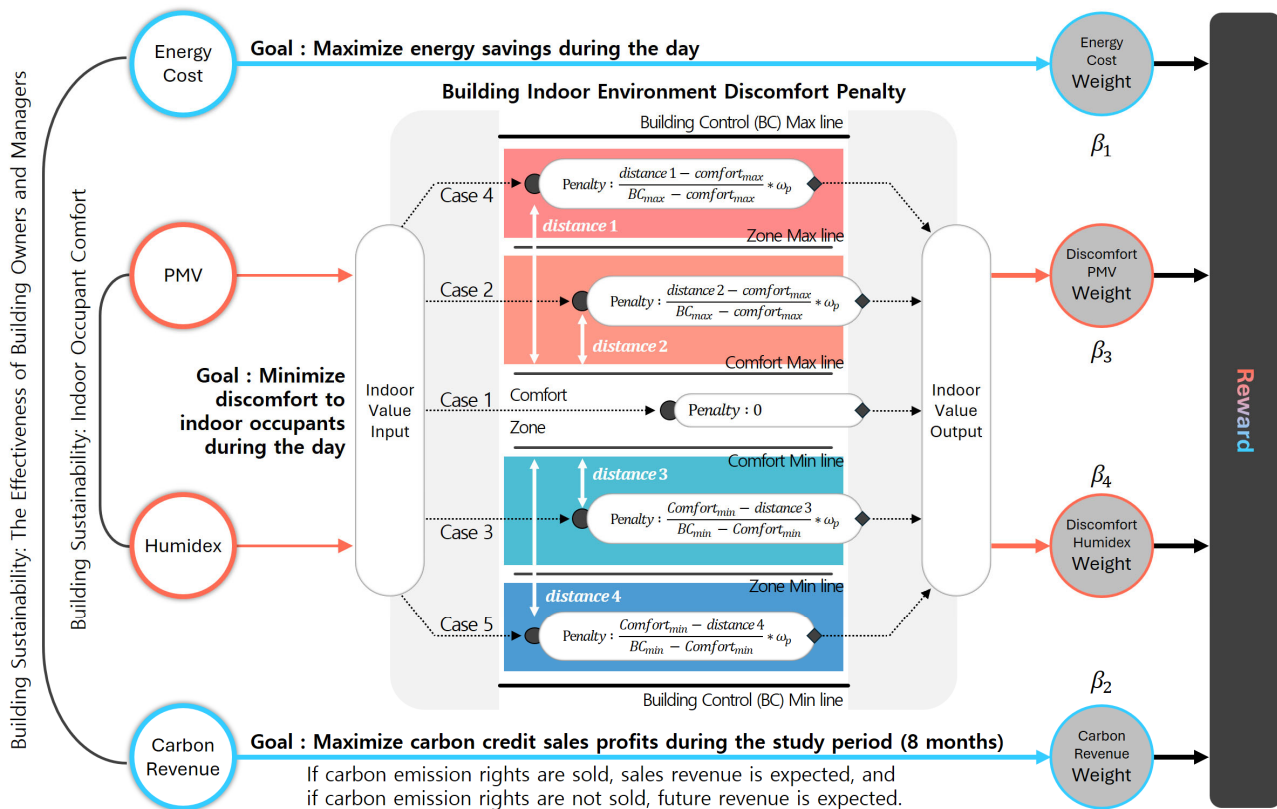


Figure 5. Deep reinforcement learning reward system.

As shown in Figure 5, defining minimum and maximum values for building control is necessary for normalization. In this study, indoor temperature is assumed to range from 15 °C to 40 °C and humidity from 30% to 100%. These ranges are used to determine the maximum and minimum values of PMV and Humidex for the target building.

The reward associated with revenue from carbon permit sales, Rev_{carbon} is defined as follows:

$$Rev_{\text{carbon}, t} = \begin{cases} CP \times P_t^{CP}, & 1 \geq CP \text{ and agent sell} \geq 1 \\ 0, & 1 \geq CP \text{ and agent sell} = 0 \\ 0, & \text{otherwise} \end{cases} \quad (23)$$

Here, Rev_{carbon} can be obtained from Equation (23), and the weighting parameter λ_2 is fixed at 1×10^{-1} , which reflects the trade-off between discomfort and carbon revenue. The number of carbon permits (CP) is determined using Equation (24), based on the energy savings computed from Equation (17). Revenue from carbon permits is then calculated by multiplying the number of permits sold, CP, by the carbon permit price at the time of sale P_t^{CP} . As Rev_{carbon} increases with higher carbon permit prices, the agent is incentivized to sell when the market price is favorable. However, a minimum level of energy savings is required to obtain a single CP. Therefore, the carbon permit component does not provide immediate rewards for the agent's actions. To address this, the agent is not allowed to sell CPs at every step; instead, the future expected revenue from CPs, derived from energy savings, is incorporated. While this expected revenue does not directly contribute to the reward function, it is reflected as part of the state transition resulting from the agent's actions.

A minimum threshold is required to trade carbon permits. CPs can only be obtained when the reduction in emissions exceeds this threshold; otherwise, no revenue is generated. To gain the right to sell in the carbon market, at least one carbon permit must be secured. One CP represents the minimum tradable unit in the carbon market and corresponds to

one metric ton of carbon dioxide equivalent (tCO_2eq). This refers to either one ton of CO_2 or the amount of another greenhouse gas with an equivalent global warming potential to one ton of CO_2 . In other words, to sell a CP, energy consumption must be reduced by the amount equivalent to one ton of CO_2 emission. According to the Energy Commission of Malaysia, the carbon emission factor is $0.774 tCO_2eq$ per MWh. Since one carbon permit corresponds to $1 tCO_2eq$, a reduction of $1000 kg CO_2eq$ is required to obtain a permit, as shown in Equation (24). Based on the emission factor, this corresponds to approximately $1290 kWh$ of electricity savings. Once accumulated energy savings reach the minimum threshold required to obtain a carbon permit, the decision to sell is made based on the current market price of CPs. Unlike other reward components, carbon permit revenue is designed to be maximized over the full training cycle, rather than at each individual step.

$$CP = \text{Carbon}_{\text{save}}^{\text{AHU}} / 1000. \quad (24)$$

The reward component related to tenant discomfort, referred to as *Discomfort*, is defined in Equations (24) and (25). Overall discomfort is derived from the average of the normalized PMV and Humidex indices. The agent controls AHUs across eight zones and receives rewards reflecting the indoor environmental quality in each zone. The discomfort component is calculated by first obtaining the PMV and Humidex indices for all eight zones. $PMV_{i,t}$ and $Humidex_{i,t}$ denote the measured/estimated values at zone i and time t . Then, the agent receives aggregated normalized values, defined as $PMV_{\text{nor}}^{\text{ZONE}}(\sum_{i \in Z}^n PMV_{i,t})$ and $Humidex_{\text{nor}}^{\text{ZONE}}(\sum_{i \in Z}^n Humidex_{i,t})$, where $n = 8$ ($Z = \{1, \dots, 8\}$ is the set of zones (so $n = |Z| = 8$)) denotes the number of zones. There are ten possible combinations representing acceptable comfort ranges in indoor environments, as shown in Table 5. The variable i_{val} represents the present value of each comfort index in a zone. The parameters $comfort_{\min}$ and $comfort_{\max}$ represent the common lower and upper bounds of the comfort range across all zones. The parameters $z_{\min}$ and $z_{\max}$ represent the zone-specific minimum and maximum observed for each comfort index. These values are derived from the PMV and Humidex indices during the nine-month observation period. Further details are described in Section 2.3.1 and Figure A1.

Table 5. Comfort ranges and environmental conditions in buildings.

	Condition Case	Range
PMV	Condition 1	$PMV_{comfort_{\min}} \leq PMV_{i_{val}} \leq PMV_{comfort_{\max}}$
	Condition 2	$PMV_{i_{val}} < PMV_{comfort_{\min}}$
	Condition 3	$PMV_{i_{val}} > PMV_{comfort_{\max}}$
	Condition 4	$PMV_{i_{val}} < PMV_{z_{\min}}$
	Condition 5	$PMV_{i_{val}} > PMV_{z_{\max}}$
Humidex	Condition 6	$Humidex_{comfort_{\min}} \leq Humidex_{i_{val}} \leq Humidex_{comfort_{\max}}$
	Condition 7	$Humidex_{i_{val}} < Humidex_{comfort_{\min}}$
	Condition 8	$Humidex_{i_{val}} > Humidex_{comfort_{\max}}$
	Condition 9	$Humidex_{i_{val}} < Humidex_{z_{\min}}$
	Condition 10	$Humidex_{i_{val}} > Humidex_{z_{\max}}$

Because a wide range of environmental conditions can arise within the building, the reward function assigns different weights to the penalty term p depending on the indoor environment. Equations (25) and (26) define the penalty term ω , which is applied when the conditions in Table 5 are violated, depending on the environmental state. Specifically,

ω_{p-low_z} , ω_{p-high_z} , ω_{h-low_z} , and ω_{h-high_z} are applied when the zone-level conditions z_{min} or z_{max} are reached but the comfort range is violated. In contrast, ω_{p-low_m} , ω_{p-high_m} , ω_{h-low_m} , and ω_{h-high_m} are applied when the conditions exceed the boundary limits of z_{min} or z_{max} . The out-of-range conditions are classified into negative (low, representing colder conditions) and positive (high, representing hotter conditions) directions, and the penalty differs accordingly. These distinctions are made because the scale of human discomfort varies with increases in temperature and humidity. To reflect this realistically, the agent is designed to receive greater penalties under conditions that are likely to cause higher discomfort [124]. If indoor conditions fall within the comfort range, no penalty is applied.

$$Disc_{PMV, t} = \frac{1}{n} \sum_{i \in Z} PMV_{i, t}^p \quad (25)$$

$$Disc_{Humidex, t} = \frac{1}{n} \sum_{i \in Z} Humidex(h)_{i, t}^p \quad (26)$$

$$PMV_i^p = \begin{cases} \text{norm}(PMV_{comfort_{min}} - PMV_{i_{val}}) \times \omega_{p-low_z}, & \text{if Condition 2} \\ \text{norm}(PMV_{i_{val}} - PMV_{comfort_{max}}) \times \omega_{p-high_z}, & \text{if Condition 3} \\ \text{norm}(PMV_{z_{min}} - PMV_{i_{val}}) \times \omega_{p-low_m}, & \text{if Condition 4} \\ \text{norm}(PMV_{i_{val}} - PMV_{z_{max}}) \times \omega_{p-high_m}, & \text{if Condition 5} \\ 0, & \text{if Condition 1} \end{cases} \quad (27)$$

$$Humidex(h)_i^p = \begin{cases} \text{norm}(h_{comfort_{min}} - h_{i_{val}}) \times \omega_{h-low_z}, & \text{if Condition 7} \\ \text{norm}(h_{i_{val}} - h_{comfort_{max}}) \times \omega_{h-high_z}, & \text{if Condition 8} \\ \text{norm}(h_{z_{min}} - h_{i_{val}}) \times \omega_{h-low_m}, & \text{if Condition 9} \\ \text{norm}(h_{i_{val}} - h_{z_{max}}) \times \omega_{h-high_m}, & \text{if Condition 10} \\ 0, & \text{if Condition 6} \end{cases} \quad (28)$$

4.4. Constraints

To ensure that the proposed reinforcement learning model does not converge to extreme policies but instead reflects realistic operational conditions, several constraints are imposed:

1. **Comfort range.** The comfort range is defined as the range within which energy savings can be achieved while minimizing tenant discomfort. In this study, the operational range for each zone was defined using minimum and maximum values observed during a nine-month data collection period. In the target building, indoor temperature and humidity varied by location, including the lobby, lower, middle, and upper floors. Indoor temperature in the building was observed to range from a minimum of 19.5 °C to a maximum of 29.4 °C, and relative humidity from 50.1% to 97.7% across weekdays and weekends. Accordingly, the indoor temperature was constrained to between 18 °C and 32 °C, and relative humidity to between 45% and 100%.
2. **Owner's requirement.** In accordance with the building owner's request, indoor temperature must be maintained within the range of 23 °C to 26 °C. The agent is required to satisfy this constraint by default.
3. **Occupancy schedule.** The target building follows a fixed office schedule, with working hours on weekdays and non-occupancy on weekends. The agent is trained to control AHUs efficiently during weekday working hours in line with this pattern.

4. **Pre-conditioning.** Since the tenant's working hours begin at 9:00, the agent can pre-adjust indoor temperature and humidity in advance in line with Constraints 2 and 3. Furthermore, between 8:00 and 9:00, just before occupancy, the agent can apply a multiplier ranging from 1 to 10 times the average rate of change in temperature and humidity. Details of the change rates are provided in Section 4.5.2.
5. **AHU operation.** The testbed building cannot operate all AHUs simultaneously due to the risk of energy overload. In practice, AHUs are controlled sequentially by floor and zone according to a scheduling strategy. However, since this study focuses on a subset of AHUs rather than the entire system, it is assumed that all selected AHUs can be operated simultaneously.
6. **Switching constraint.** Frequent power switching increases mechanical stress on the equipment and undermines the building's long-term sustainability. In addition, some AHU components also require approximately 30 min of preparation time after shutdown before restarting. Therefore, once an AHU switches from OFF to ON, the agent is restricted from turning it OFF again for at least one hour. By contrast, after switching from ON to OFF, no such restriction applies.

4.5. Algorithm Implementation

4.5.1. SAC Learning Algorithm

The agent responsible for controlling indoor comfort and reducing energy consumption is modeled to learn an optimal policy that adapts to dynamic environmental changes. The evolving environment includes variations in AHU operational states, indoor temperature and humidity, energy consumption, and carbon emissions. The Soft Actor–Critic (SAC) deep reinforcement learning is applied to maintain indoor thermal conditions while optimizing energy operations. A comprehensive explanation of the algorithm can be found in Algorithm 1 (Soft Actor–Critic for Building Energy Management). The algorithm begins by initializing key parameters: the replay buffer (for storing experiences), learning rate (which controls the speed of learning), discount factor (which determines the weight of future rewards), entropy coefficient (which balances exploration and exploitation), and the maximum number of training episodes N . The policy network π_θ , two Q-networks $Q_{\phi1}$, $Q_{\phi2}$, and replay buffer D are also initialized. Training proceeds for N episodes, with each episode corresponding to one day. The agent is trained to learn an optimal policy that maximizes the cumulative reward within each daily episode. After the initial state s_0 is set, the environment is simulated in 15 min intervals. At each step, the agent determines the ON/OFF status of the AHUs, then observes the reward r_t and the next state s_{t+1} . The transition tuple (s_t, a_t, r_t, s_{t+1}) , generated from the agent's interaction with the environment, is stored in the replay buffer D . A mini-batch is sampled from the replay buffer for training. In order to mitigate overestimation, the algorithm selects the smaller value between the two Q-networks. This minimum Q-value, together with the observed reward, is used to compute the target value y_t . The Q-networks are then updated by minimizing the loss function, which reduces the gap between predicted Q-values and the target. The policy network is then updated by minimizing the policy loss function $L\pi(\theta)$. Finally, the entropy coefficient α is adjusted to balance exploration and exploitation. This process repeats across all episodes until episode N , after which the algorithm terminates.

Algorithm 1: Soft Actor–Critic for Building Energy Management

Input: Replay buffer D , learning rate α , discount factor γ , entropy coefficient α , maximum episodes N

Output: Optimized policy π^*

Initialize policy network π_θ , Q-networks $Q_{\phi 1}$, $Q_{\phi 2}$, and replay buffer D

for each episode = 1 to N do

Initialize state s_0 (e.g., $\{E_{elec}, E_{info}, C_{info}, AHU_{state}, T_{in}, H_{in}, PMV_{zone}, Humidex_{zone}, T_{out}, H_{out}, t\}$)

for each timestep $t = 15$ to T do

Select action $a_t = \{AHU_{set}, T_{factor}, H_{factor}, CP_{sell}, CP_{ratio}, CP_{accu.}\} \sim \pi_\theta(a_t|s_t)$

Observe reward r_t and the next state s_{t+1}

Store transition (s_t, a_t, r_t, s_{t+1}) in D

Sample mini-batch of transitions (s_t, a_t, r_t, s_{t+1}) from D

Compute target value:

$$y_t = r_t + \gamma(\min(Q_{\phi 1}(s_{t+1}, a_{t+1}), Q_{\phi 2}(s_{t+1}, a_{t+1})) - \alpha \log \pi_\theta(a_{t+1}|s_{t+1}))$$

Update Q-networks by minimizing:

$$L_Q(\phi) = \frac{1}{2} \sum (Q_\phi(s_t, a_t) - y_t)^2$$

Update policy network by minimizing:

$$L_\pi(\theta) = \mathbb{E}(\log \pi_\theta(a_t|s_t) - Q_\phi(s_t, a_t))$$

Adjust the entropy coefficient α to maintain target entropy

$$L_\alpha = \mathbb{E}[-\alpha(\log \pi_\theta(a_t|s_t) + \alpha H)]$$

If termination condition:

break

end for

end for

4.5.2. Algorithm for Environmental Temperature and Humidity Changes

The following algorithm applies changes in indoor temperature and humidity based on the agent's decision to turn the AHU ON or OFF. This algorithm is developed from environmental data collected through Energy IoT devices installed in the building. Based on the analyzed data, the agent applies monthly and hourly variations in temperature and humidity to the environment, depending on the AHU ON/OFF status. The environment evolves according to the current and sustained state of the AHU, and the agent responds to the outcomes of these environmental changes.

Figure 6 illustrates the average changes in temperature and humidity at 15 min intervals, derived from data collected between 1 July 2024, and 31 March 2025. The observed patterns can be classified into three major characteristics.

First, there is a decrease phase characterized by rapid changes in temperature and humidity during the period one hour before and after the start of working hours. In this phase, three patterns, labeled as ③, ④, and ⑤, are observed at around 8:00, 9:00, and 10:00, respectively. Each of these time periods shows a greater change than the average hourly variation. In addition, there were clear differences in the magnitude of change across various zones. The 28th floor, located in the upper part of the building, exhibited the largest variations, followed by the 2nd floor in the lower part of the building. The remaining zones showed changes similar in magnitude to the overall average.

Second, a maintenance phase is observed, during which indoor temperature and humidity remain stable. In phase ⑥, despite the AHU remaining ON, both variables show minimal changes, staying close to zero.

Finally, there is an increase phase characterized by a sharp rise in temperature and humidity after working hours end. Within this phase (represented by patterns ① and ②),

temperature and humidity display diverging trends. The period around 18:00 is marked by the coexistence of both AHU ON and OFF states. This leads to a greater increase in temperature at 18:30 compared to 19:00, while humidity shows a larger increase at 19:00 than at 18:30. In addition, an irregular phase, labeled as (f), was observed, during which humidity exhibited inconsistent patterns.

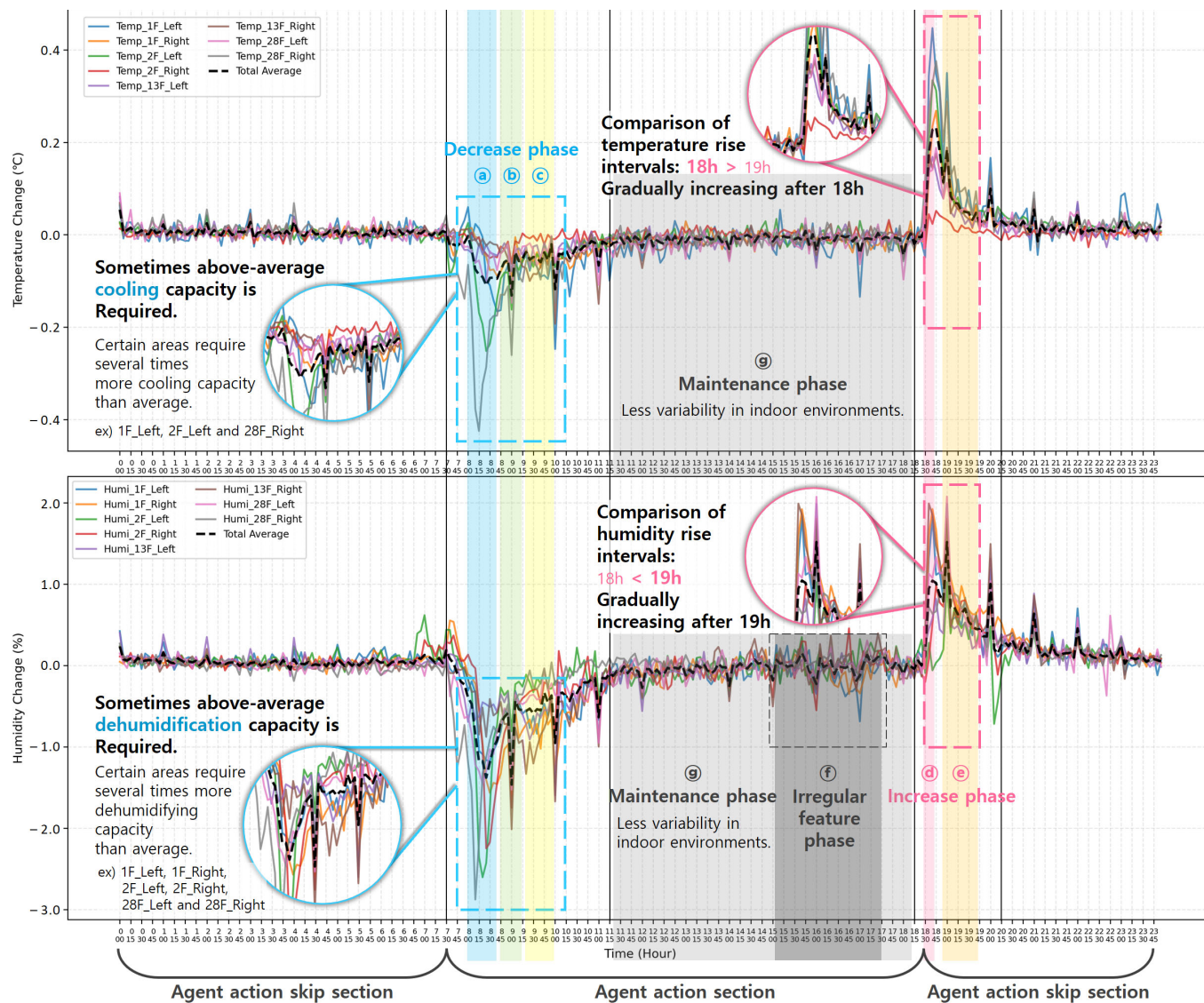


Figure 6. Control patterns for indoor comfort in the target building (15 min change rate).

An environmental algorithm was developed based on these observed variation patterns, which allows the DRL model to operate in a manner that closely resembles real-world conditions. The algorithm that models the indoor environmental changes resulting from the agent's actions is described in detail in Algorithm 2 (Temperature and Humidity Change Decision).

