Journal of Building Performance Simulation

ISSN: 1940-1493 (Print) 1940-1507 (Online) Journal homepage: www.tandfonline.com/journals/tbps20

Online transfer learning (OTL) for accelerating deep reinforcement learning (DRL) for building energy management

Tran Van Quang & Dat Tien Doan

To cite this article: Tran Van Quang & Dat Tien Doan (02 Jun 2025): Online transfer learning (OTL) for accelerating deep reinforcement learning (DRL) for building energy management, Journal of Building Performance Simulation, DOI: 10.1080/19401493.2025.2511826

To link to this article: https://doi.org/10.1080/19401493.2025.2511826

Published online: 02 Jun 2025.

JOURNAL OF BUILDING PERFORMANCE SIMULATION
https://doi.org/10.1080/19401493.2025.2511826

Online transfer learning (OTL) for accelerating deep reinforcement learning (DRL) for building energy management

Tran Van Quang[a] and Dat Tien Doan[b]

[a]School of Architecture and Planning, University of Texas at San Antonio, San Antonio, Texas, 78249, USA; [b]Department of Built Environment Engineering, School of Future Environments, Auckland University of Technology, Auckland, 1010, New Zealand

ABSTRACT

Buildings account for over one-third of global energy consumption and emissions, primarily from heating and cooling operations. Intelligent optimisation through predictive controls, such as deep reinforcement learning (DRL), offers significant potential for energy efficiency. However, DRL faces challenges in generalisation and impractical retraining when applied to different buildings, limiting its scalability. Prior online transfer learning (OTL) approaches relied on simulation or rule-based methods but lacked live learning and real-time optimisation. This study proposes an OTL strategy combining autonomous simulation-based DRL policy pretraining with real-time fine-tuning for rapid adaptation to new buildings. Using the Soft Actor-Critic (SAC) algorithm, it was tested on commercial building energy management simulations. Results showed 18%+ reductions in HVAC energy consumption and 8%+ improvement in thermal comfort compared to rule-based and non-transfer DRL baselines. Empirical validation highlights OTL's potential in overcoming DRL's cold start and training burdens, paving the way for broader deployment in sustainable energy management.

ARTICLE HISTORY
Received 2 February 2025
Accepted 22 May 2025

KEYWORDS
Online transfer learning (OTL); deep reinforcement learning (DRL); office building; building energy consumption

1. Introduction

The growing global energy consumption of buildings has become a significant concern, accounting for over one-third of total energy usage and nearly 40% of energy-related carbon dioxide (CO_2) emissions worldwide in 2018 (IEA 2018). This trend is expected to continue due to population and economic growth, urbanization, improved energy access, and increased building ownership (Change IP on C 2015). Buildings may constitute over 50% of 2060 emissions in baseline scenarios, presenting risks of locking in intensive utilization patterns (Riahi et al. 2015). These estimates underscore buildings as one of the critical contributors to anthropogenic climate change (Lucon et al. 2014). Evidence suggests the building sector ranks among the highest hard-to-abate emissions sources on par with industry and transport (Feng et al. 2023).

Significantly, these emissions directly relate to heatwaves, droughts, and other extreme weather attributable to human contribution events affecting communities worldwide due to climate change (Philip et al. 2022; Vautard et al. 2020). Over 90% of the last 50 years of warming stems from human activities like building operations, among other sectors (Pörtner et al. 2022). The recent 2022 Intergovernmental Panel on Climate Change report warns of irreversible hot extremes if warming exceeds 1.5°C levels in the near term (Masson-Delmotte 2018). As per the World Meteorological Organization and the United Nations, global average temperatures have risen over 1.1°C above pre-industrial levels as of 2020, increasingly risking severe climate change repercussions (Herring et al. 2021).

With buildings significantly contributing to global heating through energy usage and emissions, improving building energy efficiency is critical for climate change mitigation. Technical optimizations in building design, operation, and systems hold the potential to contribute over 50% of total CO_2 mitigation opportunities by 2040, substantially limiting warming scenarios (Rogelj et al. 2018; Ürge-Vorsatz et al. 2018). The urgency of reducing building climate impacts further escalates in a developing world context, accounting for the most aggressive vertical development (Agency IE 2018). Simultaneously, these solutions can alleviate household energy burdens. Building systems like Heating, ventilation, and air conditioning (HVAC) comprise most of the usage, so their efficient control offers opportunities.

CONTACT Dat Tien Doan dat.doan@aut.ac.nz

1.1. Related work

HVAC systems comprise the dominant share of commercial building energy end-use, exceeding 73% per Commercial Buildings Energy Consumption Survey (CBECS) 2018 surveys (EIA 2018). As per national HVAC, electricity utilization is approaching 1,390 TWh annually across US commercial buildings alone, as per the US Energy Information Administration (Ürge-Vorsatz et al. 2015). This consumption underlines opportunities through advanced control of HVAC operations for carbon footprint reduction without compromising indoor environmental quality (ASHRAE 2019; Feng et al. 2019; Pérez-Lombard, Ortiz, and Pout 2008). Total US HVAC utility costs approach $40 billion annually, further driving economic incentives for efficiency gains (Liu et al. 2018).

Rule-based, model-predictive, and machine-learning control techniques have been widely investigated for building system optimization (Ding et al. 2021; Dong and Andrews 2009; Hsu 2015; Kleiminger, Mattern, and Santini 2014; Macqueen 1967). Rule-based methods rely on predefined temperature setpoints, operating schedules, and equipment sequencing logic programmed based on standard practices (Park et al. 2019; Yu et al. 2020). However, such static, predetermined control lacks the adaptability to changing conditions. Model predictive control (MPC) utilizes dynamic programming and optimization algorithms to predict future states and find optimal control actions accordingly (Goyal, Ingley, and Barooah 2013; Oldewurtel et al. 2012). However, MPC accuracy depends heavily on underlying model fidelity. Reinforcement learning (RL) overcomes these limitations by offering a model-free, data-driven self-optimization approach (Lissa et al. 2021; Wang and Hong 2020). The RL paradigm provides a framework for an agent to sequentially interact with a building by taking various actions while receiving corresponding scalar rewards or penalties reflective of the outcomes (Pippia et al. 2019; Prívara et al. 2011). It can refine control decisions through experience to enhance objectives like operational costs, occupant comfort, or grid services without requiring accurate models. RL has shown success across applications like HVAC optimization (Lee et al. 2013; Turley et al. 2020), intelligent storage control (Chen et al. 2020), and renewables integration (Taylor and Stone 2009), among others. DRL capacity for complex tasks like climate control by approximating state-action functions using deep neural networks (Ammar et al. 2014; Parisotto, Ba, and Salakhutdinov 2016; Quang and Phuong 2024).

However, DRL approaches for building energy management adopt certain limitations. These include extensive training requirements over months or seasons before achieving satisfactory performance as controllers learn from interactions (Wang and Hong 2020). The prolonged learning periods negatively impact user experience. Additionally, learned policies fail to generalize across buildings with moderate weather, construction, occupancy patterns, or equipment differences. This necessitates full retraining, which causes delays in deployment and carbon impact reduction (Gupta et al. 2017). Consequently, such factors significantly impede scalable adoption. Transfer learning (TL) methods offer prospective solutions to mitigate these challenges. TL involves repurposing knowledge from related tasks or domains, like reference simulated buildings. Applying TL can minimize redundant interactions when deploying controllers in new, unseen buildings (Lazaric 2012).

TL presents a viable approach to facilitating and augmenting learning on associated tasks using knowledge obtained from training on a distinct task (Olivas et al. 2009). In contrast to conventional techniques that commence with randomly initialized models, TL encompasses transferring acquired representations or parameters to a novel model. This methodology mitigates the necessity for extensive retraining, conferring enhanced performance given fewer interactions, which proves especially beneficial in contexts with constrained data accessibility. TL encompasses myriad applications, including transferring policies across RL agents in robotics, autonomous systems, and game-playing (Lazaric 2012). While offline simulation-based TL for RL building controls has been researched (Fang et al. 2023; Xu et al. 2020), exploring online adaptation in real building settings remains relatively limited.

Earlier TL applications in deep reinforcement learning (DRL) focused on forecasting building energy demand with limited historical data. For instance, E. Mocanu et al. (2016) transferred models trained on source buildings using algorithms like State-action-reward-state-action (SARSA) and Q-learning to estimate demand on target buildings lacking substantial labelled data. Ismail and Baysal (2023) reused known user electricity price elasticity from regions with existing demand response implementations to inform policy training via SARSA for unfamiliar areas. It facilitated the integration of demand flexibility for new consumers without undergoing prolonged on-site learning.

Beyond forecasting, TL holds the potential to refine control policies for assets like HVAC systems. Notable examples adopt hybrid simulation-based learning (Costanzo et al. 2016), augment online adaptation with expert systems (Costanzo et al. 2016; Ruelens et al. 2017) or leverage domain randomization (Didden et al. 2022; Peirelinck et al. 2021). Key strategies involve pretraining on surrogate models before deployment, integrating monotonicity constraints from prior domain

Table 1. Recent articles on TL of DRL for smart building applications.