Algorithm 2: Temperature and Humidity Change Decision

Input: Zone z , Current time t , AHU status ahu_{ON} ;
 Temperature/humidity change table D ;
 Previous AHU states S_{prev} ;
 T-H change multiplier $\alpha T, \beta H$
 Output: Temperature change ΔT , humidity change ΔH
 Set temperature column $col_T \leftarrow$ Temp \$z\$ base average difference;
 Set humidity column $col_H \leftarrow$ Humi \$z\$ base average difference;
 Get previous AHU state $ahu_{prev} \leftarrow S_{prev}[z]$;
 if $07:30 \leq t \leq 18:00$ then
 if $ahu_{ON} = 0$ and $ahu_{prev} = 1$ then
 $\Delta T \leftarrow D$ [year = t_y , month = t_m , hour = 18, minute = 30] [col_T];
 $\Delta H \leftarrow D$ [year = t_y , month = t_m , hour = 18, minute = 45] [col_H];
 else if $ahu_{ON} = 0$ and $ahu_{prev} = 0$ then
 $\Delta T \leftarrow D$ [year = t_y , month = t_m , hour = 18, minute = 45] [col_T];
 $\Delta H \leftarrow D$ [year = t_y , month = t_m , hour = 18, minute = 30] [col_H];
 else if $ahu_{ON} = 1$ and $ahu_{prev} = 0$ then
 Extract all rows M from D where year = t_y and month = t_m ;
 Filter $M_T \leftarrow$ rows in M with $col_T < 0$;
 Filter $M_H \leftarrow$ rows in M with $col_H < 0$;
 $\Delta T \leftarrow \min(M_T [col_T])$ or 0.0;
 $\Delta H \leftarrow \min(M_H [col_H])$ or 0.0;
 else if $ahu_{ON} = 1$ and $ahu_{prev} = 1$ then
 $\Delta T \leftarrow D$ [year = t_y , month = t_m , hour = t_h , minute = t_m] [col_T];
 $\Delta H \leftarrow D$ [year = t_y , month = t_m , hour = t_h , minute = t_m] [col_H];
 else
 $\Delta T \leftarrow D$ [year = t_y , month = t_m , hour = t_h , minute = t_m] [col_T];
 $\Delta H \leftarrow D$ [year = t_y , month = t_m , hour = t_h , minute = t_m] [col_H];
 if $08:00 \leq t \leq 09:00$ then
 if $ahu_{ON} = 1$ then
 $\Delta T \leftarrow \Delta T * \alpha T$;
 $\Delta H \leftarrow \Delta H * \beta H$;
 return $\Delta T, \Delta H$

In anticipation of working hours, the agent is permitted to take continuous actions within the range of 1 to 3 between 8:00 and 9:00, depending on the indoor comfort level. In case of temperature extremes, the amplification factor can be expanded by a factor of three to as much as a factor of ten. Depending on this value, the ranges of temperature and humidity fluctuations during the period can be amplified by up to three times the 15 min interval average variations by month and time of day. These average variations represent the mean values across the entire space. The amplification factor was derived from analysis of the collected environmental data, which enables adjustment to account for differences in fluctuation ranges across individual zones.

The period between 8:00 and 10:00 a.m. is critical for ensuring a comfortable indoor environment for tenants. As shown in Figure 2, all AHUs are turned OFF after business hours, causing indoor temperature and humidity to rise overnight due to external conditions. Restoring indoor conditions to a comfortable range before working hours begin at 9:00 a.m. is essential, as it allows occupants to remain comfortable throughout their stay in the building. If cooling is initiated too early, it may lead to excessive AHU energy consumption, whereas activating it too late results in prolonged discomfort for the tenants.

Furthermore, even when AHUs are switched from OFF to ON, indoor temperature and humidity do not immediately return to comfortable levels. This is consistent with the findings shown in Figure 6, which indicate that building managers typically focus on indoor comfort control between 8:00 and 9:00 a.m. The period from 9:00 to 10:00 a.m. can be interpreted as a continuation of efforts to stabilize and maintain indoor comfort.

Depending on the state transitions of the AHU, four possible scenarios can occur during working hours. If the AHU remains continuously ON, as in the current schedule, the monthly and hourly average variations are applied. When the AHU switches from ON to OFF, the environmental change is determined by the period with the greatest increase in temperature and humidity. If the AHU remains OFF, the change is applied based on the period following the hour with the highest increase in temperature and humidity. When the AHU switches from OFF to ON, the change is applied according to the average variation observed during the period with the greatest decrease in temperature and humidity.

5. Results

This section presents the training and validation results of energy savings and carbon permit sales, based on the deep reinforcement learning model described earlier, while maintaining indoor comfort in the target building. Due to its geographic location near the equator, the target building experiences a year-round hot and humid tropical rainforest climate with no distinct seasonal divisions. Therefore, the results for training and validation are those that only include specific climates.

To ensure reproducibility, a fixed random seed was applied during all training and evaluation processes. Reinforcement learning algorithms inherently involve pseudo-random processes at multiple stages, including the initialization of neural network weights, the initial environmental state, and the agent's stochastic action selection. By fixing the seed, the same pseudo-random number sequence is generated in every execution, maintaining consistent experimental conditions and enabling identical results to be reproduced. This approach is particularly important for reproducibility in research validation. It also enables fair experimental comparison, ensuring that performance differences arise from changes in model architecture or hyperparameters rather than random initialization. In addition, seed fixing enhances debugging efficiency, as deterministic behavior allows potential issues to be replicated and corrected under identical conditions. Although the fixed seed setting leads to deterministic outcomes, it was intentionally adopted to guarantee experimental consistency. To quantify the reliability and robustness of the model, repeated training with various seed configurations may be required to verify policy uncertainty.

5.1. Development of the Training and Evaluation Environment

A simulation environment was developed in Python to closely resemble real-world conditions, enabling interactions with the environment based on the agent's actions using data collected from the target building. The simulation incorporated AHU energy consumption as well as zone-specific temperature and humidity data, with AHU power consumption values obtained from actual measurements at the site. The PMV for discomfort calculation was computed using the Python library `pythermalcomfort` (functions `pmv-ppd-ashrae` and `Humidex`). The primary Python libraries used to construct and develop the deep reinforcement learning environment are listed in Table A1.

The SAC algorithm was implemented using the PyTorch library within an Anaconda environment on a desktop equipped with an Intel Core i5 CPU, 64 GB of RAM, and an NVIDIA RTX 4060 Ti GPU. Under this setup, the model performed simulation loops corresponding to daily operation hours (07:30–18:30) for a total of 16,700 episodes, which required an average computation time of approximately 160 min. The Soft Actor–Critic

(SAC) model developed in this study was designed with a lightweight yet stable multi-layer perceptron (MLP) architecture. Both the actor and critic networks consist of two fully connected layers with 256 hidden units each. The actor network receives the input state and passes it through two hidden layers of 256 neurons each to estimate the mean and log standard deviation of the action distribution. The output is normalized to the range $[-1, 1]$ using the tanh activation to reflect the constraints of the action space. The network uses the ReLU activation function, and orthogonal initialization is applied to enhance training stability. The critic network takes the concatenated state–action pair as input and passes it through the same two-layer MLP structure to predict the state–action value $Q(s, a)$. Following the double Q-learning architecture of SAC, two independent Q-functions (Critic1 and Critic2) are trained in parallel, each outputting a scalar Q-value.

The Smooth L1 (Huber) loss is used to prevent divergence during training, and an additional L2 weight decay term is introduced to reduce overfitting. The overall model contains approximately 330,000 trainable parameters, which ensures efficient learning and inference for mid-scale reinforcement learning problems.

5.2. Hyperparameters for DRL

To learn a building operation strategy that ensures both energy-efficient AHU control and indoor comfort, the deep reinforcement learning parameters are configured as shown in Table A2. The agent decides every 15 min whether to turn the AHU ON or OFF. The subsequent action is determined according to the range of indoor PMV and Humidex observed over the 15 min interval. However, for equipment stability, the actions are partially constrained by the sixth constraint defined in Section 4.4.

The agent was designed to maximize its reward daily, and profit from carbon permit sales was structured to yield the maximum reward during the training period. For the PMV calculation, the air velocity was set to 0.1 m/s, which is a commonly used value for office environments. The metabolic rate was set to 1.0 met, representing the activity level of office workers using computers. Clothing insulation was set to 0.5 clo, corresponding to light summer attire such as short-sleeved shirts and thin trousers. For the target building, air temperature and mean radiant temperature were assumed to be equal, and the difference was set to zero.

Selecting appropriate learning parameters (i.e., hyperparameters) is one of the most critical steps in designing a DRL framework. These parameters must be determined according to the complexity of the problem and the dimensionality of the state space. In most DRL studies, they are identified through trial-and-error in simulation environments rather than analytical formulation [118,119]. Similarly, in this study, the hyperparameters were manually tuned based on iterative experimentation. The final configurations used for training are summarized in Table A2. Although this manual tuning approach produced stable learning performance, it does not guarantee optimality.

5.3. Training Performance

5.3.1. Reward Results by Training Progress

The agent controls the AHUs from 07:30 to 18:30, covering 11 h per day with action intervals of 15 min. For training, the deep reinforcement learning model uses eight months of data collected between 1 July 2024 and 29 February 2025. Figure 7 illustrates the training results for weekdays only, excluding weekends, with a total of 16,700 episodes. The upper graph shows the rewards obtained during training, while the lower graph depicts the ratio of energy consumption cost and carbon permit sales cost to the building's operating cost, along with episode-wise carbon permit sales cost. One episode corresponds to a

single day, and one complete training cycle, excluding weekends and holidays, consists of 167 episodes.

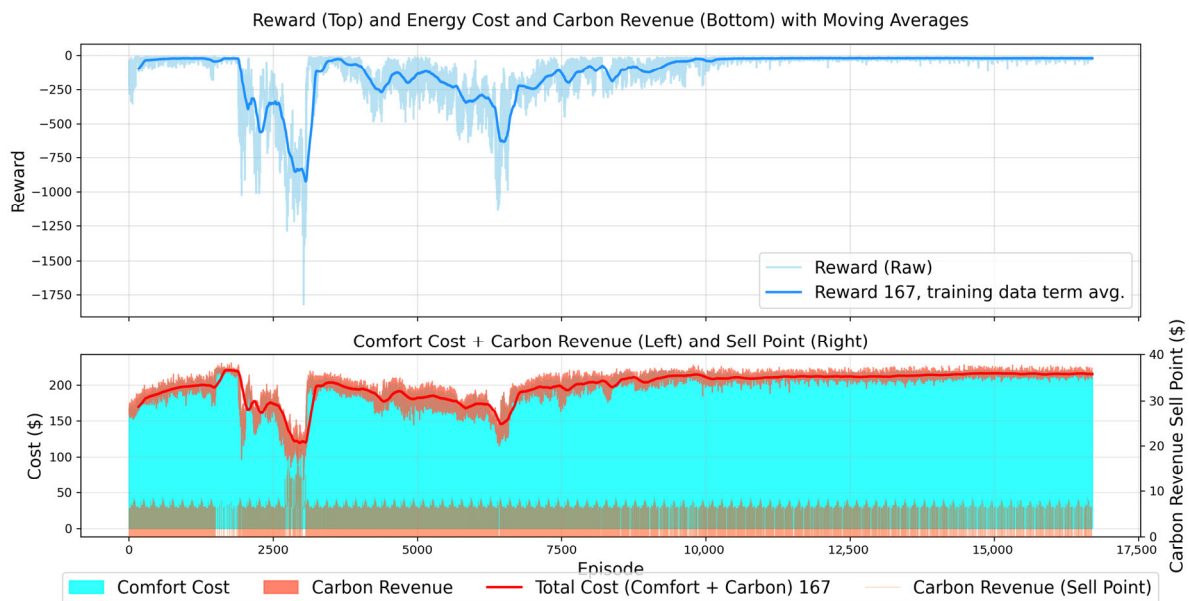


Figure 7. Learning curves and carbon-revenue trajectories of the SAC agent trained on data from 1 July 2024 to 28 February 2025.

In Figure 7, the light blue shading in the upper graph represents the results of each episode, while the blue line shows the average over 167 episodes. In the lower graph, the red shading denotes the proportion of carbon permit sales cost for each episode, and the cyan shading represents the AHU energy consumption required to maintain indoor comfort. The carbon revenue (sell point) displayed at the bottom of the lower graph indicates the realized carbon permit revenue during each episode. The red line shows the average total operating cost of the building across 167 episodes.

From the graphs, it can be seen that in the early training stage, the agent explores mainly with a focus on indoor comfort. Around 2500 episodes, the policy shifts toward CP-oriented strategies. Between about 5000 and 7500 episodes, the agent explores policies emphasizing comfort and energy savings. Up to 10,000 episodes, the policy gradually shifts toward minimizing rewards. Beyond 10,000 episodes, the rewards converge near zero, which indicates stabilization and a balance between indoor comfort, reduced energy use, and CP sales.

5.3.2. Agent Action Policy Inference

The agent in this study operates in a 29-dimensional action space: 10 dimensions for AHU control, 3 for carbon permit revenues, and 16 for temperature and humidity variation coefficients across eight zones. Policy outputs are mapped to their dimensions, from $[-1, 1)$ to $[0, 1]$ or $[1, 3]$. Figure 8 shows how the 29-dimensional actions evolve over 16,700 episodes, grouped into early, middle, and late training stages. The action dimensions responsible for AHU control are shown on the left side of the graph, ranging from 1F L to 28F R. ‘C Sell’ denotes the decision to sell carbon permits, ‘C Ratio’ represents the sales volume, and ‘C Acc’ indicates the holding period. From 1F LT to 28F RT, the dimensions correspond to temperature variation coefficients, while from 1F LH to 28F RH, they correspond to humidity coefficients. The left Y-axis represents transformed outputs of AHU and carbon permit-related actions, which are mapped to values between 0 and 1. The right Y-axis shows the transformed outputs of temperature and humidity variation

coefficients, mapped to values between 1 and 3, as these can scale up to three times the average change during 08:00–09:00.

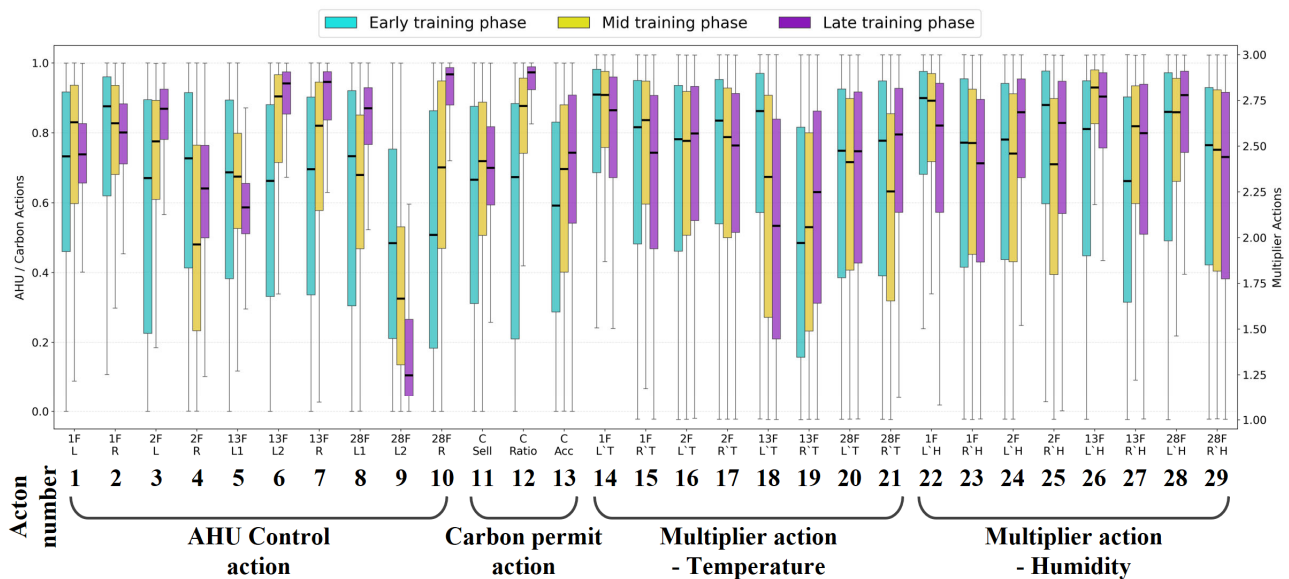


Figure 8. Changes in the 29-dimensional actions of the DRL model agent during training.

Action dimensions 1–10 correspond to AHU control, where values closer to 1.0 indicate a higher probability of switching ON, while lower values increase the probability of switching OFF. In the early stage of training, the interquartile range (IQR) spans about 0.2–0.9, with the median mostly between 0.4 and 0.8, indicating effective exploration. In the middle stage, the IQR narrows and the median rises above 0.6, suggesting that the policy becomes more conservative, shifting from exploration toward exploitation. In the late stage, the policy converges to a higher proportion of ON states than OFF states to maintain indoor comfort. An exception is observed for 28F L2, where the OFF proportion gradually increases from the middle stage, converging to frequent OFF switching in the late stage. Overall, the AHU control policy converges to ON-dominant switching for most units, with 28F L2 converging to OFF-dominant switching.

Action dimensions 11–13, which are associated with CPs, exhibit an IQR of about 0.2–0.9 during the early training stage, with the median around 0.6, which indicates effective exploration. Among them, C Sell determines whether to sell the carbon permits held (with values closer to 1.0 corresponding to a higher probability of selling). During the early and middle stages, the median stays near 0.65 while the IQR gradually narrows. In the late stage, the policy shifts toward favoring sales, which shows that selling becomes the preferred strategy. C Ratio is the action dimension that determines whether to sell one carbon permit or multiple permits, while C Acc represents the holding period, which indicates whether CPs are retained for a short or long duration. Similarly to C Sell, both C Ratio and C Acc display diverse outputs in the early training stage, reflecting active exploration. In the middle stage, C Ratio converges toward policies favoring larger sales volumes than in the early stage, whereas C Acc maintains moderate exploration but gradually shifts toward a more conservative approach. In the late stage, C Ratio converges to a policy of selling most available permits, while C Acc converges toward retaining permits instead of selling them immediately after acquisition. Overall, the policy for generating revenue from carbon permits converges to a strategy of retaining them for a period and then selling a large block of accumulated permits at once.

Action dimensions 14–21 and 22–29 determine the weighting of indoor temperature and humidity variations, respectively, when the AHUs are switched ON between 08:00

and 09:00. In the early stage of training, dimensions 14–21 for temperature and 22–29 for humidity show IQRs greater than 1.5, with most medians between 2.50 and 2.75. In the middle stage, policy changes are more pronounced in dimensions 22–29 than in 14–21, revealing zone-specific patterns where the medians either rise or fall depending on the zone. In the late stage, the weightings converge within 2.25–2.75 across zones to maintain comfortable indoor conditions after 09:00. The zone-specific increases and decreases in weighting reflect the need to avoid discomfort, since excessive adjustments during 08:00–09:00 could cause overly uncomfortable or excessively cold indoor conditions later in the day. Moreover, frequent ON/OFF switching is restricted to prevent extreme cooling. Consequently, the weightings for temperature and humidity variations during 08:00–09:00 converge to policies designed with consideration of comfort in the subsequent midday and afternoon periods.

Figure 8 illustrates the temporal evolution of the policy decision patterns across the early, middle, and late stages of training. The policies generated by the agent during its interaction with the environment were recorded and analyzed using histograms and kernel density estimates (KDE) to visualize the progression and stabilization of the action distributions over time. Therefore, this interpretation is derived from the observed policy outputs. Integrating explainable AI (XAI) techniques such as SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Gradient-weighted Class Activation Mapping (Grad-CAM) could further enhance the transparency and reliability of the model by providing interpretable insights into the agent's decision-making process.

- **Histograms and Kernel Density Estimates (KDE):** A histogram represents the distribution of data by dividing it into multiple intervals (bins) and visualizing the frequency within each interval. The kernel density estimate (KDE) provides a smooth, continuous approximation of the probability density function, allowing for an intuitive understanding of the overall shape and tendency of the data distribution.
- **SHAP (SHapley Additive exPlanations):** SHAP is a method based on game theory's Shapley values that quantitatively measures the contribution of each input feature to the model's prediction. By considering all possible combinations of input variables, it calculates the average marginal contribution of each feature to the prediction result. This enables interpretation of how each state variable influences the policy's decision-making process.
- **Feature Importance Analysis:** Feature importance analysis is a method used to evaluate how much each input variable contributes to a model's predictions or decisions. By quantifying the relative influence of individual features on the output, this analysis helps identify which variables play the most significant roles in determining model behavior. In the context of reinforcement learning, feature importance analysis can reveal which environmental factors or state variables most strongly affect the agent's policy and decision-making process, thereby improving the interpretability and transparency of the model.
- **LIME (Local Interpretable Model-agnostic Explanations):** LIME is a model-agnostic approach that fits a simple surrogate model around a specific prediction to explain why the model produced that outcome. It provides local interpretability by approximating the complex model's behavior near a particular instance.
- **Grad-CAM (Gradient-weighted Class Activation Mapping):** Grad-CAM is a technique used in deep neural networks that visualizes which parts of the input contributed most to the model's decision using gradients of the output. Although widely applied in computer vision, it can also be used in policy analysis to generate visual heatmaps that highlight which states or actions were most influential in the agent's decision-making process.

5.3.3. DRL Model Training Stability Analysis

Figure 9 illustrates the evolution of key training metrics including actor loss, critic losses, gradient norms, entropy regularization (α and $\log \alpha$), and entropy–target gaps during training. To ensure the reliability and reproducibility of the proposed SAC-based control framework, we analyzed the stability of training dynamics across multiple internal metrics. The SAC algorithm inherently relies on entropy-regularized policy updates; therefore, both actor and critic loss trajectories, along with entropy coefficients (α), are critical for assessing convergence and exploration balance.

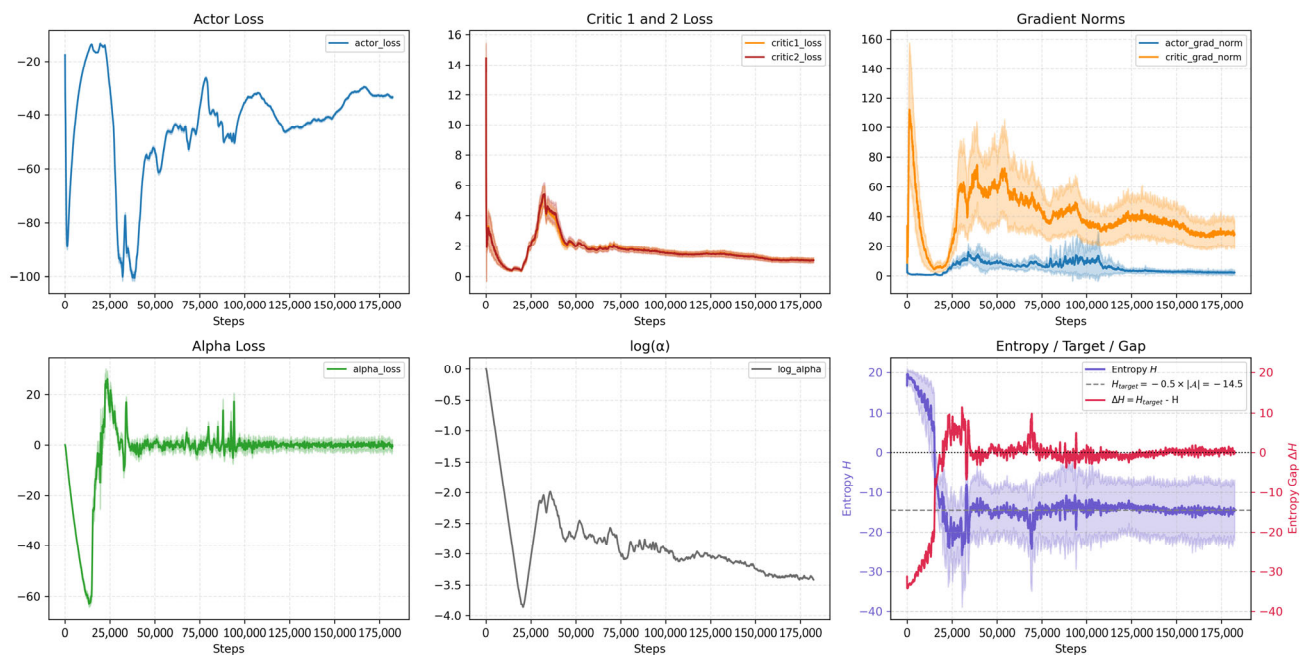


Figure 9. Loss curve according to the training process of the SAC algorithm.

For training efficiency and faster convergence, the target entropy value was set to half of the action dimension. Although it is common practice to set the target entropy as equal to the full action dimension, we adopted a smaller target to enhance training stability and improve policy consistency [125,126].

After approximately 75 k–100 k steps, all monitored metrics reached quasi-stationary states, indicating that the SAC agent achieved policy convergence with consistent gradient flow, stable value estimation, and balanced entropy control. These characteristics collectively validate the robustness of the proposed training configuration and its suitability for continuous control in complex building energy environments.

- **Actor and Critic Losses:** The actor loss initially fluctuated during the early exploration phase (0–30 k steps) due to random policy sampling but gradually stabilized after approximately 50 k steps. Both critic networks exhibited highly consistent loss trajectories that converged toward near-zero values, confirming stable Q-function learning and the absence of value divergence. The convergence of critic 1 and critic 2 also implies reliable minimization of the double-Q bias.
- **Gradient Norms:** The gradient norms of the actor and critic networks show that the initial gradients peaked sharply but rapidly decreased and stabilized thereafter. This pattern reflects effective suppression of gradient explosion through gradient clipping (threshold = 5.0 for critic, 1.0 for actor) and adaptive learning rate scheduling. The overall decline in gradient magnitudes indicates the diminishing updates as the policy converged.

- Entropy Regularization: The α -loss exhibited transient oscillations before converging to zero, while $\log \alpha$ gradually stabilized at around -3.0 . This corresponds to a steady adaptive temperature that regulates the balance between exploration and exploitation. The result confirms that the automatic entropy tuning mechanism in SAC functioned properly, allowing the agent to maintain an adequate level of policy stochasticity throughout training.
- Entropy–Target Gap: The entropy closely followed the target entropy, and the entropy gap converged near zero after ~ 40 k steps. This demonstrates that the stochastic policy successfully adjusted its entropy level to meet the target distribution, implying stable entropy calibration and reliable exploration behavior.

5.4. Model Validation

5.4.1. Indoor Comfort of Buildings in Terms of PMV Index and Temperature

Figure 10 shows the results of indoor thermal comfort during training and evaluation using the SAC DRL model. To assess comfort during working hours, only the data between 09:00 and 18:30 were used. The pink line denotes the PMV comfort boundaries of the building defined earlier (-1.0 to 0.2 index). The black line shows the average indoor temperature across the eight zones, while the gray-shaded area indicates the comfort range requested by the owner. The line graphs show the daily average PMV for each zone, and the dots represent the daily average temperature for each zone. Here, temperature reflects indoor management based on PMV and Humidex, which enables the evaluation of whether the owner's comfort requirements are met. During training, PMV values mostly remained within the defined comfort range, though minor out-of-range cases were observed at some points. The three gray dashed lines indicate the time block average temperatures: ① morning (09:00–12:00), ② afternoon (12:00–15:00), and ③ late afternoon (15:00–18:30).

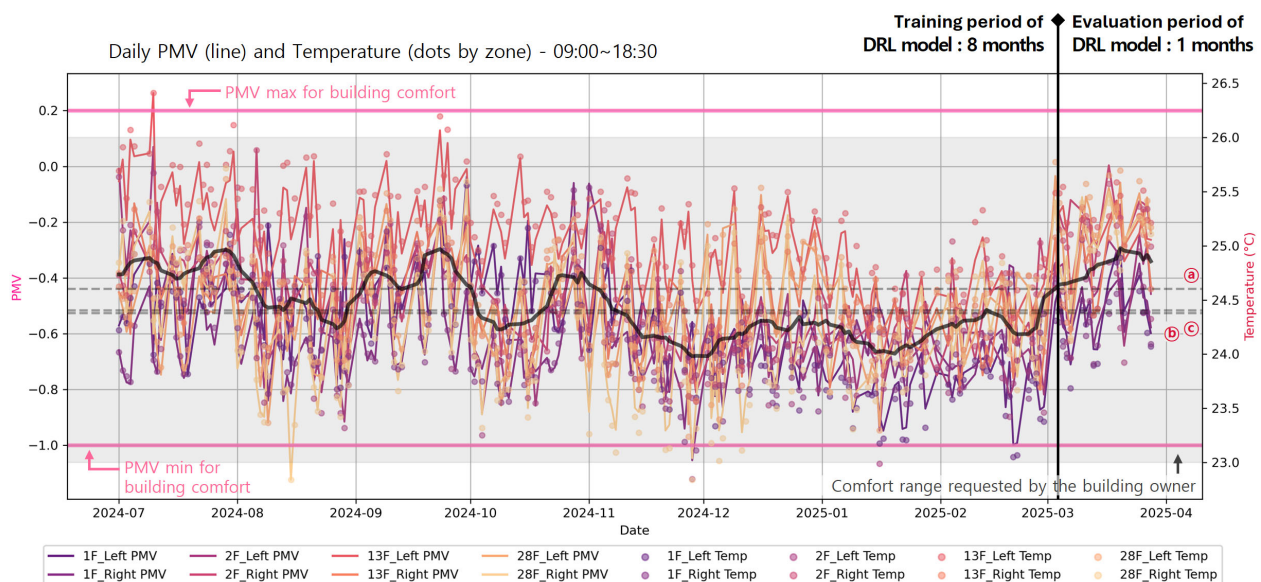


Figure 10. Trend in indoor comfort (PMV and temperature) during training and evaluation periods.

As a result of the agent's AHU control, the average indoor temperature was maintained at 24.60 °C for time ①, 24.38 °C for time ②, and 24.41 °C for time ③. Because the proportion of AHU ON states exceeds that of OFF states, the indoor temperature gradually decreases over time. This behavior indicates, as also shown in Figure 2, that the control strategy closely resembles actual building operation. The zone- and time-specific average indoor temperature and PMV are summarized in Table 6. At the initial operating point of 07:30, both PMV and temperature fall outside the comfort range. In the initial period, the left

and right zones of the 1st floor and the left zone of the 2nd floor met the temperature comfort range; however, as shown in Table 7, the high relative humidity results in PMV values outside the range. The fact that the average indoor temperature between 09:00 and 12:00 stays within the comfort range shows that the indoor environment was effectively regulated through PMV before working hours.

Table 6. Results of DRL (1 July 2024–28 February 2025): indoor temperature and PMV.

Time/ Zone	Average Temperature (°C)/PMV								Avg.
	1F_L	1F_R	2F_L	2F_R	13F_L	13F_R	28F_L	28F_R	
07:30	26.06/ 0.31	25.92/ 0.24	26.00/ 0.29	26.20/ 0.35	26.84/ 0.53	26.25/ 0.34	26.64/ 0.47	26.26/ 0.39	26.27/ 0.36
09–12	24.45/ −0.46	24.36/ −0.50	24.34/ −0.50	24.59/ −0.43	25.12/ −0.28	24.67/ −0.40	24.78/ −0.38	24.50/ −0.45	24.60/ −0.43
12–15	24.13/ −0.60	24.00/ −0.66	24.19/ −0.58	24.47/ −0.49	25.18/ −0.28	24.45/ −0.51	24.45/ −0.59	24.18/ −0.59	24.38/ −0.53
15–18:30	24.00/ −0.66	23.81/ −0.74	24.18/ −0.60	24.64/ −0.44	25.25/ −0.26	24.36/ −0.56	24.23/ −0.63	24.00/ −0.67	24.30/ −0.57

Table 7. Results of the DRL model (1 July 2024–28 February 2025): indoor humidity and Humidex.

Time/ Zone	Average Humidity (°C)/Humidex								Avg.
	1F_L	1F_R	2F_L	2F_R	13F_L	13F_R	28F_L	28F_R	
07:30	85.2/ 36.4	84.4/ 36.0	85.8/ 36.4	83.1/ 36.3	75.8/ 36.2	80.6/ 36.0	78.7/ 36.3	85.3/ 36.9	82.4/ 36.3
09–12	67.8/ 30.4	66.8/ 30.1	67.9/ 30.3	65.5/ 30.3	59.1/ 30.0	64.4/ 30.2	61.3/ 29.9	66.8/ 30.3	64.9/ 30.2
12–15	64.9/ 29.4	63.8/ 29.0	65.2/ 29.6	63.0/ 29.6	56.9/ 29.7	61.5/ 29.3	58.3/ 28.8	63.8/ 29.3	62.2/ 29.3
15–18:30	63.5/ 28.9	62.3/ 28.5	63.6/ 29.3	62.1/ 29.8	55.4/ 29.6	60.0/ 28.9	56.6/ 28.2	62.3/ 28.7	60.7/ 29.0

Using the DRL model trained over eight months, its application to a new period in March that was not included in training shows that the PMV values in all zones remained within the comfort range. In addition, the indoor temperature was consistently maintained within the requested range, as illustrated in Figure 10. The zone-specific distributions of PMV and temperature are provided in Figure A2.

From Figure 10, the similarity of the PMV average line, the temperature scatter plot, and the average temperature line within the comfort range can be seen, with temperature exerting a relatively strong influence on PMV. This confirms that temperature control is the key factor in maintaining indoor comfort.

5.4.2. Indoor Comfort of Buildings in Terms of Humidex Index and Humidity

Figure 11 illustrates the results of indoor humidity comfort during the training and evaluation periods using the SAC DRL model. The blue lines denote the Humidex comfort boundaries (27–33), and the black line represents the average indoor humidity across the eight zones. The gray area indicates the humidity range observed during working hours in the actual data, set between 50% and 80%. The line graphs show the daily average Humidex for each zone, while the dots indicate the daily average relative humidity for

each zone. The three gray dashed lines represent the time-of-day average relative humidity across all zones.

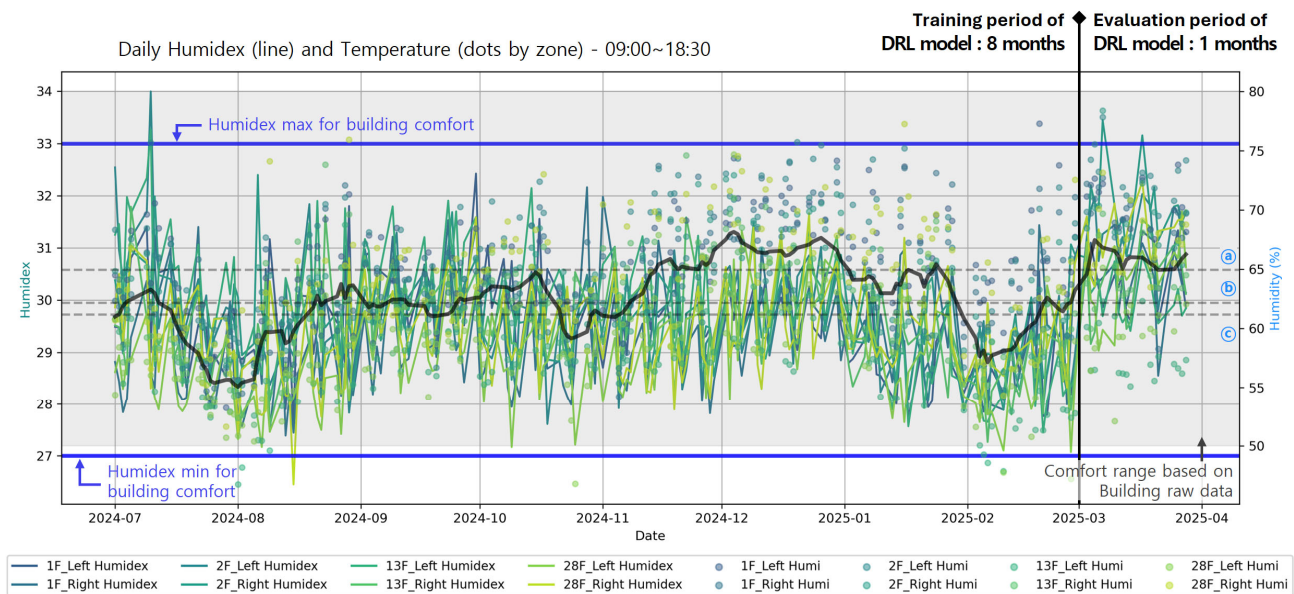


Figure 11. Trend of indoor comfort (Humidex and humidity) during the training and evaluation periods.

The average relative humidity was controlled at 64.94% for ①, at 62.17% for ②, and at 60.74% for ③. The average humidity and Humidex values for each zone and time are shown in Table 4. At 07:30, both average relative humidity and Humidex fell outside the comfort range. Although the relative humidity in the right zone of the 13th floor and the left zone of the 28th floor were close to the comfort range, Table 3 shows that the high indoor temperature resulted in Humidex values outside the comfort range. From 09:00 to 12:00, both the average relative humidity and Humidex remained within the comfort range, which indicates that humidity-related comfort was effectively controlled. In Figure 11, the distribution of the humidity scatter plot shows patterns that are either similar to or higher than the Humidex average lines. The reasons for these results are given in Section 5.4.5.

The application of the DRL model trained over eight months to March shows that the Humidex stayed within the comfort range on most days. However, the right zone of the 2nd floor produced values outside the comfort range in early and mid-March. This was found to be due to switching the AHU mode to OFF during time ③ to save energy. Consequently, as the end of the workday approached, the indoor relative humidity rose compared with other days. The zone-specific distributions of Humidex and humidity are provided in Figure A3.

In Tables 6 and 7, the left zone of the 28th floor and the right zone of the 2nd floor both show the same initial Humidex value (36.3 at 07:30) but display contrasting environmental characteristics. The relative humidity in the 2nd floor right zone is 4.4% higher than in the 28th floor left zone, while the indoor temperature is 0.44 °C lower. In this case, the PMV is predicted to be higher in the 28th floor left zone due to its higher temperature, making it the least comfortable space. During the 09:00–12:00 period, the right zone of the 1st floor and the left zone of the 2nd floor both show a PMV of −0.50 and similar indoor temperatures. However, in the same period, the relative humidity in the 1st floor right zone is 1.1% lower than in the 2nd floor left zone, resulting in a lower Humidex value. Consequently, when the Humidex is the same, the zone with a higher temperature has a higher PMV and is therefore less comfortable, whereas when the PMV is the same, the zone with a higher humidity has a higher Humidex and is therefore less comfortable. These findings indicate

that indoor comfort is more strongly influenced by temperature, and when temperature levels are the same, humidity becomes the differentiating factor.

5.4.3. Relationship Between Indoor Comfort and Energy Savings

In this section, the training results on indoor comfort and energy saving rates are analyzed. Figure 12 shows the daily extent and frequency of out-of-range cases across the eight zones over the eight-month training period, together with the corresponding daily energy saving rates. The direction of out-of-range conditions (whether toward hot or cold) is not distinguished. The daily average comfort across the eight zones is calculated using Equations (29) and (30). Out_{avg} represents the proportion of out-of-range cases of PMV or Humidex among the eight zones $|Z|$ during working hours (09:00–18:30), over the total time steps $|T|$. The zone-specific out-of-range instance, $non_{z,t}$, is counted as 1 if either the PMV or Humidex in zone z at time t falls outside the comfort range, and 0 if both indices remain within the range. Because the agent can take an action every 15 min, a total of 312 data points per day are considered. Energy savings are calculated using Equation (3).

$$Out_{avg} = \frac{1}{|T|} \sum_{t \in T} \left(\frac{1}{|Z|} \sum_{z \in Z} non_{z,t} \right), \quad (29)$$

$$non_{z,t} = \begin{cases} 1, & \text{if } (PMV_{z,t} \notin [-1.0, 0.2]) \text{ or } (Humidex_{z,t} \notin [27, 33]) \\ 0, & \text{otherwise} \end{cases}, \quad (30)$$

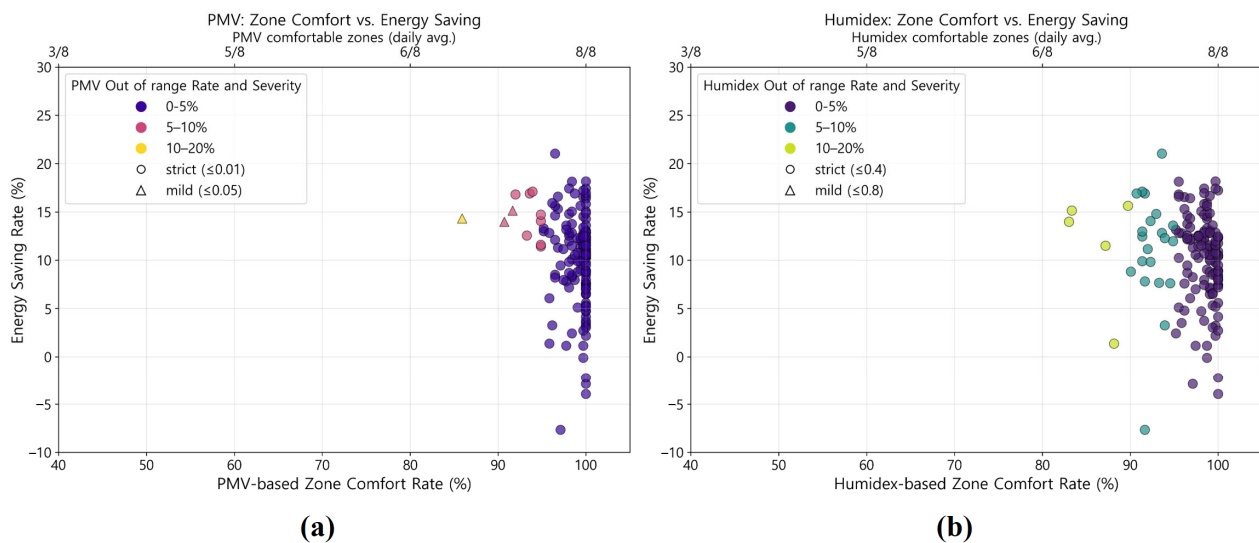


Figure 12. (a) Relationship between daily average PMV and energy saving rate for the eight zones; (b) relationship between daily average Humidex and energy saving rate for the eight zones.

As shown in Figure 12a,b, the daily average out-of-range cases across the eight zones generally fall within 0–5%. This corresponds to 1–15 of the 312 daily data points where either PMV or Humidex did not meet the comfort range. The daily violation levels indicate that both indices remain within strict comfort criteria. This demonstrates that the comfort ranges defined by PMV and Humidex were successfully maintained in most zones during working hours. At this point, the strict criterion for PMV is defined as the change in PMV caused by a 0.1 °C variation in temperature under constant Humidex, based on training data. The strict criterion for Humidex is defined as the change in Humidex caused by 1% variation in relative humidity under constant PMV. Examining the relationship between PMV, Humidex, and the energy saving rate shows that Out_{avg} increases with the energy saving rate. This highlights how the balance between energy savings and indoor comfort

should be set. When the energy saving rate exceeds 10%, six of the eight zones remain within the comfort range on average, while two zones fall outside. In addition, 5–20% of the daily working hours fall outside the comfort range. For energy saving rates below 10% and for negative saving intervals where more energy was consumed than the baseline, all eight zones remained within the comfort range. Comparing violation frequencies shows that PMV was more strictly controlled than Humidex.

To quantify the relationship between energy savings and indoor comfort, Pareto frontier, sensitivity analysis, and trade-off curve analyses were performed, as shown in Figure 13. The trade-off curve in Figure 13 was constructed using the last 167 episodes after 16,700 training episodes, representing the states visited by the converged SAC policy. The Pareto-front-based method seeks a set of non-dominated solutions, where no objective can be further improved without degrading at least one other objective [127]. Instead of applying mathematical optimization methods, we filtered samples with both high comfort and high energy-saving performance using the recorded PMV and Humidex values after training. The Pareto frontier reveals a range of operating points that balance conflicting objectives between comfort and energy efficiency. As illustrated in Figure 13a, nine Pareto-optimal solution sets were identified. The combined comfort index (PMV + Humidex) varied from 95.03% to 100%, while the energy-saving rate ranged from 4.15% to 21.05%, indicating substantial trade-off dynamics. Among these, six operating policies achieved energy savings while maintaining comfort within the acceptable PMV and Humidex ranges. The Pareto slope (-2.047) indicates that a 1% decrease in comfort corresponds, on average, to a 2% improvement in energy-saving performance. Beyond point P8, comfort continues to decline, but energy savings show minimal further improvement, suggesting a limit in the efficiency–comfort trade-off.

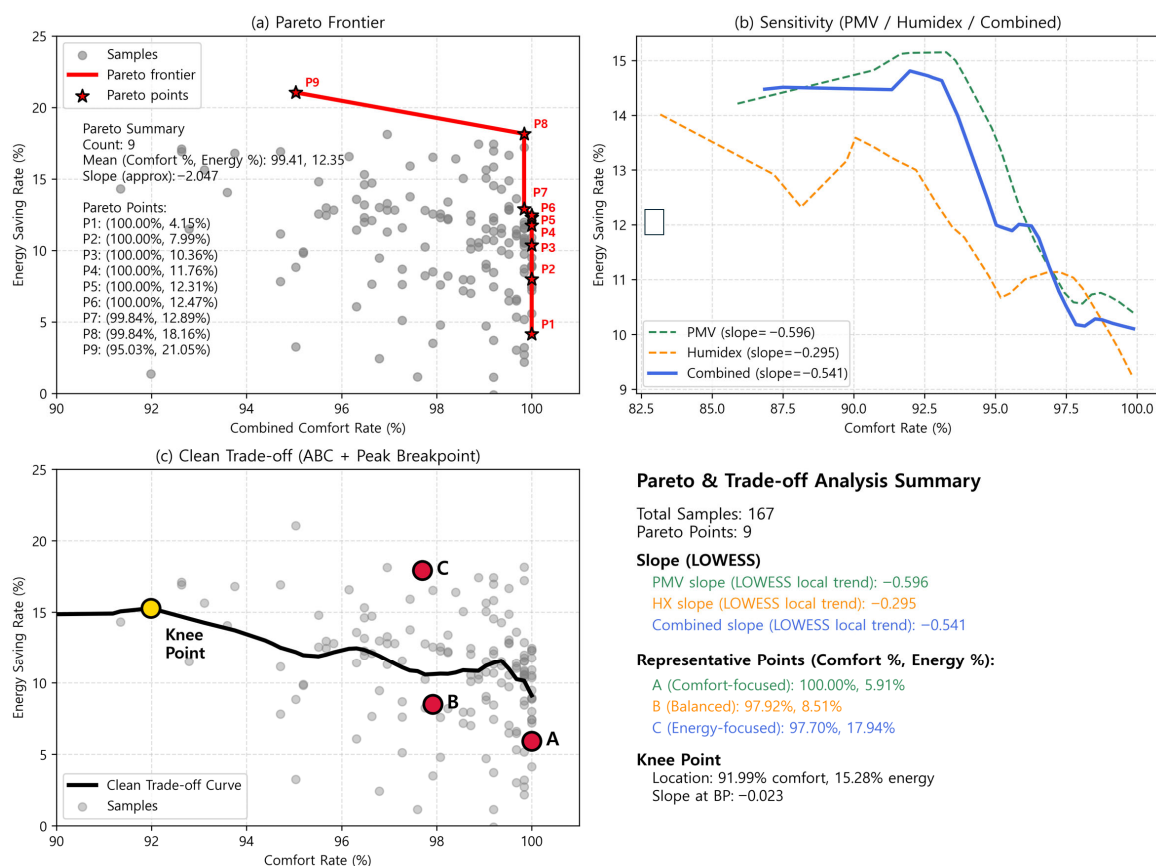


Figure 13. Trade-off relationship between energy savings and indoor comfort.