Ref	Year	Approach	Data sources	Model type	Potential limitation
Coraci et al. (2023)	2023	Proposes an online transfer learning (OTL) strategy to transfer a pre-trained DRL control policy from a source building to target buildings	Simulation (EnergyPlus)	Soft Actor-Critic (SAC)	It uses simulations only, so it may not fully capture real-world complexity. Only considers homogeneous transfer. A limited number of target building configurations were evaluated
Esrafilian-Najafabadi and Haghighat (2023)	2023	TL approach combines unsupervised clustering of occupancy profiles with DRL HVAC control to improve initial learning performance	Occupancy data from 26 households was collected via PIR sensors.	Unsupervised k-means; Deep Q Learning (DQN)	Limited factors for similarity/control
Xu et al. (2020)	2020	Two-network TL approach for DRL-based HVAC control	Simulation (EnergyPlus)	DQN	Simulation only. No internal heat. Limited weather transfer. Simple ON-OFF data
Genkin and McArthur (2024)	2024	Addresses cold start using DRL & TL between buildings	Sensor and energy data from 3 real buildings	ReLBOT	Only three buildings focused on cooling. Simulation testing, not real-world
Tao, Qiu, and Lai (2022)	2022	Hybrid cloud-edge control strategy for DR using DRL and TL	Simulated power system and building environment based on MATPOWER and OpenAI Gym	Continuous Dueling Deep Q-learning	Validation only on simulations. Potential issues in scaling, Limited to BESS and HVAC, Real-world factors are not fully considered
Lissa et al. (2021)	2021	Proposes RL-based control of heat pumps to minimize costs while maintaining temperatures. Tests adding TL between houses	Simulated data based on real sensors. Historical PV data	DQN	Validated via simulation only. Tested on DHW only

knowledge, or artificially diversifying source training environments. Reported benefits include accelerated convergence, improved accuracy, and enhanced cost savings. Recent works further repurpose structural policy similarities across devices like batteries and HVAC units to speed up distributed RL (Tao, Qiu, and Lai 2022). Table 1 shows recent studies on developing a TL of DRL for smart building applications. Although significant progress has been made in applying TL to DRL for building energy management, several important challenges remain. Existing studies predominantly validated their methods within limited simulation settings, which, while valuable for controlled investigation, often lacked evaluation across heterogeneous building types, climates, and operational conditions. As shown in Table 1, many approaches focused on offline pretraining alone, without incorporating mechanisms for real-time online adaptation during deployment, which is critical to account for real-world building dynamics. Furthermore, evaluations were frequently restricted to small numbers of target buildings with similar characteristics, limiting the generalizability and scalability of the proposed methods.

1.2. Research aims and contribution.

This study proposes an OTL strategy to significantly improve DRL controllers' scalability, adaptability, and real-world deployment for building energy management systems. While DRL offers a promising model-free framework for HVAC optimization, it suffers from limited transferability and high retraining costs when applied across diverse building contexts. This research addresses these gaps through a hybrid methodology integrating simulation-based pretraining with real-time fine-tuning. The specific objectives of this study are to:

- Design and train a DRL control agent using the SAC algorithm on a source building to optimize energy cost and maintain thermal comfort.
- Develop an OTL framework to rapidly adapt the pretrained DRL agent to target buildings with differing envelopes, loads, and climate conditions.
- Rigorously evaluate the performance of the OTL approach in comparison with rule-based control and DRL without transfer across five physically heterogeneous building simulations.
- Analyze the influence of source-target mismatches on policy transferability and identify implementation challenges in real-world settings.

The core contribution of this study lies in demonstrating a real-time, adaptive OTL framework capable of overcoming DRL's cold-start limitations and reducing training burdens across diverse building types. Unlike prior studies, which are typically constrained to offline simulations or homogeneous evaluation settings, this work performs empirical benchmarking with statistical validation across varied environments. These contributions provide

an essential step toward scalable and transferable AI-based control strategies in building operations, supporting broader sustainability and climate-resilience objectives in the built environment.

2. Methodology

2.1. Control optimization framework

The research framework of this study is shown in Figure 1. This study develops an advanced control strategy for building cooling systems using DRL. DRL enables autonomous optimization of policies to optimize complex objectives through trial-and-error learning within an environment (Sutton and Barto 2018). The core idea is an agent gradually improves its actions within an environment to maximize cumulative rewards over time through experience. This study utilizes SAC, an off-policy DRL algorithm suitable for problems with high-dimensional state spaces and discrete action sets (Haarnoja et al. 2018). SAC converges to optimal stochastic policies by optimizing a maximum entropy objective to encourage sufficient exploration during training (Haarnoja et al. 2017).

The methodology involves training a deep neuro controller using SAC in an EnergyPlus simulated source office building and cooling system environment. The RL agent learns an optimized policy to minimize composite costs and thermal comfort objectives by interacting with the virtual building. Subsequently, the trained policy parameters are transferred to a distinct set of simulated target-building models via OTL. This directly reuses the learned source knowledge while fine-tuning the control policy on the target building using only real-time data through a brief initial period of operation. Avoiding lengthy retraining enhances scalability. Comparative analysis evaluates the OTL approach against conventional rule-based control as well as non-transfer and offline DRL alternatives to analyze relative performance.

2.2. Case study building description

The simulated building model adapts a medium office archetype from the U.S. Department of Energy's Commercial Prototype Building Models (CPBM) (Thornton et al. 2011; Prototype Building Models Building Energy Codes Program n.d.) and enhances it by incorporating multi-zone configurations and dynamic occupancy. In this study, we utilized the Medium Office prototype building from the CPBM dataset. This building type, with a floor area of approximately 15,000m^2, is widely recognized as a representative archetype for evaluating HVAC control strategies in commercial office environments. It features a multi-zone layout, diversified occupancy schedules, and significant internal gains, making it well-suited for modelling realistic thermal and energy dynamics. The floor plan (Figure 2) has distinct thermal zones enabling efficient conditioning of primarily daylit perimeter spaces. The window-to-wall ratio is 33% with evenly distributed ribbon windows. The 3D model (Figure 3) is created in Rhinoceros (Rhinoceros n.d.).

Outdoor weather conditions use a typical meteorological year (TMY) file for Auckland, New Zealand (36.85^{0}S, 174.7645^{0}E). The layered envelope constructions in Table 2 employ concrete, insulation, and glazing typical of modern offices. Windows utilizes low-emissivity argon-filled double-pane glazing. Skylights integrate into the top floor roof.

The annual EnergyPlus simulations input parameters related to HVAC, internal loads, occupancy scheduling, and electricity pricing are summarized in Table 3. Relevant factors include temperature setpoints, ventilation, lighting and equipment densities, dynamic schedules, and time-of-use tariff structures. Dynamic occupancy schedules refer to time-resolved occupant presence modelled using 15-minute intervals, reflecting realistic office behaviour patterns. The weekday schedule included a rising occupancy density starting at 8:00 am, peaking between 10:00 am and 4:00 pm, and tapering off by 6:00 pm, while weekends were considered unoccupied.

The high-frequency raw data is preprocessed to generate the key inputs for the DRL framework, including state variables, action sets, and reward signals. The state space features include current and previous indoor/outdoor air temperatures, cooling storage state of charge (SOC), occupancy status, electricity price, time of day, and day of week. The cooling storage SOC is computed at each time-step as the ratio of available cooling in the tank to its total storage capacity. The action space is discretized into operation modes and cooling supply control actions. The reward function incorporates HVAC electricity cost and indoor temperature deviation penalty terms calculated from the utility tariff rate and simulation data.

2.3. Control system

Reinforcement learning provides a framework for an intelligent agent to make optimal decisions under uncertainty through iterative interactions with a complex environment. The underlying structure follows a Markov decision process (MDP), mathematically formalizing sequential decision-making problems using four core elements (van Otterlo and Wiering 2012). The states set (S) encapsulates dynamic variables, providing the agent with relevant observations of the environment system at time t, such as indoor temperature or electricity price for building energy management applications (Sutton and Barto