Sensitivity analysis was conducted to examine how changes in comfort affect variations in the energy-saving rate. Figure 13b illustrates how the energy-saving rate responds to decreases in comfort, showing that the PMV index (−0.596) has a stronger influence on energy savings than the Humidex index (−0.295). This indicates that energy savings are more sensitive to PMV than to Humidex. The sensitivity analysis was performed using Locally Weighted Scatterplot Smoothing (LOWESS), a form of local weighted regression. LOWESS performs regression only within a local neighborhood of each point, assigning higher weights to nearby observations and lower weights to distant ones, thereby capturing local trends in the data. Because the relationship between energy and comfort is nonlinear and may vary across time segments due to weather and operational conditions, local regression was used instead of global regression to better capture these localized patterns [128]. For this reason, the slope presented in Figure 13b should be interpreted as a trend, rather than an absolute energy-saving rate per unit of comfort change.

Figure 13c shows the trade-off curve that characterizes the relationship between indoor comfort and the energy-saving rate. A represents a comfort-oriented operation strategy, B reflects a balanced strategy between comfort and energy savings, and C corresponds to an energy-saving-oriented operation strategy. These points (A, B, and C) represent representative building operation strategies depending on the priority of the decision-maker. The black curve corresponds to the nonlinear trend obtained using LOWESS-based local regression, which smooths the average behavior of data segments to reveal the underlying pattern. From the trade-off curve, we observe that when comfort exceeds approximately 92%, additional increases in comfort generally reduce the energy-saving rate. Near 99% comfort, a sharp drop in energy savings is observed. A knee point appears at around 92% comfort, which represents the turning point where the marginal efficiency between comfort and energy savings changes most significantly [129,130]. This point indicates where the reduction in energy savings per unit increase in comfort becomes the largest. In other words, it identifies the region where the system no longer improves efficiently. Managing comfort at the knee point ($\approx 92\%$) enables an energy-saving rate of approximately 15%, making it an economically advantageous operating point in terms of efficiency. However, the knee point is not a global optimum; it is a local optimum derived from the curvature of the nonlinear trade-off curve. While point C yields the highest energy-saving performance overall, it does not represent the efficiency turning point. Instead, C highlights maximal energy savings, whereas the knee point indicates where the system is most sensitive to efficiency changes. To identify this trade-off knee point, we computed the curvature of the LOWESS-smoothed energy–comfort curve. The curvature was derived using Equations (31)–(34).

$$f'(x_i) \approx \frac{y_{i+1} - y_{i-1}}{x_{i+1} - x_{i-1}}, \quad (31)$$

$$f''(x_i) \approx \frac{\frac{y_{i+1} - y_i}{x_{i+1} - x_i} - \frac{y_i - y_{i-1}}{x_i - x_{i-1}}}{x_{i+1} - x_{i-1}}, \quad (32)$$

$$k_i = \frac{f''(x_i)}{(1 + (f'(x_i))^2)^{\frac{3}{2}}}, \quad (33)$$

$$i_{\text{kneepoint}} = \arg \max_i k_i, \quad (34)$$

where $f'(x_i)$ and $f''(x_i)$ denote the first and second derivatives of the smoothed curve. Through the first derivative, the slope at each point is obtained, and the second derivative captures changes in that slope. The curvature k_i is computed using these derivatives. The

point with the maximum curvature corresponds to $i_{\text{kneepoint}}$, where the marginal gain in energy savings rapidly diminishes relative to the decrease in comfort.

5.4.4. Multiple Perspectives on the Training Results

Figure 14 shows the analysis of the deep reinforcement learning model results from three different perspectives. A total of 7515 raw data points from the training period and 810 raw data points from the evaluation period were used. The circular markers represent results from the training period, while the star-shaped markers represent results from the evaluation period. Panels (a), (c), and (f) show the time-of-day distributions of PMV, while panels (b), (d), and (g) present the time-of-day distributions of Humidex.

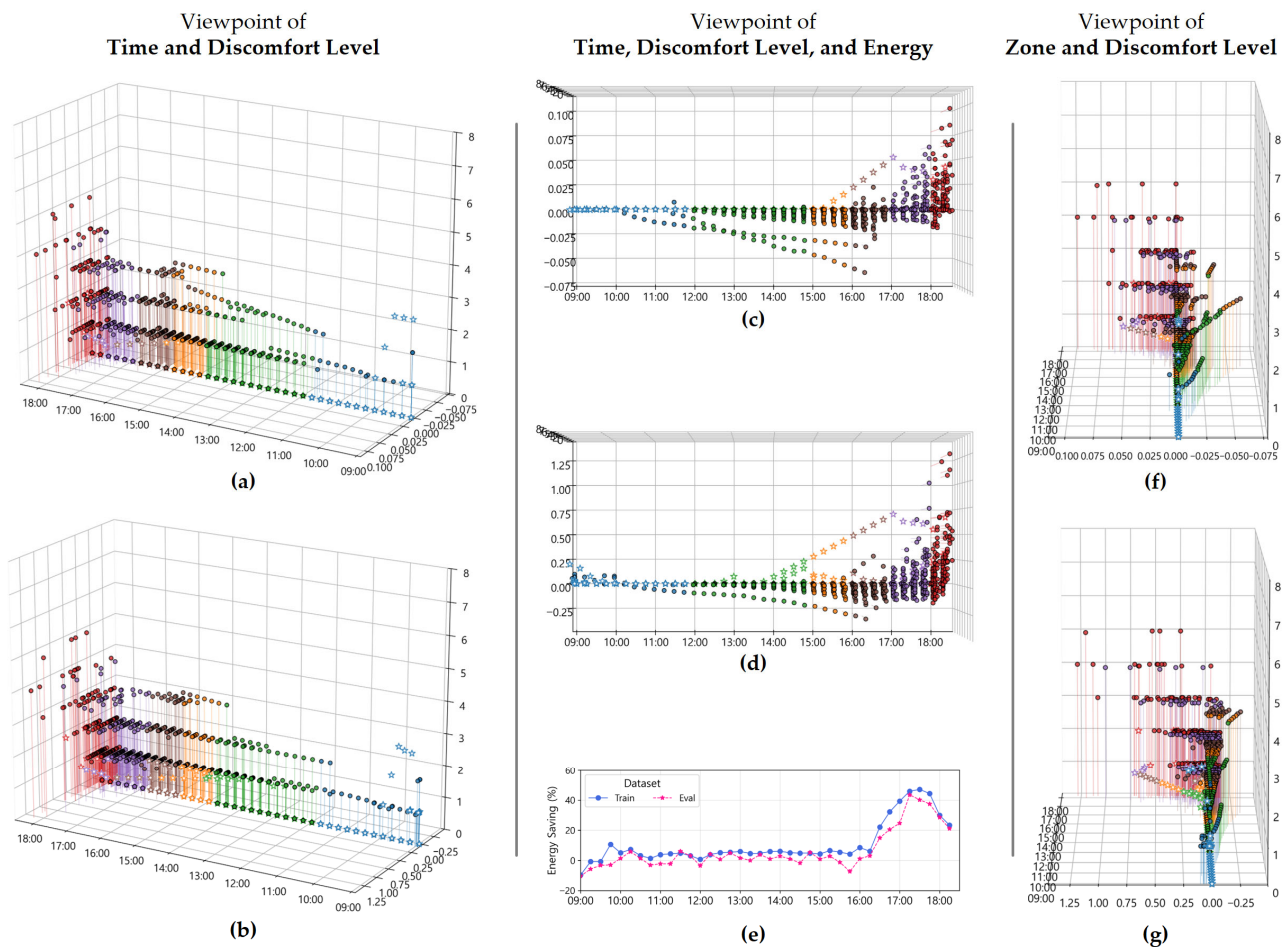


Figure 14. A three-dimensional interpretation of DRL model training and evaluation; (a,c,f) PMV violation graphs by time zone; (b,d,f) Humidex violation graphs by time zone; (e) Energy saving rate by time zone; Time-period colors: Morning 9–12 (blue), Lunch 12–15 (green), After 15–16 (orange), After 16–17 (brown), After 17–18 (purple), After 18–18:30 (red). Circle markers represent training data, and star markers represent evaluation data.

From panels (a) and (b), it is observed that as the workday ends, the number of zones falling outside the comfort range gradually increases. After 18:00, PMV exceeded the range by up to 0.1 in the warmer direction, while Humidex exceeded it by up to 1.25 in the more humid direction during certain times.

Panel (c) represents a top view of panel (a), and panel (d) a top view of panel (b). Through these, the AHU control policies learned by the DRL model can be examined in more detail. In panel (c), PMV reaches the comfort range by 09:00, while in panel (d), Humidex converges slightly later. Up to 15:00, both training and evaluation results generally

stay within the comfort range. However, as the day progresses, both PMV and Humidex tend to deviate negatively. This suggests that the model learned that maintaining operation (even with penalties) was preferable to switching the AHUs OFF. After 17:00, negative out-of-range cases decreased, and positive deviations appeared. In the evaluation results, such occurrences were observed at times earlier than 17:00, unlike in the training results. This can be explained by panel (e), which presents the time-of-day average energy saving rate. Between 16:00 and 18:30, the results indicate that the DRL model learned that it was more advantageous to reduce energy consumption by switching AHUs OFF, even at the cost of minor comfort violations in the warmer direction.

Panels (f) and (g) show both the extent of out-of-range occurrences of indoor comfort and the number of zones affected simultaneously. The results indicate that in most cases, only one zone fell outside the comfort range during morning and lunchtime periods. From the afternoon, two to three zones were affected, and on some days up to five zones were affected after 18:00. Consequently, the DRL model emphasized indoor comfort earlier in the day, but after 15:00, control shifted toward energy savings over comfort.

The PMV-PPD index derived for each space from these results is illustrated in Figure A4. Rather than adopting the typical PMV comfort range of -0.5 to $+0.5$, we defined a customized range of -1.0 to $+0.2$ based on the analyzed indoor environmental data and the local office culture in Malaysia. As a result, the PPD curve is observed to be slightly skewed to the left rather than symmetrical. This PMV range can be adjusted by analyzing the difference between the predicted dissatisfaction (PPD) and the perceived comfort level through an on-site occupant survey after model application, and adjusting the PPD range by adjusting the gap between the predicted value and the actual occupant experience [131].

5.4.5. Conditional Linear Regression Analysis

A regression analysis was conducted to better understand the interaction between temperature and humidity. Regression analysis examines relationships among variables to predict how changes in one variable affect another. In this study, the relationship between temperature and humidity is quantified under the condition that the Humidex stays within the comfort range. As noted in Section 1.3.2, perceived comfort varies depending on humidity, even when temperature remains the same. Likewise, PMV and Humidex assess comfort differently, although both incorporate temperature and humidity. Consequently, identical PMV or Humidex values may correspond to different temperature/humidity conditions. In this study's DRL model, temperature and humidity are not used directly in the reward function; only PMV and Humidex must meet the comfort criteria. To satisfy them, various combinations of temperature and humidity may arise. A linear regression analysis was therefore conducted to examine how the DRL agent satisfies the PMV and Humidex. The regression is quantified using Equation (35). Rather than the entire dataset, the sample is restricted to specific Humidex intervals. As with Sections 5.4.1 and 5.4.2, the regression analysis divides working hours into three time blocks, and data were extracted for cases where Humidex falls within the range of 27–33.

$$\text{Humidity}_Z(t|s) = \beta_{0,z,s} + \beta_{1,z,s}\text{Temp}_z(t) + \epsilon_{z,s}(t), \text{ Humidex} \in [27, 33]. \quad (35)$$

Here, $\text{Humidity}_Z(t|s)$ denotes the humidity at time t in zone z . $\beta_{0,z,s}$ is the baseline level, indicating the average starting point of humidity in percentage terms. $\beta_{1,z,s}$ represents the average sensitivity of humidity to temperature changes in a given zone during the three working hour periods s within the total working hours t . This coefficient indicates how much humidity changes, on average, for a 1°C change in temperature. $\epsilon_{z,s}(t)$ denotes the error term with a mean of zero. Figure 15 illustrates the relationship between temperature and humidity changes when the Humidex index remains within the comfort range.

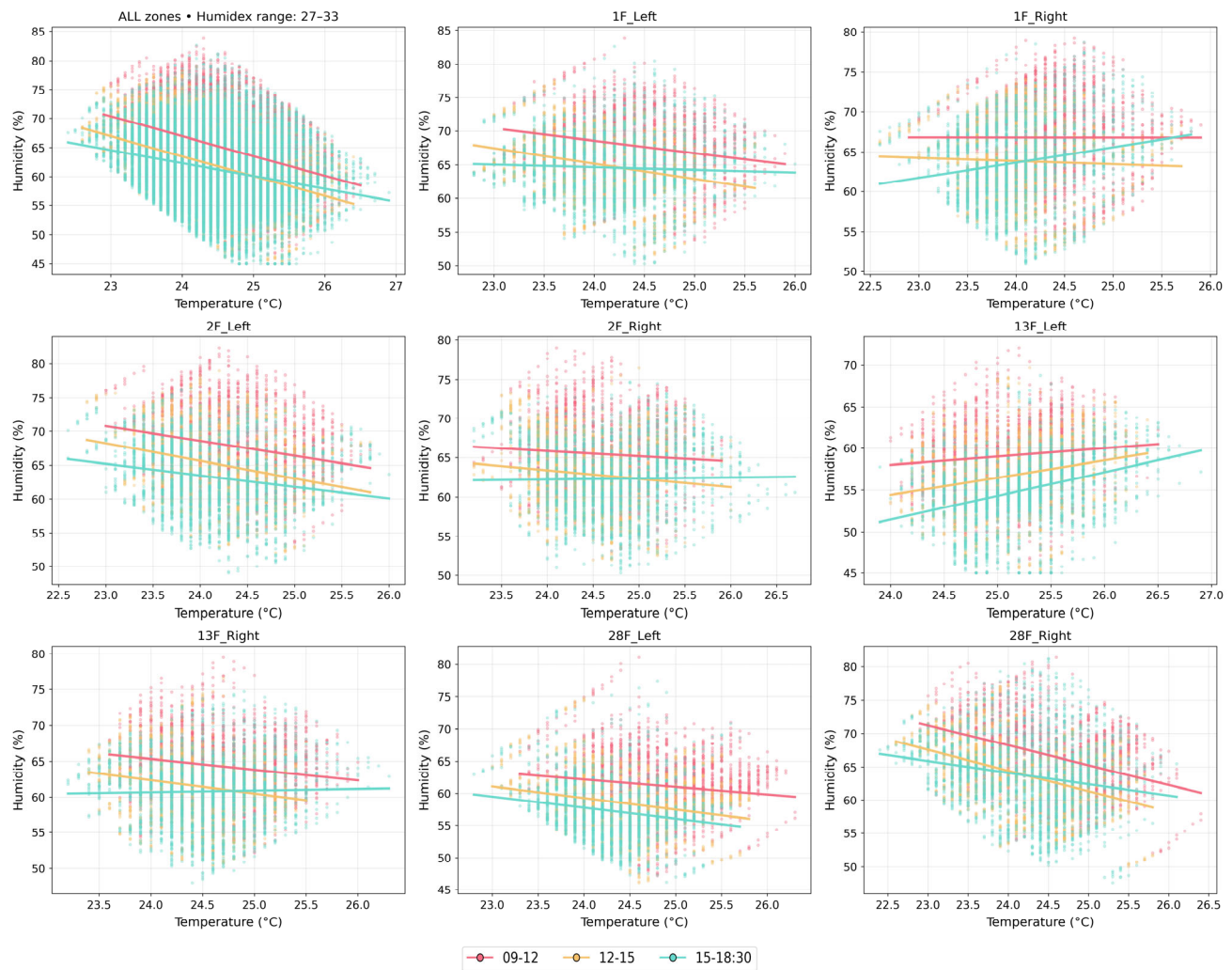


Figure 15. Linear relationship between temperature and humidity comfort range.

The slopes in Figure 15 confirm a negative correlation, with humidity decreasing as temperature increases. These results indicate that the DRL model regulated temperature and humidity in opposite directions to maintain the comfort ranges of both PMV and Humidex. Unlike other zones, the 13th floor left zone shows a positive correlation between temperature and humidity. As shown in Tables 6 and 7, this zone records the highest average temperature and the lowest humidity among the eight zones. For this case, the DRL model appears to have learned to maintain PMV just below the comfort boundary after 09:00, while reducing Humidex.

On average, as the day progresses, the slopes of the regression lines for the morning (blue) and afternoon (orange) periods either remain similar or become steeper. In contrast, the regression line for the evening period (green) shows the opposite slope. This outcome suggests that the DRL model scheduled AHU control primarily to satisfy indoor comfort during the morning and afternoon, while shifting toward energy-saving operation at lunchtime and in the evening before closing. Consequently, the analysis indicates that the DRL model converged to a policy that regulates temperature and humidity in opposite directions to satisfy the comfort ranges of PMV and Humidex, while applying time-specific strategies to achieve energy savings.

5.4.6. Carbon Permit Revenue

Figure 16 illustrates the DRL model results for carbon permit trading. The upper panel shows the DRL selling policy, while the lower panel compares nine alternative models. The number of CPs that can be sold per day is capped at three, since large fluctuations directly affect the reward function and hinder convergence. Therefore, the carbon permit selling policy is designed to maximize profitability under this restriction. Permit generation is calculated using Equations (17) and (24).

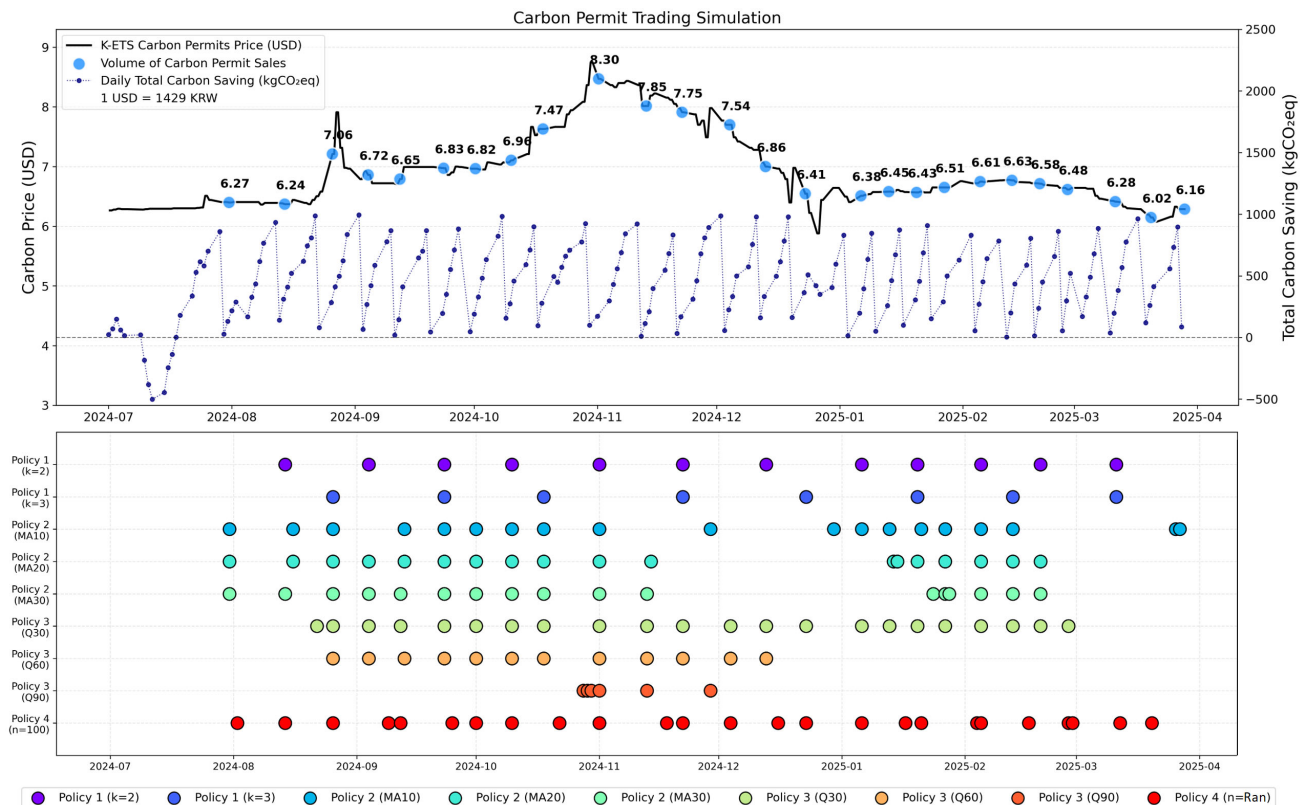


Figure 16. Training and simulation results of carbon permit sales in the DRL model.

Circle markers indicate the times when carbon permits were sold, while dashed circles mark the end of each day. For carbon permit trading, the K-ETS prices from the Korean market were applied. As a result of implementing energy-saving policies while maintaining indoor comfort, the average cycle required to generate one carbon permit through energy savings was about 8–9 days. Although only ten AHUs were included for training and evaluation in this study, applying the model to all 60 AHUs in the building is expected to shorten the permit generation cycle.

Next, we evaluated whether the DRL trading policy in Figure 16 achieved the highest revenue. For comparison, the trained policy was tested against nine arbitrary benchmark policies. These include conditional selling, moving average, quantile-based, and random strategies. The conditional approach predetermines the number of CPs to sell before trading. The moving average approach sells only when the price exceeds the average price. The quantile-based approach sells only when the permit price meets or exceeds the upper quantile threshold. The random approach executes sales at random. The conditional approach corresponds to Policy 1. Because the agent strategy converged to selling one permit at a time, Policy 1 was designed to sell only when at least two or three permits had been accumulated. Policy 2 applies the moving average approach, where sales are triggered by comparing prices with short-, medium-, or long-term moving averages.

The short-term benchmark is the 10-day moving average, the medium-term the 20-day moving average, and the long-term the 30-day moving average. The DRL model itself employs a 5-day moving average, selling when the price is above this threshold. Policy 3 corresponds to the quantile-based approach, in which sales are made when the unit price of a carbon permit exceeds the 30th, 60th, or 90th percentile thresholds. Policy 4 represents the random approach, under which sales are executed without restriction. Table 8 summarizes comparative results for the DRL model and the benchmark selling policies.

Table 8. Results of the carbon permit trading policy of the DRL model and nine selling policies.