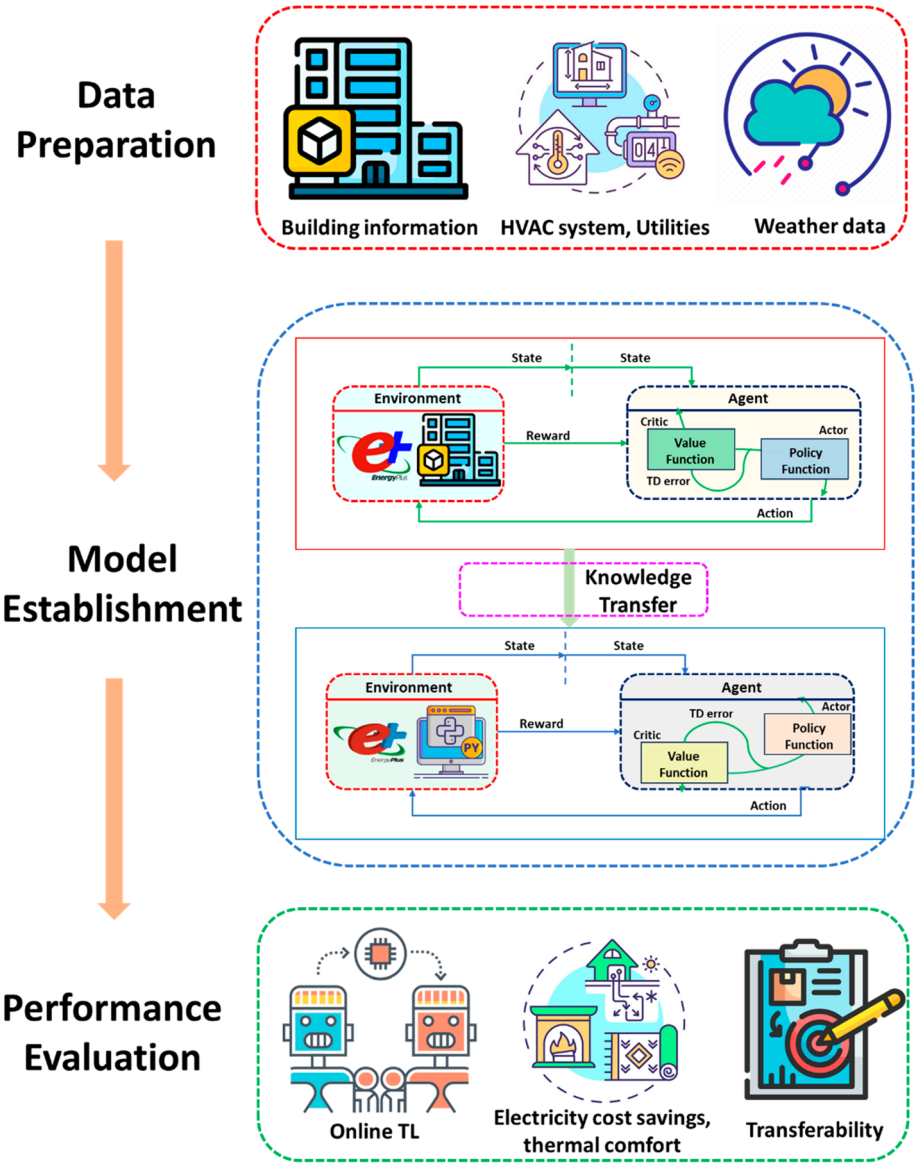


Figure 1. The framework of the proposed research.

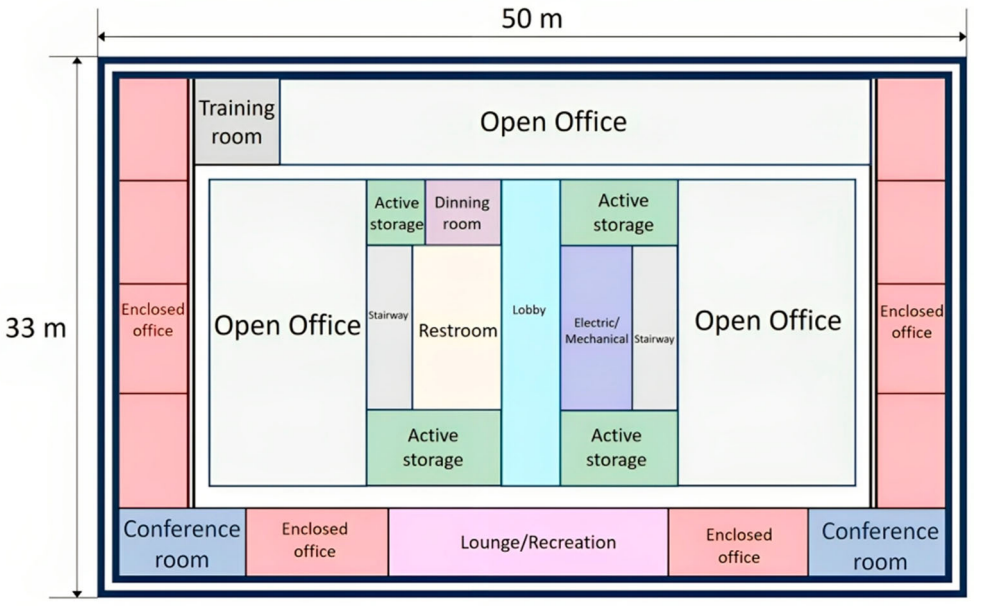
2018). The actions set (a) defines the possible decisions the agent can take to influence the environment at each time-step, whether altering HVAC operation modes or cooling supply regulation. The transition probability (p) governs the likelihood of progressing from one state s_t to the next state s_{t+1} based on the current state, selected action, and innate environment dynamics.

$$p(s_{t+1}, r|s_t, a) = \Pr\{R_{t+1} = r, S_{t+1} = |S_t, A_t\} \quad (1)$$

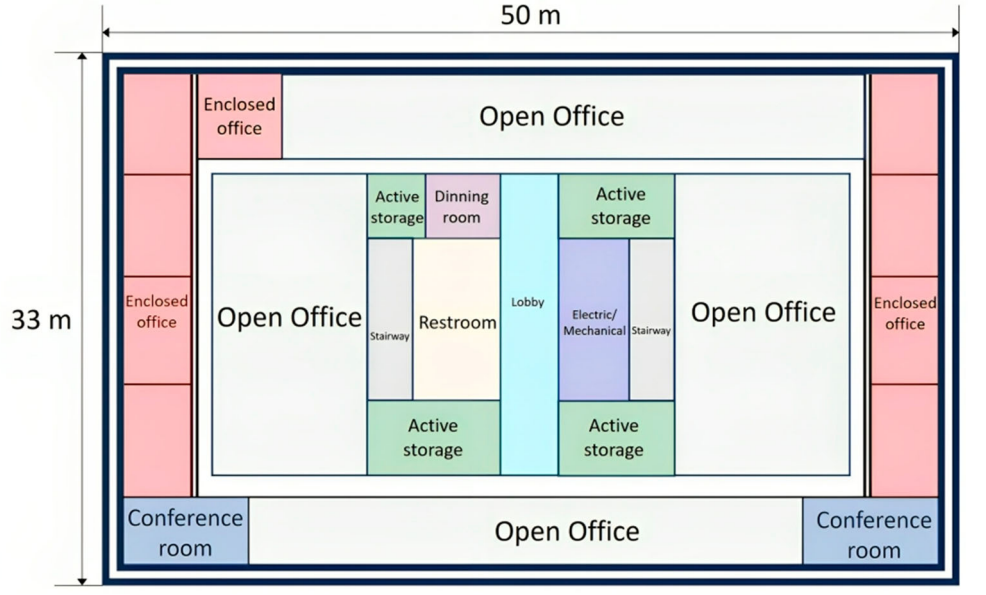
Finally, the reward function (R) provides scalar feedback to the agent quantifying the subjective desirability of action outcomes, imposing penalties on undesirable consequences like thermal discomfort and incentives for preferred behaviours like minimizing operational expenditure (Bellman 1952). By sequentially interacting with the environment system, the reinforcement learning agent aims to determine the optimal policy denoted π(a|s) that maps state to the actions expected to yield the highest attainable cumulative reward over the long term. This maximization relies on approximating the optimal state-action value function Q(s, a) for each decision using Bellman equations (Shteingart and Loewenstein 2014; Sutton and Barto 2018), considering immediate rewards plus appropriately discounted future rewards reachable from the following state by optimally acting thereafter.

$$q_* = \sum_{s_{t+1}, r} p(s_{t+1}, r|s_t, a) = \left[r + \gamma \max_a q_*(s_{t+1}, a)\right] \quad (2)$$

The control algorithm utilizes the SAC technique – an off-policy maximum entropy RL algorithm well-suited



(a) Ground floor



(b) Middle and Top floor

Figure 2. Zone configuration of the simulated medium office building (adapted from Zhihong Pang et al. 2020).

for problems with continuous action spaces and sparse rewards (Haarnoja et al. 2018). SAC has state-of-the-art performance on various constant control tasks. The underlying framework follows a MDP formulation comprising the critical elements of state space, action space, and reward function (Shteingart and Loewenstein 2014; Sutton and Barto 2018). As shown in Figure 4, the agent initially observes the environment state s_t, which captures relevant information like occupancy, temperature, and time. It then selects a control action a_t from the available discrete set based on its learned policy π. The environment transitions to a new state s_{t+1} while providing a scalar reward r_t as feedback on the action's desirability. The new state and reward update the agent's understanding, allowing it to improve its decisions over successive interactions.

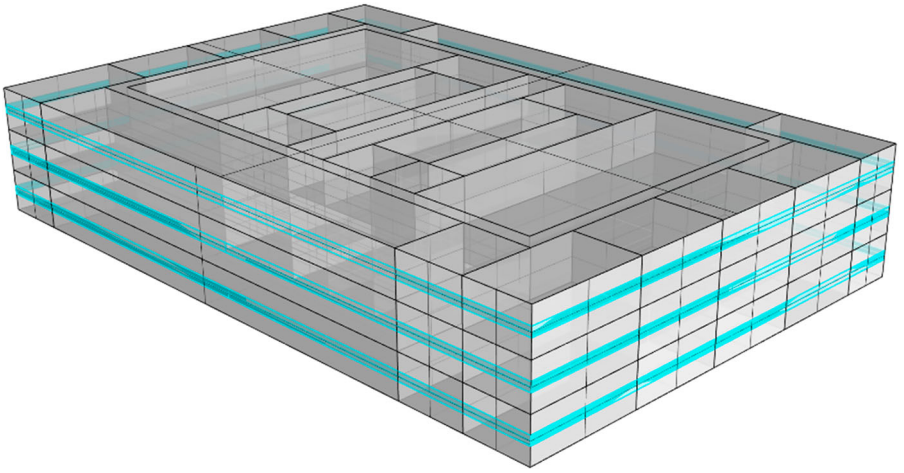


Figure 3. Medium office building model.

Table 2. Envelope component properties.

Component	Thickness (m)	Conductivity (W/mK)	Density (kg/m^3)	Specific Heat (J/kgK)
External wall	0.2	1.13	1850	1000
Internal wall	0.0125	0.16	800	1090
Ceiling	0.2	1.13	1850	1000
Floor	0.15	1.13	1850	1000
Window	0.006	0.9	2500	840

2.3.1. State space

The state space S consists of variables providing relevant observations of the environment system to the RL agent. For the building cooling system control problem, the state space S includes:

$$S = \{T_{in}(t), T_{in}(t-1), T_{in}(t-2), T_{out}(t), SOC(t), SOC(t-1), P(t), P(t+1), \ldots, P(t+24), O(t), O(t+1), \ldots O(t+24), T(t), D(t)\} \quad (3)$$

Indoor air temperature (Tin) – current and previous time-steps; Outdoor air temperature (Tout); Cooling storage tank SOC – current and prior time-steps; Electricity price (P) – current and predicted future prices; Occupancy status (O) – current and expected future occupancy; Time of day (T); Day of week (D).

The indoor temperature over previous time-steps accounts for thermal inertia. SOC over previous time-steps helps assess storage dynamics. Electricity price and

Table 3. Simulation parameters.

Parameters	Value	Typical range	References/Justification
Location	Auckland, New Zealand	N/A	Leverages TMY3 weather file for Auckland Int. Airport (ASN).
Time-step	15 min	1 min to 1 h	Captures thermal dynamics faster than hourly data while limiting computation vs 1 min data.
Total period	1 year	Months to years	Annual simulation to capture seasonal variations (Hoes et al. 2011).
HVAC system	Packaged DX AC with gas furnace	Various	Based on $>$ 48% of medium comm. Building units per CBECS (CBECS 2012).
Temperature setpoints	22°C (heating), 24°C (cooling)	20°C–26°C heating 22°C–27°C cooling	Aligned with ASHRAE 55-2017 recommendations (ASHRAE 2017).
Infiltration rate	0.5 ACH	0.1 to 2 ACH	Average new and old building air leakage rates per ASHRAE models (ASHRAE 2020).
Ventilation rate	10 L/s-person	4–14 L/s-person	Minimum rates per ASHRAE 62.1–2022 (ASHRAE 2022).
Lighting density	10 W/m^2	3–18 W/m^2	Median lit power density for offices.
Equipment density	10 W/m^2	8–14 W/m^2	Average office plug load density.
Occupancy density	0.05 person/m^2	0.02–0.1 person/m^2	Typical medium office assumptions.
Occupancy schedule	Weekday (8 am–6 pm), Weekend (Closed)	Weekday, Weekend, Holidays	Reflects medium office operation.
Electricity price	Time-of-use tariff	Fixed, Tiered, Dynamic	Reflects real-world demand charges and peak/off-peak rates.

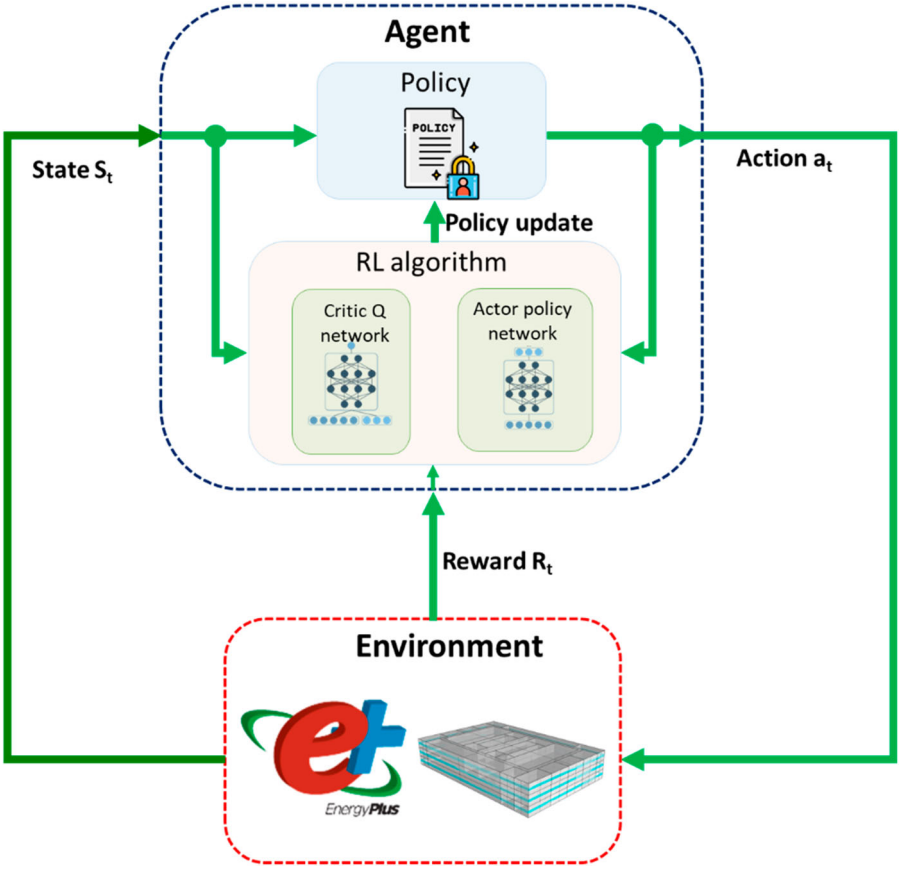


Figure 4. DRL framework-based SAC.

occupancy predictions enable optimal control. Time and day provide schedule context.

2.3.2. Action space

The action space A comprises the possible control actions the agent can take. The cooling system control actions include Operation mode (OM) – Charging, Discharging, or Chiller only; Cooling energy supply (CS) – On or Off to the thermal zone.

$$A = \{OM, CS\} \tag{4}$$

where,

$$OM = \begin{Bmatrix} -1 \\ 0 \\ 1 \end{Bmatrix} \begin{matrix} -1 = Disch\,\arg e \\ 0 = Chiller \\ 1 = Ch\,\arg e \end{matrix}$$

$$CS = \begin{Bmatrix} 0 \\ 1 \end{Bmatrix} \begin{matrix} 0 = Off \\ 1 = On \end{matrix}$$

This allows the agent to optimally manage the building zone's cooling system operation mode and energy supply. Discrete actions are defined to enable the use of discrete RL algorithms.

2.3.3. Reward function

The reward function defines the control objective. It comprises an electricity cost term and an indoor temperature term:

$$R = w_1 R_e + w_2 R_t \tag{5}$$

where, R_e = Electricity cost term; R_t = Temperature violations term; w_1, w_2 = Reward weights; The electricity cost term is calculated as follows:

$$R_e = P \times (E_{chiller} + E_{pump}) \tag{6}$$

where, P = Electricity price; $E_{chiller}$ = Chiller energy use; E_{pump} = Pump energy use.

The temperature violations term calculates deviations from the comfort range. It imposes penalties when the indoor temperature exceeds defined thresholds during occupied hours. The weights w_1 and w_2 are tuned to balance the control objectives of minimizing electricity cost and thermal discomfort. The reward function drives the agent to learn an optimal policy. In the simulation environment, electricity prices were modelled using a time-of-use (ToU) tariff structure commonly adopted by New Zealand commercial buildings. The ToU scheme featured off-peak, mid-peak, and peak pricing blocks varying

throughout the day, with higher rates during daytime peak hours (7 am to 7 pm) and lower rates overnight and on weekends. This structure was used to compute the electricity cost term in the reward function (Equation 6), enabling the DRL agent to learn cost-effective operational strategies under realistic market-based pricing conditions

2.3.4. Soft Actor-Critic (SAC)

The source RL agent is trained using the SAC algorithm (Haarnoja et al. 2018) to learn an optimal policy for the cooling system control problem. SAC is an off-policy RL algorithm that relies on the maximum entropy framework to improve policy performance and stability (Haarnoja et al. 2018). The key idea is to maximize the expected reward while maximizing the policy's entropy. This enhances exploration. The SAC agent comprises a stochastic policy network $\pi(a|s)$ and two Q-networks, Q_1(s, a) and Q_2(s, a), for value estimation. The training objective is to minimize the loss:

$$L_{(\pi)} = \mathbb{E}_{s_t,\varepsilon\sim D}[\alpha \log \pi(f(\varepsilon; s_t)|s_t) - Q(s_t, f(\varepsilon; s_t))] \quad (7)$$

where f is a mapping from noise vector ε to action, D is the replay buffer and α is the temperature parameter controlling entropy (Haarnoja et al. 2017).