Policy	Total Carbon Permit Unit	Sold Carbon Permits	Total Revenue (USD)/(Rank)	Average Carbon Price (ACP)/(Rank)	Weighted Average Carbon Price (WACP)/(Rank)	Coefficient of Variation (CV)/(Rank)
DRL	26	26	176.24 (1)	6.92 (7)	0.55 (9)	2.49 (1)
Policy 1 (k = 2)	26	24	167.27 (5)	6.94 (6)	0.58 (5)	3.82 (7)
Policy 1 (k = 3)	26	24	168.04 (4)	6.98 (4)	0.60 (4)	4.73 (9)
Policy 2 (MA10)	26	25	172.17 (3)	6.86 (9)	0.53 (10)	3.33 (4)
Policy 2 (MA20)	26	22	152.22 (7)	6.73 (10)	0.57 (7)	3.42 (5)
Policy 2 (MA30)	26	22	152.22 (8)	6.90 (8)	0.58 (6)	3.53 (6)
Policy 3 (Q30)	26	23	161.33 (6)	7.02 (3)	0.61 (3)	2.78 (3)
Policy 3 (Q60)	26	14	103.25 (9)	7.36 (2)	0.78 (2)	4.22 (8)
Policy 3 (Q90)	26	12	100.07 (10)	8.37 (1)	0.96 (1)	6.18 (10)
Policy 4 (Random)	26	25	173.55 (2)	6.95 (5)	0.56 (8)	2.55 (2)

Total carbon permit unit indicates the number of CPs generated over the entire period, while “sold carbon permits” refers to the number sold under each policy. In some policies, not all generated permits were sold because the conditions were not met. Total revenue represents the final earnings, and the average carbon price (ACP) is defined as total revenue divided by the number of permits sold, representing the average selling price per permit. In terms of total revenue, the DRL model achieved the best performance, whereas for ACP, Policy 3 (Q90)—which yielded the lowest total revenue—showed the highest performance. The weighted average carbon price (WACP) is calculated using Equation (36), where the price of each sold permit is weighted by its share during the selling period. A value closer to 1 indicates that sales occurred mainly in higher price ranges. The DRL model completed most of its sales in the upper 55% of the price distribution, ranking ninth among the policies in terms of quantile-based performance.

$$WACP = \frac{\sum (w_t) \times (rev_{day})}{\sum rev_{day}}. \quad (36)$$

Here, rev_{day} is the total revenue generated on days when carbon permit sales occurred, and w_t represents the price quantile of a single carbon permit at the time of sale. In Figure 16, the price distribution w_t ranges from USD 5.88 to USD 8.79. The WACP value of the DRL model ranked sixth among the ten policies, indicating that sales were made at relatively higher prices on average. The reason for not achieving the highest prices is attributed to the relatively uniform selling pattern. The coefficient of variation (CV) is defined as the standard deviation divided by the mean, which allows for comparison between how

consistently each policy generated revenue. Equation (37) expresses this ratio of standard deviation to mean.

$$CV = \frac{\sqrt{\sum_{day=1}^n \left(rev_{day} - \sum_{day=1}^n rev_{day} / n \right)^2 / (n - 1)}}{\left| \sum_{day=1}^n rev_{day} / n \right|}. \quad (37)$$

Here, n is the total of 185 days, combining both the training and evaluation periods. The average number of carbon permits sold over 185 days is divided by the standard deviation over the same period. The DRL model recorded sales on 26 out of 185 days, which results in an average of 0.97 and a standard deviation of 2.42. Accordingly, the CV was the lowest among all policies, which means that the DRL model showed the most stable carbon permit trading performance. Consequently, the DRL model converged to a strategy that prioritized stability over maximizing the average unit price or concentrating sales in high-price intervals. This result demonstrates that the DRL-based carbon permit trading policy is robust.

5.5. Comparative Analysis Before and After Applying DRL

This section compares two indicators of indoor comfort—temperature and humidity—before and after applying the DRL model to the target building. It also evaluates the proposed indoor comfort management model, which combines PMV and Humidex, against an alternative model that applies only PMV without Humidex. Finally, it compares the total building operating costs over the entire nine-month period, including both training and evaluation phases.

5.5.1. Comparison of Temperature, Humidity, PMV, and Humidex

Figure 17 compares the PMV and Humidex results before and after applying the DRL model, which shows whether the model effectively managed indoor temperature and humidity at appropriate levels and whether it improved indoor comfort. For comparison, weekends and holidays were excluded, and the analysis was limited to 09:00–18:30. Control represents the baseline schedule applied by the building manager. DRL (PMV + Humidex) denotes the proposed model in this study, which uses both indices to predict and evaluate indoor comfort. DRL (PMV) refers to the model that applies only the PMV index, which is commonly used to evaluate indoor comfort. In this case, Humidex was removed only from the reward system but remained in the state variables. Table 9 defines the abbreviations: C denotes Control, D1 denotes DRL (PMV + Humidex), and D2 denotes DRL (PMV).

When comparing PMV results before and after applying the DRL model, Control (C) exhibited the largest variability across all zones. The interquartile range (IQR) was 0.5, the widest among the cases, with major fluctuations observed on the 1st and 28th floors. The variation range did not consistently remain within the comfort zone, and the deviation (the difference between the IQR and the whisker low/high) was also substantial in all zones. In contrast, both DRL models showed lower variability and smaller IQR values than Control. While they did not always remain strictly within the comfort zone, the deviations were acceptable. For Control, zone-specific PMV values varied widely, making it difficult to guarantee comfort. By contrast, D1 and D2 maintained PMV within -0.5 to -0.4 , which can be interpreted as more reliable indoor comfort.

From the perspective of PMV, both D1 and D2 maintained Q1 and Q3 within the comfort range. However, in D2, the first-floor right zone failed to maintain comfort, whereas D1 consistently maintained comfort across all zones. The mean and median values of D1 and D2 were at similar levels, indicating comparable average PMV performance.

However, whisker low, whisker high, Q1, Q3, and IQR values showed greater variability, which means that D2 provided a weaker assurance of comfort than D1.

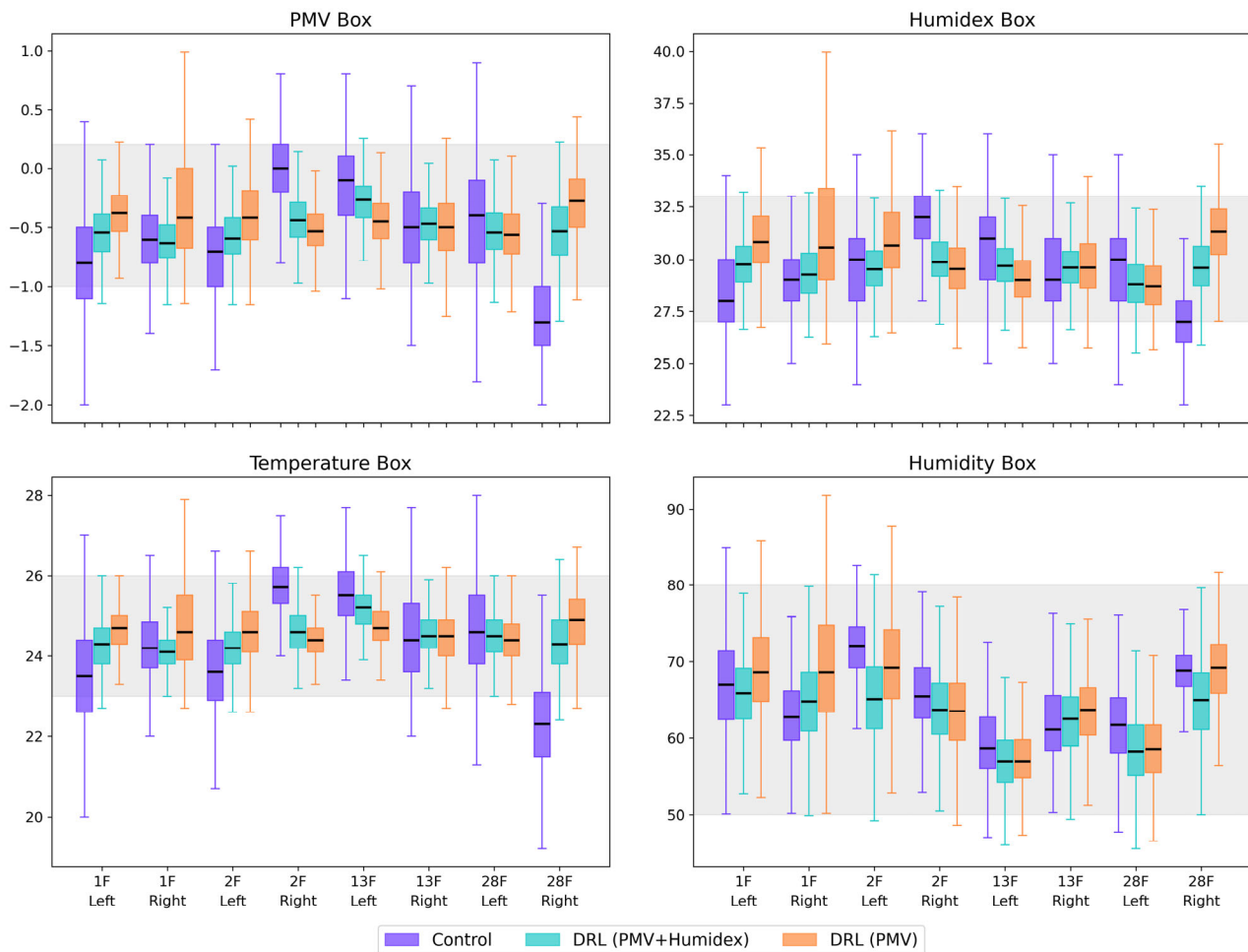


Figure 17. Comparison of indoor comfort between schedule-based control and the DRL model.

Table 9. Average statistics of eight zones under different building indoor operation policies.

Index	Whisker Low			Q1 (25%)			Median			Q3 (75%)			Whisker High			Inter-Quartile Range, (Q3-Q1)			Mean		
	C	D1	D2	C	D1	D2	C	D1	D2	C	D1	D2	C	D1	D2	C	D1	D2	C	D1	D2
PMV	−1.5	−1.1	−1.2	−0.8	−0.65	−0.61	−0.6	−0.50	−0.44	−0.3	−0.35	−0.24	0.5	0.09	0.32	0.5	0.30	0.38	−0.5	−0.48	−0.42
Humidex	24.6	26.3	26.1	28.1	28.7	29.0	29.5	29.5	30.0	30.8	30.4	31.4	34.4	33.0	34.9	2.6	1.73	2.38	29.8	29.6	30.2
Temp.	21.6	23.0	22.9	23.6	24.1	24.1	24.2	24.5	24.6	25.0	24.9	25.1	27.1	26.0	26.4	1.43	0.80	0.93	24.3	24.5	24.6
Humi.	52.5	49.1	50.7	61.7	59.3	61.2	64.7	62.9	64.8	68.2	66.2	68.7	78.0	76.4	79.9	6.56	6.86	7.50	65.3	62.8	65.1

On examining the Humidex index, C, D1, and D2 maintained Q1 and Q3 within the comfort range of 27–33. All three policies showed mean and median values around 29–30, indicating adequate comfort. This corresponds to the generally accepted range of 20–29, showing that no major discomfort occurred in terms of Humidex before or after DRL application. However, clear differences appeared in variability. C and D2 did not maintain uniform IQR values across zones and showed wider intervals extending beyond the comfort range. In contrast, D1 maintained the comfort range within acceptable limits. The whisker lows for C and D2 were generally lower than D1, while their whisker highs were higher. As with the PMV results, such high variability can undermine indoor assurance.

In terms of Humidex, D2 exhibited a larger IQR than D1, which indicates a greater number of values outside the comfort range. Although the mean and median values were

at comparable levels, zone-specific analysis revealed higher variability toward more humid conditions in the first floor, the left side of the second floor, and the right side of the 28th floor. This was further reflected in higher whisker highs than those of C, which suggests that tenants would perceive the indoor environment as more humid than usual. In contrast, D1 maintained a uniform Humidex distribution across all zones. However, because D2 relied only on PMV, which indicates consistent comfort levels across zones, there were still spatial discrepancies in Humidex. This implies that even if tenants experience similar thermal sensations, differences in perceived humidity may occur between zones and ultimately lead to variations in overall indoor comfort.

Third, when comparing temperature ranges, C, D1, and D2 all maintained mean, median, Q1, and Q3 values within the comfortable range of approximately 24 °C. Similarly to the PMV index, the baseline case C showed Q1 and Q3 values within the 23–26 °C range, but its overall variation was larger than in the other cases. A notable aspect is that temperature is evaluated as a single variable, whereas PMV is derived from multiple parameters. As a result, cases were observed where temperature remained within the comfortable range but PMV did not, and vice versa, with differing extents of out-of-range conditions. For example, in C, the right side of the first floor and the left side of the second floor had PMV values within the upper comfort limit, while the temperature exceeded it. Conversely, on the right side of the second floor, the PMV in D1 remained within the range, whereas in D2 it slightly fell below it. However, for temperature in the same zone, D1 exceeded the upper range, while D2 remained within it. These findings highlight the limitations of relying solely on temperature and support the use of PMV as a more comprehensive indicator of thermal comfort.

Temperature ranges for indoor operation are recommended differently depending on the country, season, or building. While these ranges also serve energy conservation, they are regarded as reasonably comfortable for occupants. For this reason, many buildings manage indoor conditions primarily through temperature control. As a result, PMV (0.99) and Humidex (0.91) were strongly correlated with temperature, as illustrated in Figure 3. This relationship is further confirmed in Figure 17, where zone-level comparisons of PMV, Humidex, and temperature reveal similar statistical patterns. As shown in the correlation analysis in Figure 3, Humidex is strongly influenced by temperature. However, from the humidity perspective, Humidex (0.69) responds more sensitively to variation than PMV (0.44).

Finally, comparing humidity before and after DRL application revealed patterns different from those of temperature. Unlike PMV and Humidex, which tend to follow the rise and fall of temperature, humidity did not show a consistent trend. In C, D1, and D2, humidity remained within the comfortable range. However, when Humidex was applied, D1 maintained distinctly lower humidity levels compared with C and D2. Without considering Humidex, as shown in Table 7, C and D2 exhibited similar values in most metrics. Unlike the other indices, humidity-related results in D1 and D2 tended to remain close to C. Although large statistical differences were not observed, D1 cases maintained humidity lower than C. These results indicate that while humidity itself did not undergo a major dynamic change before and after DRL application, Humidex contributed to improved comfort management, with an average humidity reduction of 2.5%. This suggests that predicting and managing indoor comfort may be more effective when multiple indices are combined, rather than relying on a single index, while also accounting for climate and building characteristics. In baseline operation, the two first-floor spaces, the left side of the second floor, and both 28th floor spaces tended to remain below the comfort range, whereas the right side of the second floor and both 13th floor spaces tended to remain above it. In

particular, on the 28th floor, the left space exhibited the highest proportion of deviations in both directions, while the right space maintained the lowest levels across all eight spaces.

The proposed DRL model, incorporating PMV and Humidex, reduced variability in indoor conditions that previously differed by floor and zone. Unlike D2, which led to some floors being perceived as too cold and others as too hot, the DRL model managed all eight spaces within the comfort range. This section has focused on a comparative analysis of indoor comfort before and after model application, which demonstrates that the DRL-based approach provides more effective comfort management than conventional operator-based building control. The following section addresses the second objective of this study by evaluating the effect of the proposed DRL model on reducing total building operating costs while maintaining comfort.

5.5.2. Analysis of Building Operating Costs

The final operating costs before and after applying the DRL models are presented in Table 10. Weekends and holidays were excluded, and only weekdays within the controllable operating hours of the AHUs were considered. The actual energy consumption represents the estimated use of the ten AHUs, calculated using Equation (1). The calculation-based AHU consumption reflects estimated use derived from the AHU control history during operable hours (07:30 to 18:30). SAC-DRL consumption refers to simulated energy use generated by the model during training and evaluation. The discrepancy between actual consumption and calculation-based AHU consumption arises because variations in cooling and dehumidification were not captured in the per-unit power values. During the training period, an exchange rate of USD 1 = MYR 4.4703 was applied.

Table 10. Comparison of operating expenses for a schedule-based building and a DRL-based intelligent building.

	Total	A Schedule-Based Building (a)	A DRL Model-Based Intelligent Building		Savings (a, b)	Saving Rate (%), (a, b)
			PMV + Humidex (b)	PMV		
Energy	Energy consumption (kWh)	488,235	453,859	450,889	34,376	7
	Energy cost (USD), (c)	39,865	37,060	36,818	2805	7
Carbon	Carbon emission (kgCO ₂ eq)	377,894	351,286	348,988	26,607	7
	Carbon permits generation	-	26	28	-	-
	Carbon permits sales	-	26	27	-	-
	Carbon revenue (USD), (d)	-	176	183	-	-
Building operation cost (USD), (c–d)		39,865	36,884	36,185	2981	7.5

Regardless of type, the DRL approach yielded more efficient results in both energy use and cost compared with the baseline. Differences among the DRL models were small, although the PMV-only model showed slightly greater savings. As indicated in Figure 17, this occurred because the PMV-only model violated comfort conditions more frequently than the combined model, while emphasizing energy savings. The proposed DRL model combining PMV and Humidex achieved approximately 7% reductions in energy consumption compared with the schedule-based building. Energy costs, calculated from electricity charges, were also reduced by 7%, and carbon emissions decreased by the same proportion. The resulting carbon permits from these savings amounted to 26, all sold, generating revenue of about USD 176. This revenue equals the cost of operating ten AHUs for about 8 h and 30 min. By simple projection, extending the same operation over nine months with 60 AHUs would yield about USD 1056, which is sufficient to operate all AHUs in a 28-floor high-rise for about 45 min. With greater emission reductions, even larger revenues can be expected.

Carbon permit revenues cannot be obtained under conventional building operation and thus represent a newly created income source. However, although it may be highly profitable, depending on the emissions trading market, it can be considered an additional revenue source due to energy savings. Consequently, the final operating cost was calculated by subtracting the carbon permit revenue from the electricity charges. As a result, operating cost was reduced by 7.5%, or USD 2981, over the nine months. This calculation was based on 10 AHUs, which represent about 17% of the total 60 units installed. Although carbon permit revenue accounted for only 6% of the total cost savings of USD 2981, its contribution could increase significantly by expanding AHU control and enhancing the overall reduction in operating costs.

The reported revenue of USD 176 assumes the K-ETS price band (USD 6–10 per tCO₂). Under the higher EU price band (USD 60–80 per tCO₂), the implied revenue would be substantially larger; this should be interpreted as a hypothetical sensitivity rather than an observed outcome. This discrepancy in profitability between markets highlights the need to distinguish the policy mechanisms and economic valuation embedded in each carbon trading system when designing building operational strategies.

The DRL model proposed in this study, which combines PMV and Humidex, and the model using PMV alone showed different outcomes in terms of comfort and energy savings. As shown in Section 5.5.1 and Table 10, there is a contrasting relationship between energy savings and indoor comfort. The extent and frequency of deviations from the comfort boundary determine the level of energy savings. Conversely, when a target for energy savings is fixed, the resulting comfort conditions are also determined. From an environmental perspective, energy-saving targets can be set based on the climate of the building's location as well as operational purposes and policies. While such targets directly contribute to economic benefits, they cannot entirely exclude the provision of comfort for tenants or occupants. Therefore, coordination between environmental considerations and economic objectives is required to achieve a balanced outcome.

As environmental conditions and human factors change, optimization should be understood in relative rather than absolute terms. It involves developing and maintaining building operation policies that balance the differing positions of stakeholders. Such an approach enables both the advancement and sustainability of buildings.

6. Discussion

This study investigated the building sector as a key field in the transition toward carbon neutrality. The objectives of carbon neutrality include not only reducing or eliminating emission sources but also ensuring sustainability. In buildings, energy consumption is

the primary source of carbon emissions, and reductions can be achieved through energy savings. Building sustainability, however, requires balanced relationships among stakeholders, and reducing building energy consumption faces economic, environmental, and social limitations, because buildings are inherently complex spaces involving diverse and interrelated interests.

First, economic limitations arise from conflicting interests between stakeholders such as owners, managers, and tenants. Second, environmental limitations stem from physical constraints of time, space, and infrastructure. Third, social limitations originate from institutional systems such as carbon markets and from policies shaped by governments, corporations, and citizens. Within the economic dimension, owners seek to minimize financial losses by reducing operating costs, tenants demand thermal comfort, and managers must reconcile both requirements. These human relationships are linked to physical constraints, since reducing costs requires energy savings. In the target office building, AHU, which ensures tenant comfort, accounted for 90% of energy use. Because office buildings operate on fixed schedules of days and hours, opportunities for energy savings are temporally and spatially limited. While reducing purchased electricity and increasing renewable generation could theoretically enhance cost reduction, retrofitting or expanding infrastructure is restricted in existing buildings. Moreover, expenditures on new equipment impose an additional burden on costs.

To address these limitations, we proposed a new operation and management approach designed to satisfy all stakeholders. The method consists of three strategies focusing on cost-effectiveness, comfort, and environmental/social dimensions. First, for the economic aspect, a novel revenue model was introduced by reducing energy use and converting the savings into tradable carbon permits. Second, for comfort assurance, an intelligent operation model based on DRL was proposed to achieve energy savings and indoor comfort simultaneously. In addition, to more precisely reflect the climatic characteristics of the building, PMV and Humidex indices were combined. Third, to strengthen social diffusion of carbon neutrality and to enable more environmentally friendly operation, a scenario was developed in which individuals could also participate in carbon trading. The corresponding revenue and contributions to emission reductions were evaluated under this scenario.

The application of a SAC-based DRL to the Malaysian office building demonstrated that energy-efficient control can be achieved without sacrificing indoor thermal comfort. The model succeeded in reducing operational costs by 7.5%, primarily through reduced electricity consumption. Through the combined application of the PMV and Humidex indices, the model maintained indoor temperature effectively within the conventional comfort range of 23–26 °C, while controlling humidity at a more comfortable level than the baseline operation. This demonstrates that integrating both indices enables more precise and climate-sensitive comfort management compared to traditional temperature-based control strategies. Although only USD 176 of carbon revenue was generated—equivalent to approximately 6% of the total cost savings—this confirms that carbon trading serves as a supplementary economic incentive rather than a primary driver in office buildings.

The energy savings achieved through the proposed method can contribute to grid stability, while reduced carbon emissions support national carbon neutrality. By maintaining comfort, this approach also improves tenant satisfaction. These contributions align stakeholder interests: owners avoid passing additional energy costs to tenants, tenants are less likely to leave due to higher satisfaction, and managers can operate buildings more intelligently through the DRL model. Together, these effects enhance building sustainability.

This study had a limitation in linking Algorithm 2: Temperature and Humidity Change Decision with actual energy consumption. The collected data represented an integrated energy profile at the distribution board level, which makes it difficult to extract AHU-

specific consumption and apply control-based values. Therefore, instead of accounting for increased consumption during intensified cooling or dehumidification, the analysis relied on simplified energy consumption values based on AHU ON/OFF status per unit. Future research should integrate energy consumption data with AHU control and indoor environmental data to derive more refined energy profiles. Based on such profiles, algorithms will be applied to capture actual consumption changes according to control levels.

In this study, we proposed a building operation and management approach tailored specifically to hot and humid climates. Malaysia, located near the equator, is characterized by the absence of distinct seasonal variations throughout the year; therefore, only summer-like conditions were considered. However, to generalize and broaden the applicability of the proposed method, it must be adaptable to various climate conditions. Future research will extend this approach to reflect the characteristics of the climate in spring, summer, autumn, and winter.

The emissions trading market has been gradually expanding, with an increasing number of sectors being incorporated into its scope [113]. Market prices fluctuate significantly depending on economic, environmental, and social factors, and such volatility directly affects the profitability and decision-making of the proposed building operation model. Therefore, the DRL agent must be capable of adjusting its energy-saving and comfort-control strategies in response to changing carbon market prices [132]. Since carbon prices vary widely across national ETSs, the balance between energy savings, occupant comfort, and economic benefits must be dynamically adapted according to the scale and characteristics of each market [133]. Furthermore, linking the proposed model with renewable energy trading schemes—such as the export of surplus energy beyond mandatory supply obligations in the UK or Renewable Portfolio Standards (RPS) in the United States—highlights its potential scalability and applicability in future building energy management frameworks [108,134,135].

In this study, a single-agent SAC framework was employed to control all AHUs simultaneously, which enhances global coordination and simplifies implementation. However, this approach increases the dimensionality of the action space and may limit adaptability for zone-specific comfort preferences [50]. Future research will extend this framework to a multi-agent structure, where each zone or piece of equipment is assigned an individual agent. This allows localized learning, cooperative or competitive behaviors among zones, and better scalability. Multi-Agent SAC will be considered as a potential algorithm for this extension [47].