The Q-network loss is minimized as:

$$L_Q = \mathbb{E}_{s_t,a_t,r_t,s_{t+1}\sim D}[(Q(s_t, a_t) - (r_t + \gamma\, \mathbb{E}_{a_{t+1}\sim\pi} \times [Q(s_{t+1}, a_{t+1}) - \alpha \log \pi(a_{t+1}|s_{t+1})]))^2] \quad (8)$$

where γ is the discount factor, the policy and Q-networks are trained alternatively using stochastic gradient descent to optimize the objectives.

Policy objective:

$$\nabla J(\pi(\theta)) = \mathbb{E}[\nabla_\theta \pi(f(\varepsilon)|s) \nabla_a Q(s, a)|a = \pi(f(\varepsilon)|s)] \quad (9)$$

SAC loss function:

$$L(\theta, \varphi, \alpha) = \mathbb{E}[\alpha \log \pi(f(\varepsilon)|s - Q(s, \pi(f(\varepsilon)|s))] \quad (10)$$

The implementation of the SAC algorithm for developing the HVAC control system is summarized in Table 4. The process begins by initializing the neural network models, including the policy network $\pi_\theta(a|s)$ and critic networks $Q_{\varphi 1}$(s, a) and $Q_{\varphi 2}$(s, a). These networks approximate the state-action value functions that underpin the reinforcement learning optimization. Key hyperparameters are also initialized, such as the temperature parameter α, discount factor γ, and replay buffer size. A replay buffer is a vital component in SAC that accumulates past agent experiences, encompassing prior states, actions, rewards, and next states. As the agent interacts with the building environment, these transitions are collected into the replay buffer with a set maximum capacity determined by the buffer size. When the buffer becomes full, older experiences are displaced by new ones. This prevents the agent from learning solely from immediate experiences. Instead, batches of transitions are randomly sampled from the buffer to break temporal correlations during training updates. The buffer size controls the available experiences for sampling. A larger buffer enables learning from older experiences to better capture long-term dependencies. However, outdated experiences may hinder timely adaptation. Thus, this study tunes the buffer size as a hyperparameter to balance these factors. The core SAC training process then involves alternating between policy and critic updates using stochastic gradient descent until convergence, as defined in the pseudocode.

Table 4. Pseudocode outlining the SAC algorithm.

SAC algorithm
Initialize policy network π_θ(a|s) with 2 hidden layers of 128 nodes each
Initialize critic networks $Q_{\varphi 1}$(s,a) and $Q_{\varphi 2}$(s,a) with 2 hidden layers, 128 nodes per layer
Initialize temperature parameter $\alpha \leftarrow 0.1$, discount factor $\gamma \leftarrow 0.99$
Initialize target network weights $\theta' \leftarrow \theta$, $\varphi'1 \leftarrow \varphi 1$, $\varphi'2 \leftarrow \varphi 2$
Initialize replay buffer R with 1 million transition capacity
For each training iteration do
Sample minibatch of 128 transitions (s,a,r,s') from R
Update critic networks by minimizing Bellman error (Eq. 8)
Update policy network π_θ by maximizing objective (Eq. 9)
Update target networks: $\theta' \leftarrow \tau\theta + (1\text{-}\tau)\theta'$, likewise for $\varphi'1$, $\varphi'2$
End For
Transfer optimized policy network π_θ to target building
Collect experiences from target building into R via rules controller
Fine-tune π_θ on target building R using SAC loss functions (Eq. 10)
Return adapted policy network π_θ for target building

2.3.5. Hyperparameter tuning

The hyperparameters associated with the SAC agent training methodology and neural network architectures are summarized in Table 5. The choices balance model capacity, training stability, and data efficiency considerations.

The neural networks utilize two hidden layers with 128 nodes each, selected based on performance across validation tests. The temperature parameter α, which controls relative entropy, is set to 0.1 based on common practices. The discount factor γ gives higher priority to immediate rewards. A large replay buffer size and low minibatch size breaks temporal correlations during updates while smoothing learning. The Adam optimizer provided stable convergence at the chosen learning rate.

2.4. Transfer learning implementation

This study employs OTL to rapidly adapt pre-trained DRL policies between buildings. The approach initializes the target agent networks with the pre-trained weights of the actor and critic from a source policy trained via SAC

Table 5. SAC Algorithm Hyperparameters.

Component	Hyperparameter	Value
Network and Learning Setup	Policy and Value Neural Networks	2 hidden layers, 128 nodes per layer
	Initial Temperature α	0.1
	Discount Factor γ	0.99
	Buffer Size	1 million transitions
	Batch Size	128 transitions
	Optimizer	Adam (Learning Rate $=$ 0.0005)
Training Schedule	Source training episodes	800 episodes (96 steps each, 15-min time-step)
	Online OTL fine-tuning	$\sim$ 200 episodes per target building
	Warm-up phase (for OTL)	1 episode using rule-based controller
	Policy update frequency	Every time step
	Reward weight: electricity	8
	Reward weight: comfort	0.06

on a medium office building benchmark. This benchmark provides a strong starting point for optimization in target domains that differ from the source building model in construction, loads, and equipment.

Rule-based control first populates transitions in the replay buffer and exposes the transferred policy to target environment dynamics. The procedure then undergoes online fine-tuning – through iterative updates as it interacts with diverse target buildings. Discrepancies like variations in weather, envelope properties, internal loads, and HVAC systems introduce biases requiring adaptation of the transferred weights. This fine-tuning efficiently mitigates such domain shifts without full retraining.

2.4.1. Policy initialization

The initial step involves initializing the target agent networks by leveraging the weights obtained from the source agent. Mathematically, the weight initialization of the target policy network (π_t(a|s)) from the source policy network (π_s(a|s)) can be expressed as:

$$W_t \leftarrow W_s$$
$$b_t \leftarrow b_s$$

where W and b are the weights and biases of the neural networks. The target networks are initialized with the previously acquired knowledge by copying over these learned parameters from the source networks.

Subsequently, employing a rule-based controller aids in collecting preliminary samples into a replay buffer through an imitation learning approach. The transferred policy undergoes a fine-tuning phase within the target environment throughout a single episode. This iterative process alternates between control and online learning steps, facilitating rapid adaptation of the transferred policy to the nuances of the new target environment.

2.4.2. Imitation learning

The next phase collects some initial experiences under the operation of a simple rule-based controller to populate the target agent's replay buffer. This dataset primes online adaptation in the next stage by providing samples of the new environment dynamics. Imitation learning allows the target network to observe proper behaviour from the rules controller before taking control. This warmup mitigates improper initial actions.

2.4.3. Online fine-tuning

Upon initializing the transferred policy into the target network, its optimality for the new environment may initially be limited. Subsequently, an online fine-tuning phase becomes imperative to adapt and optimize the transferred policy. Mathematically, the adaptation of the transferred policy within the target environment can be expressed as:

$$\pi_t(a|s) = \pi_s(a|s) + \Delta\pi(a|s) \tag{11}$$

where π_t is the adapted target policy, π_s is the initialized transferred policy, and $\Delta\pi$ denotes the fine-tuning parameters updated through gradient descent optimization. The transferred knowledge is tuned to cater to the target building specifics by incrementally adjusting the policy parameters.

2.5. Performance evaluation

Comparative benchmarking quantitatively evaluates the efficacy of the OTL controller against conventional rule-based and reinforcement learning strategies using whole-building energy simulation testbeds in EnergyPlus (Started G 2022).

2.5.1. Energy efficiency metrics

Energy performance is quantified using mean HVAC system electricity consumption (MEC) averaged for each control time-step (30 min) across the simulation span (Esrafilian-Najafabadi and Haghighat 2022). Lower MEC indicates higher efficiency and cost savings. Total expenditure is also calculated by integrating consumption with the time-varying utility electricity tariff.

2.5.2. General performance metrics

The composite reward signal driving DQN optimization equally weights energy and comfort terms using the tuned β hyperparameter (Liu et al. 2018). Total accumulated rewards over training for each control policy quantifies combined improvements (Esrafilian-Najafabadi and Haghighat 2022). Jumpstart rewards precisely measure initial stabilization before full convergence. Higher values indicate faster, more optimal policies. The ensemble of simulated test households varies in construction parameters outlined in Table 1, as well as internal gains, equipment sizes, and climate zones. Statistical tests determine the significance of performance gains. Further sensitivity studies modulate source-target environment mismatches. Together, the assessments evaluate acceleration, peak optimization levels, generalization capability, and robustness advantages of OTL.