Because the target building is a high-rise office building actively used by many people, the trained DRL model could not be applied directly on-site. Operational decisions in real buildings are made by the owner and are typically based on traditional and hierarchical governance models. To assess the reliability and safety of the developed DRL model, a governance framework that enables effective communication among owners, facility managers, and occupants is essential. Additionally, simulations and validations are required for all periods of the year, including the rainy season. In the next stage, the governance model is deployed, expanding the data verification period, and the DRL model is expected to be applied in practice to evaluate its performance in terms of energy savings, carbon emission reductions, comfort, and tenant satisfaction. Evaluation adopts a safer and more reliable model through comparative benchmarking between various existing controllers and reinforcement learning models. In this study, we did not perform extensive hyperparameter optimization. Due to limited computational resources and the absence of parallel computing infrastructure, it was not feasible to explore a wide range of hyperparameter combinations. As a result, only a manually selected set of configurations were tested.

We acknowledge that applying systematic hyperparameter tuning methods—such as Bayesian optimization, grid/random search, population-based training, or using the web service module Optuna—could potentially yield better-performing configurations than those adopted in this work. Future research will incorporate automated hyperparameter optimization to improve model performance and ensure reproducibility.

While the proposed SAC-based DRL model demonstrates clear potential for improving building energy efficiency and supporting carbon neutrality, its real-world applicability within current policy frameworks remains subject to critical constraints. Under the Korean Emissions Trading Scheme (K-ETS), emission reductions must undergo strict Measurement, Reporting, and Verification (MRV) procedures, and the transaction costs and price volatility inherent in carbon markets could limit the direct profitability of building-level trading. Similarly, although Malaysia's Nationally Determined Contributions (NDCs) emphasize carbon-neutral development, the country has not yet implemented a formal emissions trading scheme (ETS). However, it is currently in the planning stage of introducing carbon pricing mechanisms, including the carbon tax proposal outlined in the 2025 national budget. Therefore, further research and validation are required to examine the practical effectiveness and integration of the proposed model within Malaysia's future ETS framework. Furthermore, integrating artificial intelligence into national energy or carbon markets requires not only technological capability but also regulatory transparency, accountability, and governance frameworks to ensure responsible use. Thus, while the proposed model supports carbon-neutral operations, its integration with existing policy mechanisms demands institutional adaptation, regulatory innovation, and ethical safeguards to achieve long-term sustainability.

7. Conclusions

This study proposes a DRL-based operational management model using the Soft Actor–Critic algorithm, which simultaneously considers energy savings, indoor thermal comfort, and carbon permit revenue. After nine months of training and simulation, the proposed model reduced total building operating costs by 7.5% compared to conventional operation. Energy consumption and carbon emissions were each reduced by 7%. In addition, USD 176 of carbon revenue was generated through carbon permit trading, corresponding to approximately 6% of total cost savings. Therefore, carbon revenue should be interpreted as a supplementary benefit rather than a primary economic driver. In terms of indoor thermal environment, the model maintained a similar temperature level to the baseline while reducing humidity, thereby improving humidity-related thermal comfort.

Although the absolute amount of carbon revenue is relatively small, the significance of this study lies in its empirical demonstration of an integrated operational framework that combines energy cost reduction, occupant comfort preservation, and carbon market-based economic value. Furthermore, the results indicate the potential to extend this framework toward the following:

- Multi-agent DRL architectures;
- Integration with renewable energy and energy trading markets (e.g., FIT, RPS, VPP);
- Real-world deployment supported by governance-based operational strategies.

Author Contributions: Conceptualization, K.C., M.-i.C., and S.P.; methodology, K.C., H.J., and G.Y.; software, K.C., Y.B., and G.Y.; validation, K.C., Y.B.; formal analysis, K.C.; investigation, K.C., H.J., and G.Y.; resources, K.C. and Y.B.; data curation, K.C.; writing—original draft preparation, K.C.; writing—review and editing, K.C. and M.-i.C.; visualization, K.C.; supervision, S.P.; project administration, S.P.; funding acquisition, S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-RS-2024-00436248, 50) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant, funded by the Korean government (MOTIE) (RS-2024-00398346, ESS Big Data-Based O&M and Asset Management Technical Manpower Training).

Data Availability Statement: Data available on request due to restrictions. Since the training data for deep reinforcement learning were collected from a real-world building, their use is partially restricted due to privacy and legal considerations.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1 presents the simulation environment conditions used in the deep reinforcement learning model of this study. Python and the Anaconda virtual environment were employed for the development of both the simulation environment and the DRL model.

Table A1. Main libraries used in environment setup and development environment.

Library	Description	Version
python	-	3.9.18
pytorch	Deep Learning Framework	2.5.1
pytorch-cuda	Metapackage for using CUDA (GPU) in PyTorch	12.1
pythermalcomfort	Thermal comfort calculations such as PMV, PPD, and Humidex	3.3.0
forex_python.converter	Real-time exchange rate inquiry and currency conversion	1.9.2
gym	Library providing reinforcement learning simulation environment	0.26.2
GPU	NVIDIA GeForce RTX 4060 Ti	-
CPU	13th Gen Intel(R) Core(TM) i5-13600KF(3.50 GHz)	-
RAM	64.0 GB	-
OS	Windows 11, x64	-

Table A2 presents the hyperparameters and the reward weight ratios employed in the DRL model of this study. In addition, it includes the parameter settings required to predict the PMV index as part of the indoor comfort evaluation.

Table A2. Parameters used for DRL training and reward.

Symbol	Description	Values
γ	Discount factor	0.98
α	Reward temperature	Auto
τ	Target network update	0.01
-	Replay buffer size	1×10^6
-	Batch size	256
-	Learning rate: Actor	3×10^{-4}
-	Learning rate: Critic	3×10^{-4}
-	Learning rate: Alpha	1×10^{-4}
-	Policy	Gaussian
-	Policy update	4 (one step, 15 min)
-	Agent action choice term	15 min (one step)
ω_{p-low_z}	PMV comfort-range out penalty	80
ω_{p-high_z}		
ω_{p-low_m}	PMV min-max out penalty	160
ω_{p-high_m}		

Table A2. Cont.

Symbol	Description	Values
ω_{h-low_z}	Humidex comfort-range out penalty	80
ω_{h-high_z}		
ω_{h-low_m}	Humidex min–max-range out penalty	160
ω_{h-high_m}		
λ_3	Discomfort PMV weight	0.3
λ_4	Discomfort Humidex weight	0.3
λ_1	Energy weight	0.3
λ_2	Carbon permit revenue weight	0.1
-	Training data	1 July 2024–29 February 2025
-	Test data	1 March 2025–31 March 2025
-	1 episode	One day
-	Air velocity	0.1 m/s
-	Metabolic rate	1.0 met
-	Clothing insulation	0.5 clo
-	tr_offset	0
-	(Mean radiant temperature—dry-bulb temperature)	

Table A3 presents the detailed results corresponding to Figure 12 in the main text.

Table A3. Detailed data on indoor comfort and energy savings from the DRL model training results.

Out-of-Range	PMV					Humidex				
	0–5%	5–10%	10–20%	20–30%	30%	0–5%	5–10%	10–20%	20–30%	30%
Daily count	154	11	2	0	0	135	26	6	0	0
Daily count rate (%)	92.2	6.6	1.2	0	0	80.8	15.6	3.6	0	0
Strict	152	9	0	0	0	135	26	5	0	0
Mild	2	2	1	0	0	0	0	1	0	0
Moderate	0	0	1	0	0	0	0	0	0	0
Severe	0	0	0	0	0	0	0	0	0	0
Step count	6006	429	78	0	0	5265	1014	234	0	0
Out-of-range step	55	29	13	0	0	69	76	36	0	0
Out-of-range step rate (%)	0.92	6.76	16.67	0	0	1.31	7.50	15.38	0	0
Energy saving 0–5%	14	0	0	0	0	12	1	1	0	0
Energy saving 5–10%	42	0	0	0	0	36	6	0	0	0
Energy saving 10–15%	69	6	1	0	0	65	9	2	0	0
Energy saving 15–20%	18	4	0	0	0	17	3	2	0	0
Energy saving 20–25%	1	0	0	0	0	0	1	0	0	0
Total daily	167					167				
Total step	6513					6513				
Total out-of-range rate (%)	1.49					2.78				

Appendix B

Figure A1 presents a statistical graph of ambience based on indoor temperature and humidity data collected over nine months from 1 July 2024 to 31 March 2025. The blue shading indicates the commonly used comfort range. In the pilot building, PMV is managed somewhat below the typical comfort range, while Humidex is managed above the typical comfort range. This management policy reflects the local office culture, so rather than forcing it into a standard category, it should be reflected in existing management policies.

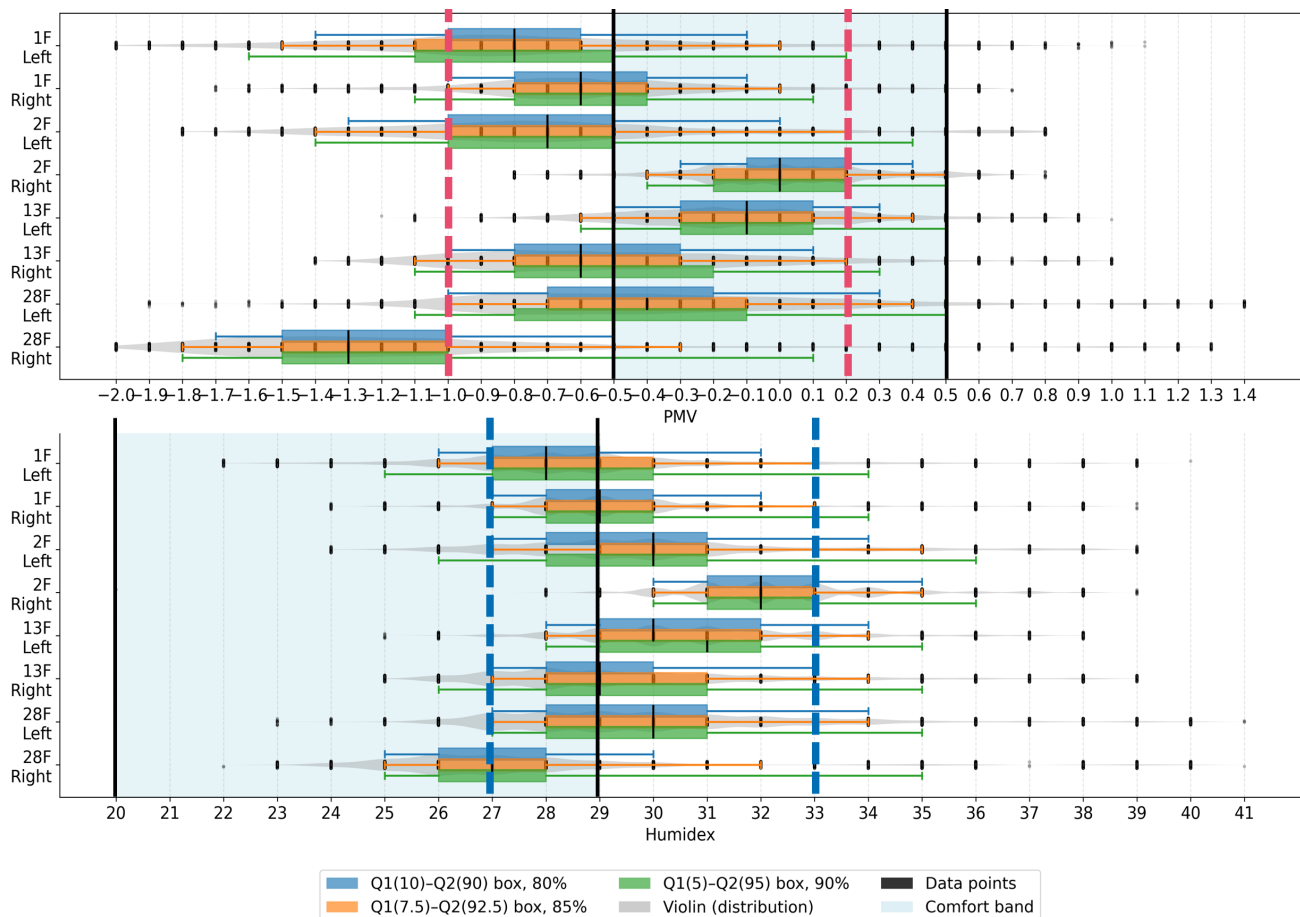


Figure A1. Setting a comfortable range based on indoor temperature and humidity statistics.

Figure A2 illustrates the zone-level trends of PMV and temperature from 1 July 2024 to 31 March 2025. The training period ran from 1 July 2024 to 28 February 2025 with 100 training iterations performed. The evaluation period spanned 1–31 March 2025, with a single evaluation run. The data shown in the figure correspond to the final training cycle of the complete set of iterations.

Figure A3 illustrates the zone-level trends of Humidex and humidity from 1 July 2024 to 31 March 2025. The training period ran from 1 July 2024 to 28 February 2025, with 100 training iterations performed. The evaluation period spanned 1–31 March 2025, with a single evaluation run. The data shown in the figure correspond to the final training cycle of the complete set of iterations.

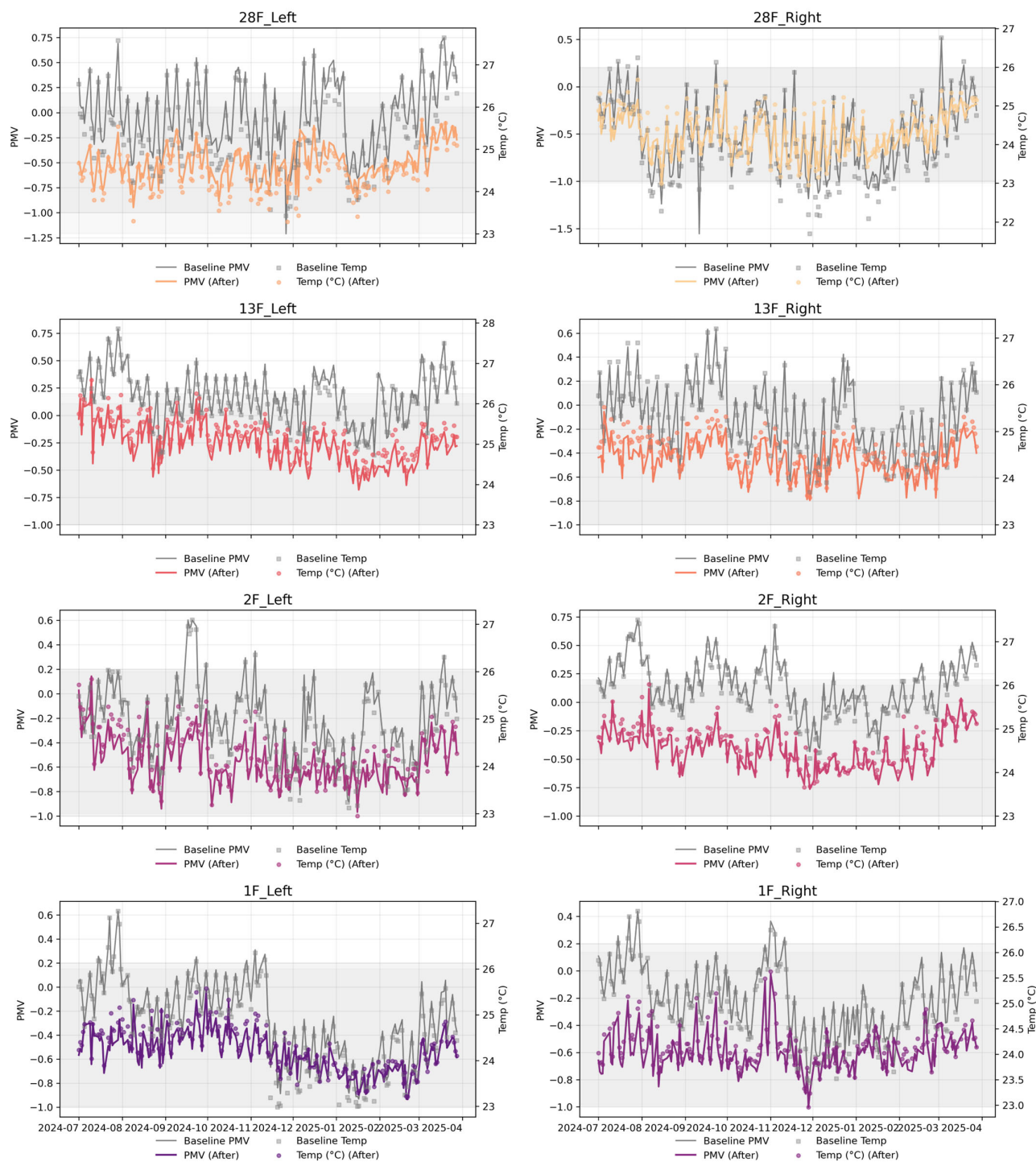


Figure A2. Zone-wise daily means (09:00–18:30), PMV, and temperature comparison with baseline.

Figure A4 illustrates the thermal comfort between PMV and PDD based on the learned results. For each space, we can predict how indoor comfort changed during the morning, lunch, and afternoon. Overall, the temporal correlation analysis revealed that comfort-related factors were more dominant in the morning hours, whereas the relationship between energy use and time became stronger in the afternoon. In particular, we can infer that indoor comfort was sacrificed more for energy savings on the 2nd and 13th floors, which correspond to mid-rise buildings, compared to lower and upper floors.

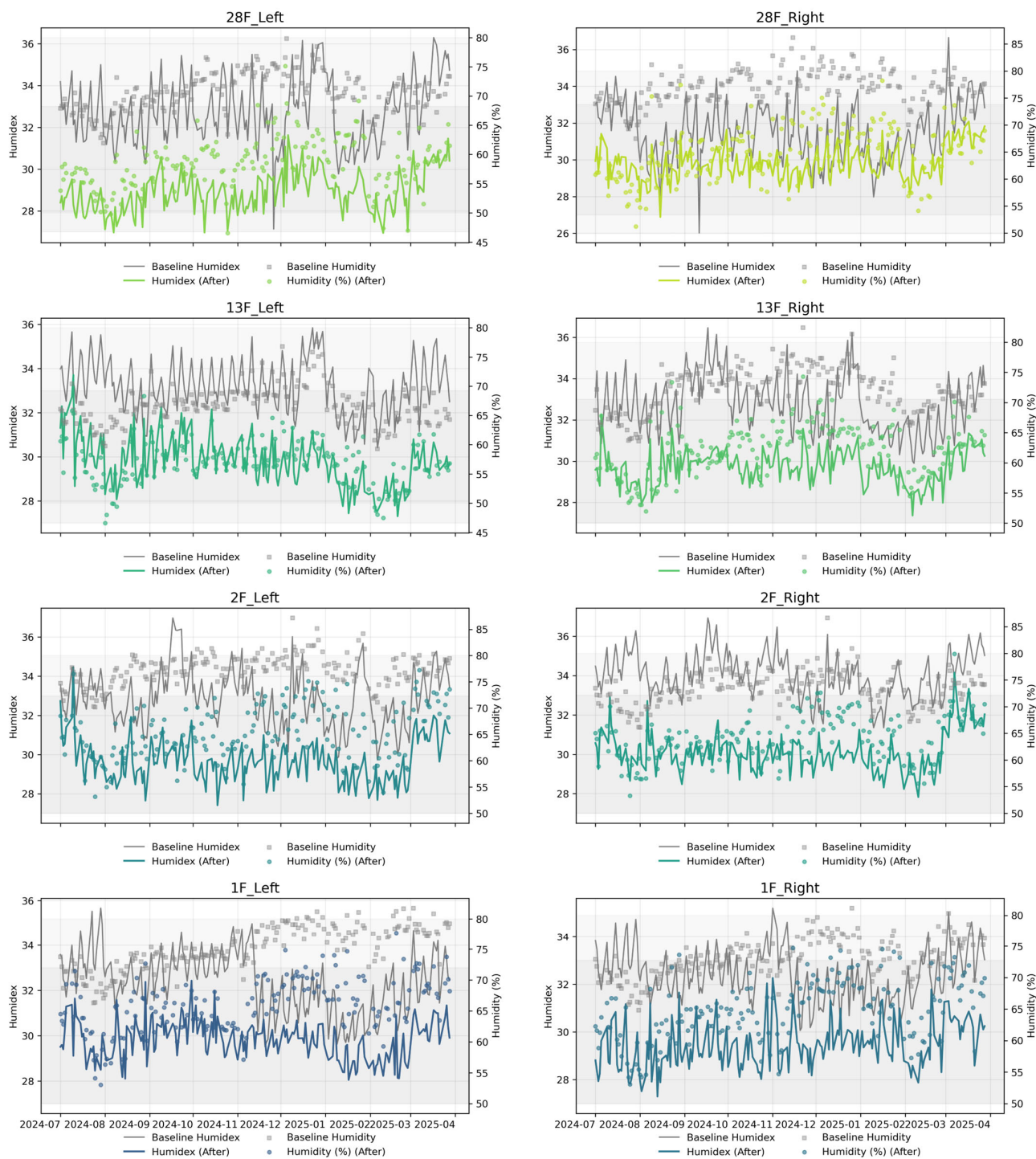


Figure A3. Zone-wise daily means (09:00–18:30), Humidex, and humidity comparison with baseline.

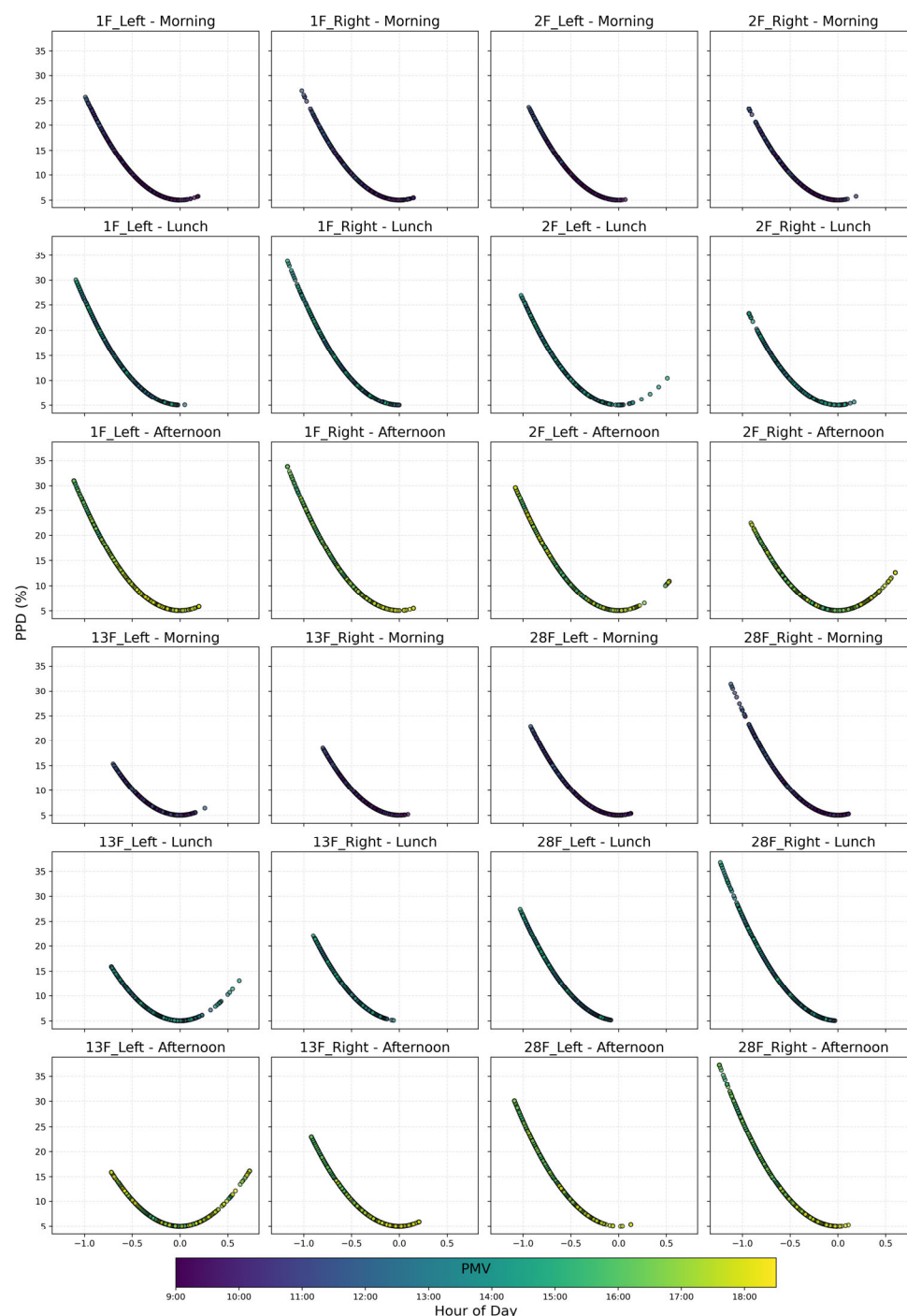


Figure A4. Thermal comfort based on predicted mean vote (PMV) and predicted dissatisfaction (PPD) index based on learning results.

Appendix C

The thermal comfort of building occupants is calculated using two types of variables: the parameters of indoor occupants, and the indoor environmental variables of the building [136]. Indoor occupant parameters are categorized into clothing insulation (amount of clothing) and metabolic rate (activity level). Building indoor environmental variables are classified into air temperature, relative humidity, air velocity, and average radiant temperature. Because PMV takes into account a variety of individual circumstances and building environmental conditions, there are many ways to optimize PMV. These include methods for creating a thermal environment by controlling air temperature and humid-

ity, methods for optimizing air flow to control perceived temperature, and methods for adjusting clothing volume and heating and cooling conditions according to season and activity. In this paper, we used the Python pythermalcomfort.models library instead of directly coding the PMV calculation formula. The PMV calculation method used in this library is based on Fanger's thermal comfort model. Fanger's thermal comfort model was proposed by P. Ole. Fanger of Denmark and is currently used in the ISO7730 and ASHRAE 55 standards. The following equations, Appendix C Equations (A1)–(A12), are the basic mechanism and formula for calculating PMV [137].

$$\text{Thermal balance} = \text{Internal heat production} - \text{Loss of heat to the environment}, \quad (\text{A1})$$

$$\begin{aligned} \text{PMV} = & (0.303e^{-0.036M} + 0.028) * \{ \text{Metabolic heat production} - \text{Dry heat loss by skin diffusion} - \\ & \text{Sweating heat loss} - \text{Latent heat loss by respiration} - \text{Sensible heat loss by respiration} - \\ & \text{Radiative heat loss} - \text{Convective heat loss} \}, \end{aligned} \quad (\text{A2})$$

where

$$\text{Metabolic heat production} = M - W, \quad (\text{A3})$$

$$\text{Dry heat loss by skin diffusion} = 3.05 * 10^{-3} [5733 - 6.99(M - W) - p_a], \quad (\text{A4})$$

$$\text{Sweating heat loss} = 0.42[(M - W) - 58.15], \quad (\text{A5})$$

$$\text{Latent heat loss by respiration} = 1.7 * 10^{-5} M(5867 - p_a), \quad (\text{A6})$$

$$\text{Sensible heat loss by respiration} = 0.0014M(34 - t_a), \quad (\text{A7})$$

$$\text{Radiative heat loss} = 3.96 * 10^{-8} f_{cl} * [(t_{cl} + 273)^4 - (\bar{t}_r + 273)^4], \quad (\text{A8})$$

$$\text{Convective heat loss} = f_{cl} h_c (t_{cl} - t_a), \quad (\text{A9})$$

where M is the metabolic rate (W/m^2), W is the effective mechanical power (external work), equal to zero for most activities (W/m^2). p_a is the water vapor partial pressure (Pa). t_a is the air temperature ($^{\circ}\text{C}$), and $\bar{t}_r$ is the mean radiant temperature ($^{\circ}\text{C}$). f_{cl} is the clothing surface area index. t_{cl} is the surface temperature of clothing ($^{\circ}\text{C}$). h_c is the convective heat transfer coefficient ($\text{W}/\text{m}^2\text{K}$).

$$f_{cl} = \begin{cases} 1.00 + 1290I_{cl} & \text{for } I_{cl} \leq 0.078 \text{ m}^2\text{C}/\text{W} \\ 1.05 + 0.645I_{cl} & \text{for } I_{cl} > 0.078 \text{ m}^2\text{C}/\text{W} \end{cases} \quad (\text{A10})$$

$$t_{cl} = 35.7 - 0.028(M - W) - I_{cl}(\text{Radiative heat loss}) + \text{Convective heat loss}, \quad (\text{A11})$$

$$h_c = \begin{cases} 2.38(t_{cl} - t_a)^{0.25} & \text{for } 2.38(t_{cl} - t_a)^{0.25} > 12.1 * \text{root}(v_{ar}) \\ 12.1 * \text{root}(v_{ar}) & \text{for } 2.38(t_{cl} - t_a)^{0.25} < 12.1 * \text{root}(v_{ar}) \end{cases}, \quad (\text{A12})$$

where I_{cl} is the thermal insulation value of clothing, and v_{ar} is the relative airflow velocity (relative to the human body). The general sensory scale according to the PMV index was derived from a study of over 1300 subjects, and its criteria are shown in Table A4 below.

Table A4. PMV sensation scale.

Cold	Cool	Slightly Cool	Neutral	Slightly Warm	Warm	Hot
−3	−2	−1	0	+1	+2	+3

The metabolic rate M and clothing insulation value I_{cl} , which are essential for calculating PMV, are described in Tables A5 and A6 below. For typical office environments

using computers, a metabolic rate of 1.0 is used. More detailed information on the various metabolic rates and clothing insulation can be found in [138].

Table A5. Human metabolic rate according to type of activity.

Activity	Metabolic Rate	
	W/m ² °C	Met
Reclining	46	0.8
Seated, relaxed	58	1.0
Sedentary activity (Office, dwelling, school, laboratory)	70	1.2
Standing, light activity (Shopping, laboratory, light industry)	93	1.6
Medium activity (Shop assistant, domestic work, machine work)	116	2.0
Walking on the level:		
2 km/h	110	1.9
3 km/h	140	2.4
4 km/h	165	2.8
5 km/h	200	3.4

Table A6. Clothing insulation values.

Activity	Metabolic Rate	
	W/m ² °C	Met
Nude	0.000	0.0
Shorts	0.015	0.1
Typical tropical clothing ensemble: briefs, shorts, open-neck shirt with short sleeves, light socks, and sandals	0.045	0.3
Light summer clothing: briefs, long light-weight-trousers, open-neck shirt with short sleeves, light socks, and shoes	0.080	0.5
Light working ensemble: light underwear, cotton work shirt with long sleeves, work trousers, woolen socks, and shoes	0.110	0.7
Typical indoor winter clothing ensemble: underwear, shirt with long sleeves, trousers, jacket or sweater with long sleeves, heavy socks, and shoes	0.160	1.0
Heavy traditional European business suit: cotton underwear with long legs and sleeves, shirt, suit including trousers, cotton underwear with long legs and sleeves, shirt, suit including trousers, jacket and waistcoat, woolen socks, and heavy shoes	0.230	1.5

Humidex is an index used to measure how hot a person feels in hot and humid weather. It is intended to express the combined effects of warm air and humidity. Humidex is an index that combines air temperature and relative humidity and indicates how hot it feels to a person. Humidex was developed to take into account the climate characteristics of Canada. As the complex physical impacts of high temperature and humidity become more severe, the need to understand and quantify them emerges. To this end, Canada's Atmospheric Environment Service developed Humidex for hot and humid environments. Malaysia, where the target building in this study is located, has climate characteristics similar to those that led to the development of Humidex in Canada. The basic equation for Humidex uses temperature and dew point temperature, as shown in Equation (A13) [110].

$$\text{Humidex} = T - h, \quad (\text{A13})$$

$$h = \frac{5}{9}(e - 10), \quad (\text{A14})$$

$$e = 6.11[\exp\{(\frac{M_W L}{R^*}) * (\frac{1}{273.16} - \frac{1}{T_d})\}], \quad (\text{A15})$$

where e is the vapor pressure (mb); M_W is the molecular weight of water, 18.016 g/mol; L is the latent heat of vaporization, 2500 J/g; R^* is the universal gas constant, 8.314 J/molK; 6.11 refers to the saturation vapor pressure at a standard temperature of 273.16 °K; 273.16 refers to the melting point of ice; and T_d is the dew point temperature. Therefore, h represents the temperature in Celsius that must be added to the dry bulb temperature to account for the additional discomfort caused by excessive humidity. Humidex should not be given in units, but it is almost always presented with °C to help the general public understand [110]. Humidex has three methods based on Equation (A13): dew point temperature, wet bulb temperature, and relative humidity. Table A7 describes the comfort level of the Humidex index using relative humidity. Figure A2 is a Humidex heatmap plotted against temperature and relative humidity. The Humidex range set here is the same as the PMV range.

Table A7. Degree of comfort at different Humidex values.

Humidex Index (°C)	Degree of Comfort
20–29	Comfortable
30–39	Varying degrees of discomfort
40–45	Almost everyone uncomfortable
46 and over	Many types of labor must be restricted

References

1. Hamilton, I.; Kennard, H.; Rapf, O.; Amoroch, J.; Steuwer, S.; Kockat, J.; Toth, Z. *Global Status Report for Buildings and Construction—Beyond Foundations: Mainstreaming Sustainable Solutions to Cut Emissions From the Buildings Sector*; UNEP: Nairobi, Kenya, 2024.
2. Liu, Y.; Lei, P.; He, D. Endogenous Green Technology Progress, Green Transition and Carbon Emissions. *Int. Rev. Econ. Financ.* **2024**, *91*, 69–82. [\[CrossRef\]](#)
3. Park, S.; Lee, S.; Park, S.; Park, S. AI-Based Physical and Virtual Platform with 5-Layered Architecture for Sustainable Smart Energy City Development. *Sustainability* **2019**, *11*, 4479. [\[CrossRef\]](#)
4. Seto, K.C.; Churkina, G.; Hsu, A.; Keller, M.; Newman, P.W.; Qin, B.; Ramaswami, A. From Low-to Net-Zero Carbon Cities: The next Global Agenda. *Annu. Rev. Environ. Resour.* **2021**, *46*, 377–415. [\[CrossRef\]](#)
5. Chen, L.; Msigwa, G.; Yang, M.; Osman, A.I.; Fawzy, S.; Rooney, D.W.; Yap, P.-S. Strategies to Achieve a Carbon Neutral Society: A Review. *Environ. Chem. Lett.* **2022**, *20*, 2277–2310. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Braulio-Gonzalo, M.; Bovea, M.D.; Jorge-Ortiz, A.; Juan, P. Contribution of Households' Occupant Profile in Predictions of Energy Consumption in Residential Buildings: A Statistical Approach from Mediterranean Survey Data. *Energy Build.* **2021**, *241*, 110939. [\[CrossRef\]](#)
7. Chen, S.; Wu, J.; Pan, Y.; Ge, J.; Huang, Z. Simulation and Case Study on Residential Stochastic Energy Use Behaviors Based on Human Dynamics. *Energy Build.* **2020**, *223*, 110182. [\[CrossRef\]](#)
8. Gulbinas, R.; Taylor, J.E. Effects of Real-Time Eco-Feedback and Organizational Network Dynamics on Energy Efficient Behavior in Commercial Buildings. *Energy Build.* **2014**, *84*, 493–500. [\[CrossRef\]](#)
9. Du, J.; Pan, W.; Yu, C. In-Situ Monitoring of Occupant Behavior in Residential Buildings—A Timely Review. *Energy Build.* **2020**, *212*, 109811. [\[CrossRef\]](#)
10. Ruparathna, R.; Hewage, K.; Sadiq, R. Improving the Energy Efficiency of the Existing Building Stock: A Critical Review of Commercial and Institutional Buildings. *Renew. Sustain. Energy Rev.* **2016**, *53*, 1032–1045. [\[CrossRef\]](#)
11. Favero, M.; Møller, J.K.; Cali, D.; Carlucci, S. Human-in-the-Loop Methods for Occupant-Centric Building Design and Operation. *Appl. Energy* **2022**, *325*, 119803. [\[CrossRef\]](#)
12. Li, H.; Wang, Z.; Hong, T. Occupant-Centric Key Performance Indicators to Inform Building Design and Operations. *J. Build. Perform. Simul.* **2021**, *14*, 814–842. [\[CrossRef\]](#)

13. Tao, Y.X.; Zhu, Y.; Passe, U. Modeling and Data Infrastructure for Human-Centric Design and Operation of Sustainable, Healthy Buildings through a Case Study. *Build. Environ.* **2020**, *170*, 106518. [\[CrossRef\]](#)
14. Kojima, T.; Sakuma, T.; Nishihara, N.; Hayashi, T.; Munakata, J. Causal Modeling between Workplace Productivity and Workers' Satisfaction with Various Spaces in Office Buildings. *J. Asian Archit. Build. Eng.* **2017**, *16*, 409–415. [\[CrossRef\]](#)
15. Bluysen, P.M.; Roda, C.; Mandin, C.; Fossati, S.; Carrer, P.; De Kluizenaar, Y.; Mihucz, V.G.; de Oliveira Fernandes, E.; Bartzis, J. Self-reported Health and Comfort in 'Modern' Office Buildings: First Results from the European OFFICAIR Study. *Indoor Air* **2016**, *26*, 298–317. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Lala, B.; Hagishima, A. Impact of Escalating Heat Waves on Students' Well-Being and Overall Health: A Survey of Primary School Teachers. *Climate* **2023**, *11*, 126. [\[CrossRef\]](#)
17. Kok, N.; Palacios, J. Tenant Satisfaction and Commercial Building Performance. In Proceedings of the European Real Estate Society Conference-London, London, UK, 12–15 July 2023.
18. Eichholtz, P.; Kok, N.; Quigley, J.M. Doing Well by Doing Good? Green Office Buildings. *Am. Econ. Rev.* **2010**, *100*, 2492–2509. [\[CrossRef\]](#)
19. Shaikh, P.H.; Nor, N.B.M.; Nallagownden, P.; Elamvazuthi, I.; Ibrahim, T. A Review on Optimized Control Systems for Building Energy and Comfort Management of Smart Sustainable Buildings. *Renew. Sustain. Energy Rev.* **2014**, *34*, 409–429. [\[CrossRef\]](#)
20. Chen, C.; An, J.; Wang, C.; Duan, X.; Lu, S.; Che, H.; Qi, M.; Yan, D. Deep Reinforcement Learning-Based Joint Optimization Control of Indoor Temperature and Relative Humidity in Office Buildings. *Buildings* **2023**, *13*, 438. [\[CrossRef\]](#)
21. Kang, Y.; Xu, W.; Wu, J.; Li, H.; Liu, R.; Lu, S.; Rong, X.; Xu, X.; Pang, F. Study on Comprehensive Whole Life Carbon Emission Reduction Potential and Economic Feasibility Impact Based on Progressive Energy-Saving Targets: A Typical Renovated Ultra-Low Energy Office. *J. Build. Eng.* **2022**, *58*, 105029. [\[CrossRef\]](#)
22. Escrivá-Escrivá, G. Basic Actions to Improve Energy Efficiency in Commercial Buildings in Operation. *Energy Build.* **2011**, *43*, 3106–3111. [\[CrossRef\]](#)
23. Liu, Q.; Liu, Y.; Cai, W.; Du, Y. Multi-Dimensional Building Carbon Emissions Echelon Peak Target Setting in China Based on Building Types, Sources, and Indicators. *Appl. Energy* **2025**, *386*, 125532. [\[CrossRef\]](#)
24. Satola, D.; Balouktsi, M.; Luetzkendorf, T.; Wiberg, A.H.; Gustavsen, A. How to Define (Net) Zero Greenhouse Gas Emissions Buildings: The Results of an International Survey as Part of IEA EBC Annex 72. *Build. Environ.* **2021**, *192*, 107619. [\[CrossRef\]](#)
25. Hansen, K.; Breyer, C.; Lund, H. Status and Perspectives on 100% Renewable Energy Systems. *Energy* **2019**, *175*, 471–480. [\[CrossRef\]](#)
26. Kim, S.-S.; Lee, W.; Bhang, B.G.; Choi, J.H.; Lee, S.H.; Woo, S.C.; Nam, W.J.; Ahn, H.K. Return of Interest Planning for Photovoltaics Connected with Energy Storage System by Considering Maximum Power Demand. *Appl. Sci.* **2020**, *10*, 786. [\[CrossRef\]](#)
27. Kryk, B.; Guzowska, M.K. Assessing the Level of Energy Poverty Using a Synthetic Multidimensional Energy Poverty Index in EU Countries. *Energies* **2023**, *16*, 1333. [\[CrossRef\]](#)
28. Szczepanik-Scislo, N.; Scislo, L. Dynamic Real-Time Measurements and a Comparison of Gas and Wood Furnaces in a Dual-Fuel Heating System in Order to Evaluate the Occupants' Safety and Indoor Air Quality. *Buildings* **2023**, *13*, 2125. [\[CrossRef\]](#)
29. Joo, J.; Paavola, J.; Van Alstine, J. The Divergence of South Korea's Emissions Trading Scheme (ETS) from the EU ETS: An Institutional Complementarity View. *Politics Policy* **2023**, *51*, 1155–1173. [\[CrossRef\]](#)
30. Moghaddasi, H.; Culp, C.; Vanegas, J.; Ehsani, M. Net Zero Energy Buildings: Variations, Clarifications, and Requirements in Response to the Paris Agreement. *Energies* **2021**, *14*, 3760. [\[CrossRef\]](#)
31. Daggash, H.A.; Mac Dowell, N. The Implications of Delivering the UK's Paris Agreement Commitments on the Power Sector. *Int. J. Greenh. Gas Control* **2019**, *85*, 174–181. [\[CrossRef\]](#)
32. Doukas, H.; Nikas, A.; González-Eguino, M.; Arto, I.; Anger-Kraavi, A. From Integrated to Integrative: Delivering on the Paris Agreement. *Sustainability* **2018**, *10*, 2299. [\[CrossRef\]](#)
33. Zhang, S.; Wang, K.; Xu, W.; Iyer-Raniga, U.; Athienitis, A.; Ge, H.; Cho, D.W.; Feng, W.; Okumiya, M.; Yoon, G.; et al. Policy Recommendations for the Zero Energy Building Promotion towards Carbon Neutral in Asia-Pacific Region. *Energy Policy* **2021**, *159*, 112661. [\[CrossRef\]](#)
34. Scherhauer, P. Better Research through More Participation? The Future of Integrated Climate Change Assessments. *Futures* **2021**, *125*, 102661. [\[CrossRef\]](#)
35. Lu, P.; Li, Z.; Wen, Y.; Liu, J.; Yuan, Y.; Niu, R.; Wang, Y.; Han, L. Fresh Insights for Sustainable Development: Collaborative Governance of Carbon Emissions Based on Social Network Analysis. *Sustain. Dev.* **2023**, *31*, 1873–1887. [\[CrossRef\]](#)
36. Park, H.J.; Lee, Y.H.; Lee, T. Imagining Carbon-Neutral Futures: Comprehensive Narratives through Citizen Participation. *Futures* **2023**, *154*, 103274. [\[CrossRef\]](#)
37. Park, S.; Cho, K.; Choi, M. A Sustainability Evaluation of Buildings: A Review on Sustainability Factors to Move towards a Greener City Environment. *Buildings* **2024**, *14*, 446. [\[CrossRef\]](#)
38. D'Oca, S.; Hong, T.; Langevin, J. The Human Dimensions of Energy Use in Buildings: A Review. *Renew. Sustain. Energy Rev.* **2018**, *81*, 731–742. [\[CrossRef\]](#)