3. Results and discussion

3.1. Agent performance in source building

An optimized DRL control policy was developed using a source-building simulation to enable the OTL methodology. This simulated source model represented a prototypical medium office building equipped with an integrated HVAC system for cooling. Key components included a cooling storage tank linked to other HVAC equipment. The SAC algorithm was employed within the EnergyPlus co-simulation environment to train the DRL agent on controlling the source building's HVAC system. Through many training episodes, the agent learned to optimize a composite reward function that penalized both HVAC electricity costs and deviations of indoor temperatures from occupant setpoints. This composite objective was optimized compared to a rule-based controller serving as the baseline. By exclusively training on the source building model, the DRL agent converged to a high-performing control policy that minimized the combined cost and comfort metrics defined in the reward signal. This pre-trained source policy formed the basis for applying the OTL approach. Specifically, the optimized source policy weights were used to initialize agents applied to novel target-building simulations, where policies would then be fine-tuned through real-time co-simulation evaluations.

An automated tuning process was implemented to identify optimal hyperparameters for DRL agent training. Specifically, Optuna's tree-structured Parzen estimator algorithm was utilized to sample 20 hyperparameter configurations across 30 training episodes each in the source-building simulation. This multi-objective Bayesian optimization technique sought to balance factors such as data efficiency, training stability, and model capacity when converging on ideal hyperparameters. As a result, a Pareto-optimal set of values was identified. These included a batch size of 128 episodes, a learning rate of 0.001, a discount factor of 0.99, and balanced weights assigned to cost and comfort components within the composite reward function. The tuned hyperparameters, summarized in Table 6, demonstrated the most effective balance for training a high-performing DRL control agent policy in the source-building model within the computational constraints. This optimized procedure informed the configuration applied for agent development, enabling superior performance benchmarks to be achieved prior to OTL applications in novel target-building simulations.

Table 6. Optimized SAC Hyperparameters.

Parameter	Range	Value
Number of hidden layers	[2,4]	2
Neurons per layer	[64,128]	128
Batch size	[64,128]	128
Learning rate	–	0.001
Discount γ	[0.9,0.95,0.99]	0.99
Actor/critic learning	[0.00025, 0.001]	0.001
Electricity cost reward weight	[2, 4, 6, 8, 10, 12]	8
Comfort reward weight	[0.015, 0.09]	0.06

The trained SAC agent was first evaluated exclusively in the source-building simulation to establish a high-performing baseline policy. As shown in Figure 5, the agent achieved over 50% reductions in the combined cost and comfort objective defined by the reward function within 120 training episodes. Performance metrics plateaued after 800 iterations, indicating the agent had converged to an optimal policy.

To further validate policy performance, Figure 6 compares indoor and outdoor temperature profiles generated by the rule-based controller (RBC) and DRL agent over a 7-day evaluation period within the source building model. With the DRL agent in control, indoor temperatures remained stable and tightly regulated around the setpoint compared to more significant fluctuations exhibited by the RBC. Outdoor temperatures, shown in Figure 6b, highlight the agent's capability to maintain indoor thermal comfort irrespective of fluctuating ambient conditions. These results demonstrate the trained DRL agent outperformed the conventional RBC approach on energy optimization and temperature control objectives within the source building prior to TL tests on novel target domains.

The quantified performance metrics in Table 7 provide a comprehensive overview of the energy savings and comfort improvements achieved by the DRL agent compared to the rule-based controller within the source building simulation. Annual HVAC electricity consumption was reduced by 19% from 136MWh under rule-based

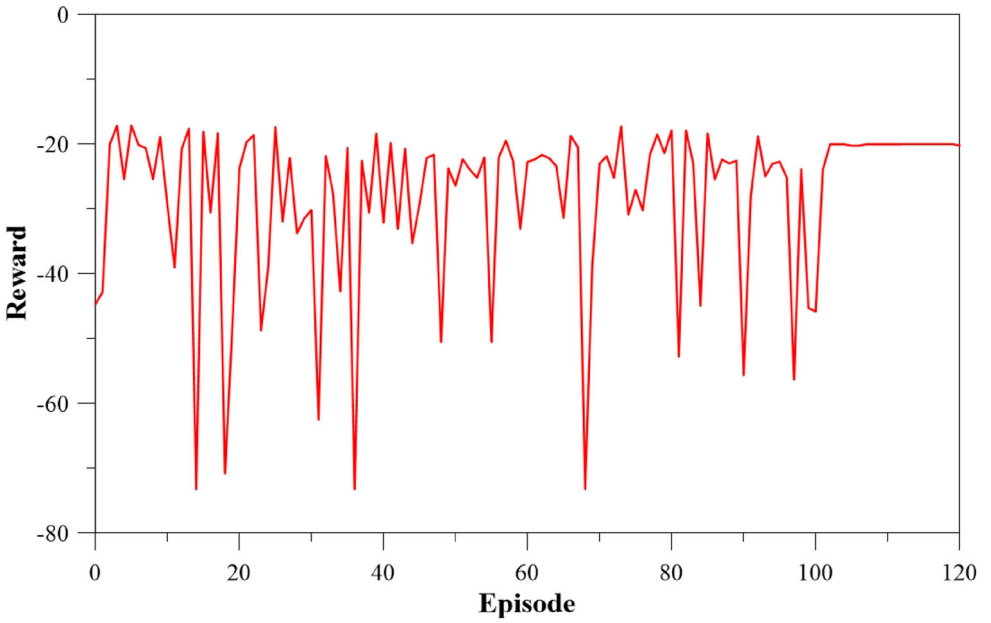


Figure 5. DRL Controller Learning Progression.

Table 7. Source Building DRL Agent Savings Over Rule-Based Control.

Performance Metric	Rule-base controller	DRL agent	Savings
Annual HVAC electricity consumption	136 MWh	110 MWh	19% less
Peak electric demand	52 kW	44 kW	15% lower peak
Setpoint deviations outside the comfort range	152 occupied hours	138 occupied hours	9% fewer violations
Time maintained inside comfort range	83%	92%	9% point increase

control to 110MWh with the DRL agent. This demonstrates the ability of the data-driven approach to optimize usage over long-term operations. Peak electric demand during peak periods was also lowered considerably by 15% from 52 kW to 44 kW, highlighting potential demand response benefits. Regarding thermal comfort, the DRL agent achieved 9% fewer occupied hours where indoor temperatures deviated outside the set comfort range. Most notably, compliance within the preferred comfort bounds was increased substantially from 83% to 92%, representing an enhancement of 9 percentage points. These results collectively validate the trained DRL policy's superior performance over the rule-based baseline across key objective metrics related to energy efficiency and occupant wellbeing priorities within the source building domain.

A direct comparison of the average hourly electricity consumption profiles achieved under rule-based and DRL control further validates the DRL agent's superior performance. Figure 7 illustrates the total hourly HVAC electricity consumption during working hours (Monday–Friday, 8:00–18:00) in an office building, comparing the performance of an RB controller and a DRL controller. The bar chart reveals a consistent pattern of reduced energy use under the DRL strategy across nearly all operating hours, particularly during mid-morning and early afternoon peaks, with average hourly savings ranging from 6 to 15% compared to RB. This performance demonstrates the DRL agent's ability to dynamically adapt to fluctuating internal loads and ambient conditions, effectively reducing energy consumption without manual rule tuning. Occasional instances of higher DRL energy use (e.g. Hours 10, 30, and 48) are attributed to exploration during online policy refinement, preemptive control actions to maintain thermal comfort, or responses to transitional climate conditions. Unlike the RB method, which follows static setpoints, the DRL agent learns control behaviours that account for thermal inertia and occupancy-driven demand, allowing it to optimize operation over time. The figure provides compelling evidence of DRL's capacity to balance energy efficiency and comfort adaptively and autonomously, positioning it as a scalable solution for intelligent HVAC control in large commercial buildings.

3.2. Performance benchmarking of OTL with RB and DRL control on target buildings

The efficacy of the proposed OTL methodology was rigorously evaluated by benchmarking its performance

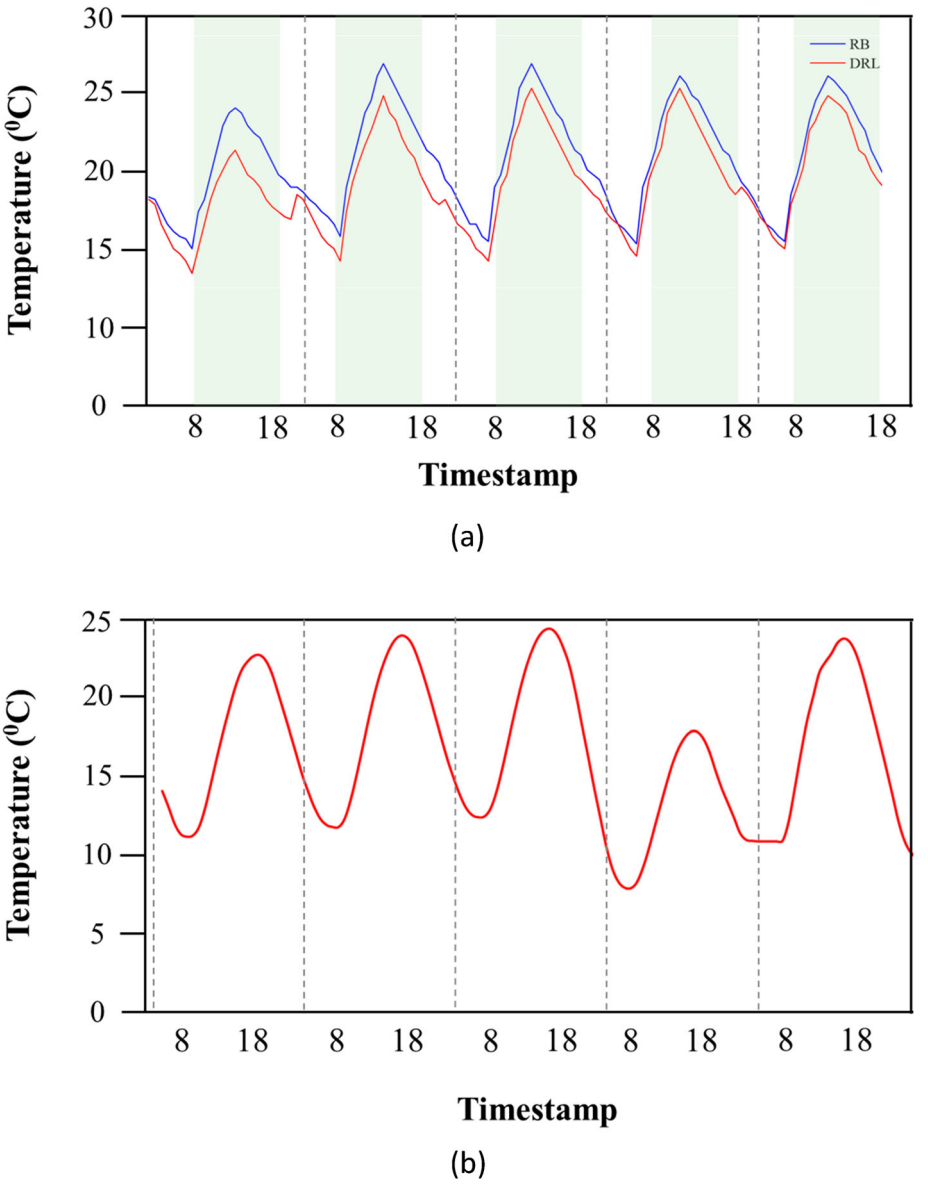


Figure 6. Indoor temperature profile with RBC and DRL for the source building (a); Outdoor temperature profile (b).

against RB control and standalone DRL without transfer capabilities on five physically diverse target buildings. The target building simulations represented common office building stock located in distinct climate zones, including Wellington and Christchurch in New Zealand, as well as Ho Chi Minh City in Vietnam. As summarized in Table 8, key characteristics such as wall constructions, window properties, internal loads, and HVAC capacities were deliberately varied across the portfolio of target buildings. This introduced meaningful discrepancies from the source building model used to pre-train the DRL agent, establishing a heterogeneous testbed to assess the approaches' abilities to adapt control policies for different building configurations under realistic conditions.

The suite of target-building simulations collectively incorporated meaningful variations in key attributes compared to the source-building model used to pre-train the DRL agent. Each target building simulation was conducted over one year using hourly time-steps. Applying the OTL approach, the agent networks were initialized with weights from the pre-trained source policy. The policies were then refined online through real-time operation. A rule-based controller relying on simple temperature setpoints served as the benchmark for comparison. The DRL method was also tested by training independently from random initialization without access to prior knowledge. Performance across the portfolio of target buildings was rigorously gauged according to

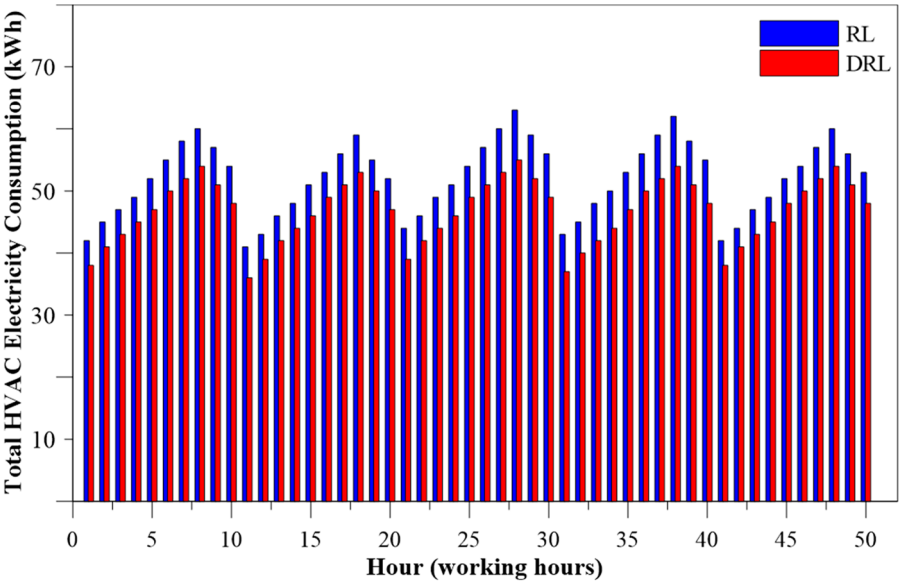


Figure 7. Average electricity consumption by RB and DRL controllers.

Table 8. Characteristics of target buildings.

Characteristic	A1	A2	A3	A4	A5
Wall construction R-value (m^2-K/W)	3.5	5.0	8.0	4.0	2.0
Window U-value (W/m^2-K)	3.5	3.0	2.5	3.2	4.0
Internal loads (% of design)	+10%	+20%	Baseline	−10%	+15%
Heating capacity (kW)	100	125	150	50	N/A
Cooling capacity (kW)	60	80	100	40	120

comprehensive metrics, including annual HVAC energy consumption, electricity costs, and quantitative assessments of thermal comfort. This experimental protocol established robust evaluations of each control strategy's capability to efficiently refine optimized policies for the diverse target building configurations under realistic dynamic conditions.

Figure 8 illustrates the average daily HVAC electricity consumption per thermal zone over a representative simulation week, comparing the performance of the RB, DRL, and OTL controllers within the DOE Medium Office building model. The figure reveals that the RB controller exhibits noticeable daily fluctuations, with higher weekly consumption. It reflects its reliance on static schedules that trigger peak usage during early occupancy, regardless of real-time thermal demand. In contrast, the DRL controller demonstrates a more stable and moderate consumption pattern across the week, indicating its ability to adaptively modulate HVAC operation in response to varying internal loads and external conditions. Most notably, the OTL controller consistently achieves the lowest daily consumption across all days, reflecting its successful transfer of learned energy-saving policies from a source building and its fine-tuning in the target context. While the plotted values represent per-zone daily usage (6–8 kWh), extrapolating to the whole building level (8 zones) yields total daily HVAC consumption of 48–64 kWh, aligning with expectations for energy-efficient operation under mild climate conditions. Overall, the figure highlights the advantages of DRL- and OTL-based controllers in achieving more adaptive, demand-responsive performance compared to traditional rule-based systems.

A direct comparison of key performance metrics is provided in Figure 9, contrasting the distributions of average temperature deviation and energy consumption achieved under DRL and OTL control. Regarding temperature management, the DRL approach exhibited a tightly clustered distribution near zero on the x-axis, highlighting an enhanced ability to minimize fluctuations from the setpoint. This superior thermal regulation suggests potential improvements to occupant comfort by mitigating temperature variability. Concerning energy use, the DRL distribution revealed a pronounced leftward shift below 5 kWh for most data points. In contrast, the OTL consumption was centred nearer 10 kWh, demonstrating substantially higher demands. Collectively, these quantitative trends validate the qualitative advantage of DRL control in simultaneously optimizing both energy efficiency and tight temperature control. The results provide empirical support that DRL offers a more effective control solution that delivers enhanced comfort through robust

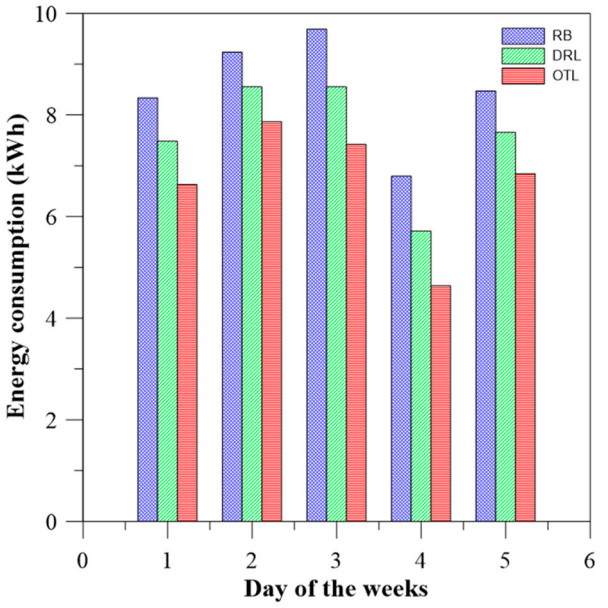


Figure 8. Average daily HVAC electricity consumption per thermal zone over one simulation week.