39. Dimoudi, A.; Tompa, C. Energy and Environmental Indicators Related to Construction of Office Buildings. *Resour. Conserv. Recycl.* **2008**, *53*, 86–95. [\[CrossRef\]](#)
40. Thormark, C. The Effect of Material Choice on the Total Energy Need and Recycling Potential of a Building. *Build. Environ.* **2006**, *41*, 1019–1026. [\[CrossRef\]](#)
41. Chau, C.K.; Xu, J.M.; Leung, T.M.; Ng, W.Y. Evaluation of the Impacts of End-of-Life Management Strategies for Deconstruction of a High-Rise Concrete Framed Office Building. *Appl. Energy* **2017**, *185*, 1595–1603. [\[CrossRef\]](#)
42. Lai, J.H.K.; Lu, M. Carbon Emission and Maintenance Cost of Commercial Buildings: Quantification, Analysis and Benchmarking. *J. Clean. Prod.* **2024**, *447*, 141459. [\[CrossRef\]](#)
43. Razzano, G.; Brandi, S.; Piscitelli, M.S.; Capozzoli, A. Rule Extraction from Deep Reinforcement Learning Controller and Comparative Analysis with ASHRAE Control Sequences for the Optimal Management of Heating, Ventilation, and Air Conditioning (HVAC) Systems in Multizone Buildings. *Appl. Energy* **2025**, *381*, 125046. [\[CrossRef\]](#)
44. American Society of Heating, Refrigerating and Air-Conditioning Engineers. *Sequences of Operation for Common HVAC Systems*; American Society of Heating, Refrigerating and Air-Conditioning Engineers: Peachtree Corners, GA, USA, 2006.
45. Pradhan, O.; Wen, J.; Chen, Y.; Lu, X.; Chu, M.; Fu, Y.; O'Neill, Z.; Wu, T.; Candan, K.S. Dynamic Bayesian Net-work-Based Fault Diagnosis for ASHRAE Guideline 36: High Performance Sequence of Operation for HVAC Systems. In Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, Coimbra, Portugal, 17–18 November 2021; pp. 365–368.
46. Sun, H.; Hu, Y.; Luo, J.; Guo, Q.; Zhao, J. Enhancing HVAC Control Systems Using a Steady Soft Actor–Critic Deep Reinforcement Learning Approach. *Buildings* **2025**, *15*, 644. [\[CrossRef\]](#)
47. Chen, H.; Sun, D.; Sun, Y.; Zhang, Y.; Yang, H. Multi-Agent Deep Reinforcement Learning-Based HVAC and Electrochromic Window Control Framework. *Buildings* **2025**, *15*, 3114. [\[CrossRef\]](#)
48. Duan, Z.; Li, B.; Zi, Y.; Yao, G. Building retrofit multiobjective optimization using neural networks and genetic algorithm three for energy carbon and comfort. *Sci. Rep.* **2025**, *15*, 38076. [\[CrossRef\]](#) [\[PubMed\]](#)
49. Lim, S.-H.; Kim, T.-G.; Yeom, D.J.; Yoon, S.-G. Robust Deep Reinforcement Learning for Personalized HVAC System. *Energy Build.* **2024**, *319*, 114551. [\[CrossRef\]](#)
50. Kodama, N.; Harada, T.; Miyazaki, K. Home Energy Management Algorithm Based on Deep Reinforcement Learning Using Multistep Prediction. *IEEE Access* **2021**, *9*, 153108–153115. [\[CrossRef\]](#)
51. Wang, H.; Chen, X.; Vital, N.; Duffy, E.; Razi, A. Energy Optimization for HVAC Systems in Multi-VAV Open Offices: A Deep Reinforcement Learning Approach. *Appl. Energy* **2024**, *356*, 122354. [\[CrossRef\]](#)
52. Burley, B.J. An Update on ANSI/ASHRAE Standard 62.1. *ASHRAE J.* **2022**, *64*, 28–30.
53. Deng, X.; Zhang, Y.; Zhang, Y.; Qi, H. Toward Smart Multizone HVAC Control by Combining Context-Aware System and Deep Reinforcement Learning. *IEEE Internet Things J.* **2022**, *9*, 21010–21024. [\[CrossRef\]](#)
54. Olesen, B.W.; Parsons, K.C. Introduction to Thermal Comfort Standards and to the Proposed New Version of EN ISO 7730. *Energy Build.* **2002**, *34*, 537–548. [\[CrossRef\]](#)
55. Deshmukh, S.; Samouhos, S.; Glicksman, L.; Norford, L. Fault Detection in Commercial Building VAV AHU: A Case Study of an Academic Building. *Energy Build.* **2019**, *201*, 163–173. [\[CrossRef\]](#)
56. Lee, J.M.; Hong, S.H.; Seo, B.M.; Lee, K.H. Application of Artificial Neural Networks for Optimized AHU Discharge Air Temperature Set-Point and Minimized Cooling Energy in VAV System. *Appl. Therm. Eng.* **2019**, *153*, 726–738. [\[CrossRef\]](#)
57. Yu, Y.; Liu, M.; Li, H.; Yu, D.; Loftness, V. Synergization of Air Handling Units for High Energy Efficiency in Office Buildings: Implementation Methodology and Performance Evaluation. *Energy Build.* **2012**, *54*, 426–435. [\[CrossRef\]](#)
58. Wang, Z.; Hong, T. Reinforcement Learning for Building Controls: The Opportunities and Challenges. *Appl. Energy* **2020**, *269*, 115036. [\[CrossRef\]](#)
59. Wang, J.; Hou, J.; Chen, J.; Fu, Q.; Huang, G. Data Mining Approach for Improving the Optimal Control of HVAC Systems: An Event-Driven Strategy. *J. Build. Eng.* **2021**, *39*, 102246. [\[CrossRef\]](#)
60. Kannari, L.; Kantorovitch, J.; Piira, K.; Piippo, J. Energy Cost Driven Heating Control with Reinforcement Learning. *Buildings* **2023**, *13*, 427. [\[CrossRef\]](#)
61. Aria, H.; Akbari, H. Integrated and Multi-Hour Optimization of Office Building Energy Consumption and Expenditure. *Energy Build.* **2014**, *82*, 391–398. [\[CrossRef\]](#)
62. Indraganti, M.; Ooka, R.; Rijal, H.B. Thermal Comfort in Offices in India: Behavioral Adaptation and the Effect of Age and Gender. *Energy Build.* **2015**, *103*, 284–295. [\[CrossRef\]](#)
63. Parkinson, T.; Schiavon, S.; De Dear, R.; Brager, G. Overcooling of Offices Reveals Gender Inequity in Thermal Comfort. *Sci. Rep.* **2021**, *11*, 23684. [\[CrossRef\]](#)
64. Shahzad, S.; Calautit, J.K.; Hughes, B.R.; Satish, B.K.; Rijal, H.B. Patterns of Thermal Preference and Visual Thermal Landscaping Model in the Workplace. *Appl. Energy* **2019**, *255*, 113674. [\[CrossRef\]](#)

65. Luo, M.; de Dear, R.; Ji, W.; Bin, C.; Lin, B.; Ouyang, Q.; Zhu, Y. The Dynamics of Thermal Comfort Expectations: The Problem, Challenge and Implication. *Build. Environ.* **2016**, *95*, 322–329. [\[CrossRef\]](#)
66. Yang, J.; Nam, I.; Sohn, J.-R. The Influence of Seasonal Characteristics in Elderly Thermal Comfort in Korea. *Energy Build.* **2016**, *128*, 583–591. [\[CrossRef\]](#)
67. Zhang, N.; Cao, B.; Wang, Z.; Zhu, Y.; Lin, B. A Comparison of Winter Indoor Thermal Environment and Thermal Comfort between Regions in Europe, North America, and Asia. *Build. Environ.* **2017**, *117*, 208–217. [\[CrossRef\]](#)
68. Yang, Z.; Becerik-Gerber, B. The Coupled Effects of Personalized Occupancy Profile Based HVAC Schedules and Room Reassignment on Building Energy Use. *Energy Build.* **2014**, *78*, 113–122. [\[CrossRef\]](#)
69. Rahman, H.; Han, H. Bayesian Estimation of Occupancy Distribution in a Multi-Room Office Building Based on CO₂ Concentrations. *Build. Simul.* **2018**, *11*, 575–583. [\[CrossRef\]](#)
70. Motuzienė, V.; Lapinskienė, V.; Rynkun, G. Optimizing Ventilation Systems for Sustainable Office Buildings: Long-Term Monitoring and Environmental Impact Analysis. *Sustainability* **2024**, *16*, 984. [\[CrossRef\]](#)
71. Gilani, S.; O'Brien, W. Exploring the Impact of Office Building Users' Modeling Approaches on Energy Use across Canadian Climates. *Energy Build.* **2019**, *197*, 68–86. [\[CrossRef\]](#)
72. Capozzoli, A.; Piscitelli, M.S.; Gorrino, A.; Ballarini, I.; Corrado, V. Data Analytics for Occupancy Pattern Learning to Reduce the Energy Consumption of HVAC Systems in Office Buildings. *Sustain. Cities Soc.* **2017**, *35*, 191–208. [\[CrossRef\]](#)
73. Chen, Y.; Liang, X.; Hong, T.; Luo, X. Simulation and Visualization of Energy-Related Occupant Behavior in Office Buildings. *Build. Simul.* **2017**, *10*, 785–798. [\[CrossRef\]](#)
74. Guo, Z.; Wang, X.; Wang, Y.; Zhu, F.; Zhou, H.; Zhang, M.; Wang, Y. The Cost-Optimal Control of Building Air Conditioner Loads Based on Machine Learning: A Case Study of an Office Building in Nanjing. *Buildings* **2024**, *14*, 3040. [\[CrossRef\]](#)
75. Lu, X.; Adetola, V.A.; Bhattacharya, S. Large-Scale Simulation-Based Parametric Analysis of an Optimal Precooling Strategy for Demand Flexibility in a Commercial Office Building. *Energy Build.* **2024**, *316*, 114284. [\[CrossRef\]](#)
76. Huang, B.; Huang, S.; Ma, X.; Katipamula, S.; Wu, D.; Lutes, R. Stochastic Scheduling for Commercial Building Cooling Systems: Considering Uncertainty in Zone Temperature Prediction. *Appl. Energy* **2023**, *346*, 121367. [\[CrossRef\]](#)
77. Kim, Y.-J. Optimal Price Based Demand Response of HVAC Systems in Multizone Office Buildings Considering Thermal Preferences of Individual Occupants Buildings. *IEEE Trans. Ind. Inf.* **2018**, *14*, 5060–5073. [\[CrossRef\]](#)
78. Guo, K.; Li, Q.; Zhang, L.; Wu, X. BIM-Based Green Building Evaluation and Optimization: A Case Study. *J. Clean. Prod.* **2021**, *320*, 128824. [\[CrossRef\]](#)
79. Batten, J.A.; Maddox, G.E.; Young, M.R. Does Weather, or Energy Prices, Affect Carbon Prices? *Energy Econ.* **2021**, *96*, 105016. [\[CrossRef\]](#)
80. Xu, H.; Pan, X.; Li, J.; Feng, S.; Guo, S. Comparing the Impacts of Carbon Tax and Carbon Emission Trading, Which Regulation Is More Effective? *J. Environ. Manag.* **2023**, *330*, 117156. [\[CrossRef\]](#)
81. Re, L.L.; Lee, C.; Cassisa, C.; Zhang, W.; Moarif, S. *Implementing Effective Emissions Trading Systems: Lessons From International Experiences*; IEA Publications: Paris, France, 2020; ISBN 92-64-96534-3.
82. Christiansen, A.C.; Arvanitakis, A.; Tangen, K.; Hasselknippe, H. Price Determinants in the EU Emissions Trading Scheme. *Clim. Policy* **2005**, *5*, 15–30. [\[CrossRef\]](#)
83. Spaargaren, G.; Mol, A.P.J. Carbon Flows, Carbon Markets, and Low-Carbon Lifestyles: Reflecting on the Role of Markets in Climategovernance. *Environ. Politics* **2013**, *22*, 174–193. [\[CrossRef\]](#)
84. Ding, G.; Deng, Y.; Lin, S. A Study on the Classification of China's Provincial Carbon Emissions Trading Policy Instruments: Taking Fujian Province as an Example. *Energy Rep.* **2019**, *5*, 1543–1550. [\[CrossRef\]](#)
85. Hare, W.; Stockwell, C.; Flachsland, C.; Oberthur, S. The Architecture of the Global Climate Regime: A Top-down Perspective. *Clim. Policy* **2010**, *10*, 600–614. [\[CrossRef\]](#)
86. Touzani, S.; Prakash, A.K.; Wang, Z.; Agarwal, S.; Pritoni, M.; Kiran, M.; Brown, R.; Granderson, J. Controlling Distributed Energy Resources via Deep Reinforcement Learning for Load Flexibility and Energy Efficiency. *Appl. Energy* **2021**, *304*, 117733. [\[CrossRef\]](#)
87. Silvestri, A.; Coraci, D.; Brandi, S.; Capozzoli, A.; Borkowski, E.; Kohler, J.; Wu, D.; Zeilinger, M.N.; Schlueter, A. Real Building Implementation of a Deep Reinforcement Learning Controller to Enhance Energy Efficiency and Indoor Temperature Control. *Appl. Energy* **2024**, *368*, 123447. [\[CrossRef\]](#)
88. Kathirgamanathan, A.; Mangina, E.; Finn, D.P. Development of a Soft Actor Critic Deep Reinforcement Learning Approach for Harnessing Energy Flexibility in a Large Office Building. *Energy AI* **2021**, *5*, 100101. [\[CrossRef\]](#)
89. Biemann, M.; Scheller, F.; Liu, X.; Huang, L. Experimental Evaluation of Model-Free Reinforcement Learning Algorithms for Continuous HVAC Control. *Appl. Energy* **2021**, *298*, 117164. [\[CrossRef\]](#)
90. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In Proceedings of the International conference on machine learning, Pmlr, Stockholmssmässan, Stockholm Sweden, 10–15 July 2018; pp. 1861–1870.

91. Park, Y.; Jun, W.; Lee, S. A Comparative Study of Deep Reinforcement Learning Algorithms for Urban Autonomous Driving: Addressing the Geographic and Regulatory Challenges in CARLA. *Appl. Sci.* **2025**, *15*, 6838. [\[CrossRef\]](#)
92. Zhao, Y.; Lin, F.; Yang, Z. Soft Actor-Critic-Based Energy Management Strategy for Hybrid Energy Storage System in Urban Rail Transit DC Traction Power Supply System. *Electr. Power Syst. Res.* **2026**, *252*, 112372. [\[CrossRef\]](#)
93. Lu, Y.; Wang, S.; Sun, Y.; Yan, C. Optimal Scheduling of Buildings with Energy Generation and Thermal Energy Storage under Dynamic Electricity Pricing Using Mixed-Integer Nonlinear Programming. *Appl. Energy* **2015**, *147*, 49–58. [\[CrossRef\]](#)
94. Lou, Y.; Yang, Y.; Ye, Y.; He, C.; Zuo, W. The Economic Impacts of Carbon Emission Trading Scheme on Building Retrofits: A Case Study with U.S. Medium Office Buildings. *Build. Environ.* **2022**, *221*, 109311. [\[CrossRef\]](#)
95. Zeng, T.; Barooah, P. An Autonomous MPC Scheme for Energy-Efficient Control of Building HVAC Systems. In Proceedings of the 2020 American Control Conference (ACC), Denver, CO, USA, 1–3 July 2020; pp. 4213–4218.
96. Avci, M.; Erkoç, M.; Rahmani, A.; Asfour, S. Model Predictive HVAC Load Control in Buildings Using Real-Time Electricity Pricing. *Energy Build.* **2013**, *60*, 199–209. [\[CrossRef\]](#)
97. Zhang, Z.; Chong, A.; Pan, Y.; Zhang, C.; Lam, K.P. Whole Building Energy Model for HVAC Optimal Control: A Practical Framework Based on Deep Reinforcement Learning. *Energy Build.* **2019**, *199*, 472–490. [\[CrossRef\]](#)
98. Zhang, Y.; Teoh, B.K.; Wu, M.; Chen, J.; Zhang, L. Data-Driven Estimation of Building Energy Consumption and GHG Emissions Using Explainable Artificial Intelligence. *Energy* **2023**, *262*, 125468. [\[CrossRef\]](#)
99. Welcome to LightGBM's Documentation!—LightGBM 4.6.0 Documentation. Available online: <https://lightgbm.readthedocs.io/en/stable/index.html> (accessed on 8 November 2025).
100. Boyce, J.K. Carbon Pricing: Effectiveness and Equity. *Ecol. Econ.* **2018**, *150*, 52–61. [\[CrossRef\]](#)
101. Narassimhan, E.; Gallagher, K.S.; Koester, S.; Alejo, J.R. Carbon Pricing in Practice: A Review of Existing Emissions Trading Systems. *Clim. Policy* **2018**, *18*, 967–991. [\[CrossRef\]](#)
102. Steinebach, Y.; Fernández-i-Marín, X.; Aschenbrenner, C. Who Puts a Price on Carbon, Why and How? A Global Empirical Analysis of Carbon Pricing Policies. *Clim. Policy* **2021**, *21*, 277–289. [\[CrossRef\]](#)
103. Biedenkopf, K.; Chen, Z.; De Clara, S.; Doda, B.; Eden, A.; Hall, M.; Heckmann, L.; Holovko, I.; Hynes, D.; Kehrer, M. *Emissions Trading Worldwide: Status Report 2024*; ICAP: New Delhi, India, 2024.
104. Rangaswamy, D.R.; Ramamurthy, K. Evaluation of Eight Thermal Comfort Indices Based on Perception Survey for a Hot-Humid Climate through a Naturally Ventilated Apartment. *J. Archit. Eng.* **2021**, *27*, 04021041. [\[CrossRef\]](#)
105. He, Y.; Li, N.; Peng, J.; Zhang, W.; Li, Y. Field Study on Adaptive Comfort in Air Conditioned Dormitories of University with Hot-Humid Climate in Summer. *Energy Build.* **2016**, *119*, 1–12. [\[CrossRef\]](#)
106. Ning, H.; Wang, Z.; Ji, Y. Thermal History and Adaptation: Does a Long-Term Indoor Thermal Exposure Impact Human Thermal Adaptability? *Appl. Energy* **2016**, *183*, 22–30. [\[CrossRef\]](#)
107. Zhang, Y.; Wang, J.; Chen, H.; Zhang, J.; Meng, Q. Thermal Comfort in Naturally Ventilated Buildings in Hot-Humid Area of China. *Build. Environ.* **2010**, *45*, 2562–2570. [\[CrossRef\]](#)
108. Hu, Y.; Chi, Y.; Zhou, W.; Li, J.; Wang, Z.; Yuan, Y. The Interactions between Renewable Portfolio Standards and Carbon Emission Trading in China: An Evolutionary Game Theory Perspective. *Energy* **2023**, *271*, 127047. [\[CrossRef\]](#)
109. Dedesko, S.; Pendleton, J.; Young, A.S.; Coull, B.A.; Spengler, J.D.; Allen, J.G. New Considerations for Representing Moisture in Indoor Thermal Conditions: Associations between Enthalpy, Cognitive Performance, and Thermal Sensations. *Indoor Environ.* **2025**, *2*, 100098. [\[CrossRef\]](#)
110. Masterton, J.M.; Richardson, F.A. *Humidex: A Method of Quantifying Human Discomfort Due to Excessive Heat and Humidity*; Environment Canada, Atmospheric Environment Service: Toronto, ON, Canada, 1981.
111. Ahmad, M.; Li, X.F.; Wu, Q. Carbon Taxes and Emission Trading Systems: Which One Is More Effective in Reducing Carbon Emissions?—A Meta-Analysis. *J. Clean. Prod.* **2024**, *476*, 143761. [\[CrossRef\]](#)
112. Haywood, L.; Jakob, M. The Role of the Emissions Trading Scheme 2 in the Policy Mix to Decarbonize Road Transport in the European Union. *Transp. Policy* **2023**, *139*, 99–108. [\[CrossRef\]](#)
113. Günther, C.; Pahle, M.; Govorukha, K.; Osorio, S.; Fotiou, T. Carbon Prices on the Rise? Shedding Light on the Emerging Second EU Emissions Trading System (EU ETS 2). *Clim. Policy* **2025**, 1–12. [\[CrossRef\]](#)
114. Edenhofer, O.; Jakob, M.; Creutzig, F.; Flachsland, C.; Fuss, S.; Kowarsch, M.; Lessmann, K.; Mattauch, L.; Siegmeier, J.; Steckel, J.C. Closing the Emission Price Gap. *Glob. Environ. Change* **2015**, *31*, 132–143. [\[CrossRef\]](#)
115. 2024 State of the Voluntary Carbon Market (SOVCM). Available online: https://3298623.fs1.hubspotusercontent-na1.net/hubfs/3298623/SOVCM%202024/State_of_the_Voluntary_Carbon_Markets_20240529%201.pdf (accessed on 1 October 2025).
116. Guo, C.; Wang, X.; Zheng, Y.; Zhang, F. Real-Time Optimal Energy Management of Microgrid with Uncertainties Based on Deep Reinforcement Learning. *Energy* **2022**, *238*, 121873. [\[CrossRef\]](#)
117. Xu, Y.; Zhao, Z.; Cheng, P.; Chen, Z.; Ding, M.; Vucetic, B.; Li, Y. Constrained Reinforcement Learning for Resource Allocation in Network Slicing. *IEEE Commun. Lett.* **2021**, *25*, 1554–1558. [\[CrossRef\]](#)

118. Wang, D.; Zheng, W.; Wang, Z.; Wang, Y.; Pang, X.; Wang, W. Comparison of Reinforcement Learning and Model Predictive Control for Building Energy System Optimization. *Appl. Therm. Eng.* **2023**, *228*, 120430. [\[CrossRef\]](#)
119. Mokhtari, R.; Montazeri, M.; Cai, H.; Heer, P.; Li, R. Price-Responsive Control Using Deep Reinforcement Learning for Heating Systems: Simulation and Living Lab Experiment. *Energy* **2025**, *337*, 138517. [\[CrossRef\]](#)
120. Havenith, G.; Fiala, D. Thermal Indices and Thermophysiological Modeling for Heat Stress. *Compr. Physiol.* **2011**, *6*, 255–302. [\[CrossRef\]](#)
121. Rana, R.; Kusy, B.; Jurdak, R.; Wall, J.; Hu, W. Feasibility Analysis of Using Humidex as an Indoor Thermal Comfort Predictor. *Energy Build.* **2013**, *64*, 17–25. [\[CrossRef\]](#)
122. Yang, L.; Yan, H.; Lam, J.C. Thermal Comfort and Building Energy Consumption Implications—A Review. *Appl. Energy* **2014**, *115*, 164–173. [\[CrossRef\]](#)
123. Agostinelli, F.; Hocquet, G.; Singh, S.; Baldi, P. From Reinforcement Learning to Deep Reinforcement Learning: An Overview. In Proceedings of the Braverman Readings in Machine Learning. Key Ideas from Inception to Current State: International Conference Commemorating the 40th Anniversary of Emmanuil Braverman’s Decease, Boston, MA, USA, 28–30 April 2017; Invited Talks. Springer: Berlin/Heidelberg, Germany, 2018; pp. 298–328.
124. Kong, D.; Liu, H.; Wu, Y.; Li, B.; Wei, S.; Yuan, M. Effects of Indoor Humidity on Building Occupants’ Thermal Comfort and Evidence in Terms of Climate Adaptation. *Build. Environ.* **2019**, *155*, 298–307. [\[CrossRef\]](#)
125. Xu, Y.; Hu, D.; Liang, L.; McAleer, S.; Abbeel, P.; Fox, R. Target Entropy Annealing for Discrete Soft Actor-Critic. *arXiv* **2021**, arXiv:2112.02852. [\[CrossRef\]](#)
126. Han, S.; Zhou, W.; Lü, S.; Zhu, S.; Gong, X. Entropy Regularization Methods for Parameter Space Exploration. *Inf. Sci.* **2023**, *622*, 476–489. [\[CrossRef\]](#)
127. Arora, J.S. *Introduction to Optimum Design*; Elsevier: Amsterdam, The Netherlands, 2004; ISBN 0-08-047025-4.
128. He, N.; Liu, L.; Qian, C.; Zhang, L.; Yang, Z.; Li, S. A Closed-Loop Data-Fusion Framework for Air Conditioning Load Prediction Based on LBF. *Energy Rep.* **2022**, *8*, 7724–7734. [\[CrossRef\]](#)
129. Antunes, M.; Estro, T.; Bhandari, P.; Gandhi, A.; Kuenning, G.; Liu, Y.; Waldspurger, C.; Wildani, A.; Zadok, E. Kneeliverse: A Universal Knee-Detection Library for Performance Curves. *SoftwareX* **2025**, *30*, 102161. [\[CrossRef\]](#)
130. Antunes, M.; Ribeiro, J.; Gomes, D.; Aguiar, R.L. Knee/Elbow Point Estimation through Thresholding. In Proceedings of the 2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud), Barcelona, Spain, 6–8 August 2018; pp. 413–419.
131. Sarbu, I.; Pacurar, C. Experimental and Numerical Research to Assess Indoor Environment Quality and Schoolwork Performance in University Classrooms. *Build. Environ.* **2015**, *93*, 141–154. [\[CrossRef\]](#)
132. Perino, G.; Ritz, R.A.; van Benthem, A.A. Overlapping Climate Policies. *Econ. J.* **2025**, *135*, 2122–2160. [\[CrossRef\]](#)
133. Van den Bergh, K.; Delarue, E.; D’haeseleer, W. Impact of Renewables Deployment on the CO₂ Price and the CO₂ Emissions in the European Electricity Sector. *Energy Policy* **2013**, *63*, 1021–1031. [\[CrossRef\]](#)
134. Bird, L.; Chapman, C.; Logan, J.; Sumner, J.; Short, W. Evaluating Renewable Portfolio Standards and Carbon Cap Scenarios in the US Electric Sector. *Energy Policy* **2011**, *39*, 2573–2585. [\[CrossRef\]](#)
135. del Río González, P. The Interaction between Emissions Trading and Renewable Electricity Support Schemes. An Overview of the Literature. *Mitig. Adapt. Strateg. Glob. Change* **2007**, *12*, 1363–1390. [\[CrossRef\]](#)
136. Enescu, D. A Review of Thermal Comfort Models and Indicators for Indoor Environments. *Renew. Sustain. Energy Rev.* **2017**, *79*, 1353–1379. [\[CrossRef\]](#)
137. ISO 7730:2025; Moderate Thermal Environments Determination of the PMV and PPD Indices and Specification of the Conditions for Thermal Comfort. ISO: Geneva, Switzerland, 2025.
138. ANSI/ASHRAE Standard 55; Thermal Environmental Conditions for Human Occupancy. ASHRAE: Peachtree Corners, GA, USA, 1992.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.