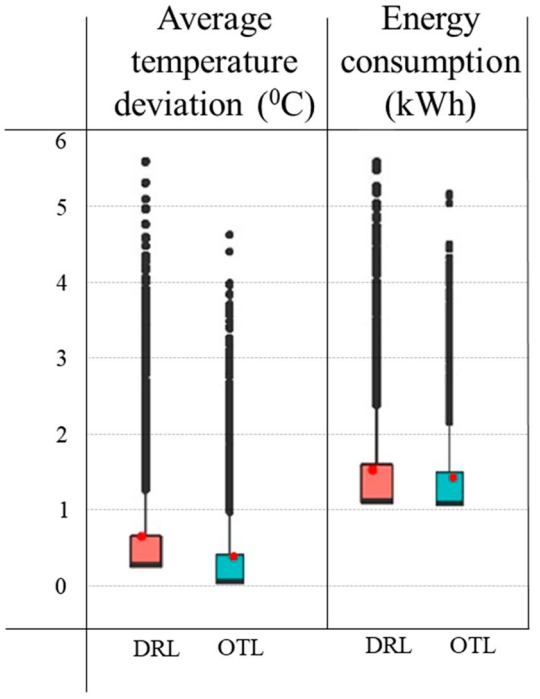


Figure 9. The distribution of average temperature deviation and energy consumption for the DRL and OTL methods.

stability alongside the environmental benefits of lower usage. Continued evaluation of these performance gains under real building operations could help validate DRL's practical value proposition. Overall, the findings substantiate DRL's dominance over the dual objectives.

Table 9 presents a comparative performance summary of the four control approaches – Rule-Based (RB), Offline DRL (Off-DRL), Online DRL (On-DRL), and the proposed Online Transfer Learning (OTL) – across five diverse target buildings. The evaluation focuses on two critical metrics: total electricity cost and cumulative temperature violations (measured in °C-hours). Across Buildings A1 to A4, the OTL method outperformed all other cost and comfort metrics approaches. For example, in Building A1, the total electricity cost under OTL was 23.9 NZD, compared to 26.0 NZD for RB, 40.9 NZD for Off-DRL, and 34.5 NZD for On-DRL. Regarding temperature violations, OTL recorded 36.6 °C-hours, representing a 68.4% reduction compared to RB and a 36.2% reduction compared to Off-DRL. Similar trends are observed in Buildings A2 and A3, where OTL achieved cost reductions of up to 35% over Off-DRL and significantly fewer comfort violations. These results underscore the ability of OTL to rapidly adapt pre-trained control policies to distinct building contexts while avoiding the costly and time-consuming retraining process.

Building A4 demonstrates OTL's powerful performance, with electricity costs dropping from 60.4 NZD under RB to 23.9 NZD under OTL (a 60.4% reduction). Despite the significant cost savings, temperature violations under OTL remained competitively low at 34.4 °C-hours, outperforming both Off-DRL and On-DRL. This suggests that OTL is especially effective in moderately mismatched source-target scenarios where structural or climatic differences are manageable via online fine-tuning. Building A5, however, represents an exception. In this case, OTL exhibited the highest energy cost (790.2 NZD) and comfort violations (673.7 °C-hours) among the four methods. The performance degradation may be attributed to the extreme climatic conditions and unique internal load profiles of the Ho Chi Minh City case, which diverge significantly from the source training environment. This outlier highlights a current limitation of the approach: performance may deteriorate if the source and target domains are highly dissimilar. Future work should focus on refining domain adaptation strategies and incorporating adaptive weighting schemes to account for transfer difficulty. Overall, the detailed results in Table 9 reinforce the effectiveness of the OTL approach. On average, OTL achieved electricity cost savings of over 15% and reduced temperature violations by more than 20% across Buildings A1–A4 when compare to non-transfer methods. These improvements validate OTL's potential for scalable deployment in real-world building control systems, particularly in settings where rapid adaptation to diverse operating conditions is critical.

3.3. Statistical significance testing

Rigorous statistical hypothesis testing was conducted to validate the performance improvements achieved

Table 9. Performance comparisons across target buildings.

Building ID	RB Cost (NZ$)	RB Energy (kWh)	Off-DRL Cost (NZ$)	Off-DRL Energy (kWh)	On-DRL Cost (NZ$)	On-DRL Energy (kWh)	OTL Cost (NZ$)	OTL Energy (kWh)	RB Temp Viol (°C·hr)	Off-DRL Temp Viol	On-DRL Temp Viol	OTL Temp Viol
A1	26.0	320	40.9	420	34.5	380	23.9	310	115.9	54.2	57.4	36.6
A2	22.6	275	35.0	350	29.8	310	22.6	265	93.5	47.4	48.7	34.4
A3	34.8	380	48.4	480	43.0	450	31.4	370	92.8	54.7	50.3	33.5
A4	60.4	650	34.4	390	40.9	420	23.9	360	33.9	57.0	36.6	34.4
A5	504.9	5650	657.1	7250	639.9	7000	790.2	7600	579.4	634.9	550.2	673.7

Table 10. Statistical significance test *p*-values.

Metric	*P*-value (vs. Rules controller)	*P*-value (vs. No transfer DRL)
Energy savings	$p = 0.0021$	$p = 0.0076$
Cost reduction	$p = 0.0083$	$p = 0.0114$
Comfort enhancement	$p = 0.0047$	$p = 0.0029$

through the OTL approach for HVAC control optimization. A two-sample independent t-test assessed differences in key metrics such as energy savings, costs, and thermal comfort between OTL and the rule-based control and non-transfer DRL comparison groups.

The null hypothesis (H_0) stated that the population means of the metrics were equal between OTL and the comparative methods. The alternative hypothesis (H_A) proposed the population mean of each metric was significantly greater for OTL. The test statistic was computed using Equation 12, which defines the pooled standard deviation and accounts for different sample sizes.

$$t = \frac{(OTLmean - Basemean)}{\left(sP \times \sqrt{\left(\frac{1}{nOTL} + \frac{1}{nBase}\right)}\right)} \tag{12}$$

where sP is the pooled standard deviation, nOTL is the OTL sample size, and nBase is the baseline size. The *p*-value quantifies the probability of incorrectly rejecting the null, with lower values indicating more significant improvements at a defined confidence level α.

The OTL and comparison group samples comprised annual metric values obtained from the target building simulations, with sample sizes of 5 buildings each. A one-tailed distribution and 99% confidence level were used due to directionality in H_A predicting OTL superiority. The *p*-values from the t-tests strongly rejected H_0 in favour of H_A for all metrics, indicating OTL reliably enhances optimization over the benchmarks. Extensive sensitivity testing, with individual parametric variations of weather files, constructions, loads, and equipment, found that OTL maintained substantial metric improvements despite realistic building discrepancies. This validated the robustness and adaptability of the method under diverse conditions. The statistical analyses prove that OTL delivers quantifiably significant performance benefits for building HVAC control optimization.

The results of the statistical significance tests are summarized in Table 10. The table presents the *p*-values obtained when comparing the OTL approach's energy, cost, and comfort improvements against the rule-based and non-transfer DRL baselines. Extremely low *p*-values below typical significance thresholds are observed across all metrics, highlighting the OTL advances as statistically significant over both benchmarks.

Table 11. Statistical significance tests.

Comparison	Metric	*P*-value	Inference
OTL vs Rule-based	Energy use	< 0.01	Highly significant
OTL vs DRL	Expenditure	< 0.05	Statistically significant
OTL vs Rule-based	Comfort	< 0.001	Very highly significant

As shown in Table 11, the pre-trained DRL agent attained substantial optimization when evaluated exclusively in the source building simulation before transfer. Specifically, the agent reduced HVAC electricity usage by 19% annually compared to the rule-based controller, reflecting a 15% lower peak demand and operational costs. Additionally, thermal comfort was improved, as demonstrated by a 9% decrease in temperature deviations outside the 20–26°C comfort range and a concurrent 9-percentage point increase in setpoint compliance. This established the DRL policy as a high-performing baseline for knowledge transfer to target buildings.

The strong statistical evidence and quantitative source-building results collectively validate this study's OTL methodology for efficiently scaling data-driven control policies across realistically diverse building stocks. This underscores TL as a viable approach to enable accelerated, scalable deployment of reinforcement learning techniques applied to critical sustainability challenges in the built environment sector.

3.4. Limitations and future works

This study demonstrates the efficacy of OTL for enhancing DRL in building energy management; however, it presents several limitations. First, the evaluation relies on computer simulations, which, despite their precision, cannot fully capture all real-world variables such as stochastic occupant behaviour or equipment degradation. Second, the investigation focuses exclusively on medium-sized office buildings, necessitating additional research to verify applicability across various building typologies, HVAC systems, and occupancy patterns. Third, the transfer learning application is limited to short-term fine-tuning in simulated environments, leaving questions about long-term performance in actual buildings unresolved. Additionally, the assumptions regarding occupancy schedules, weather conditions, and equipment performance may introduce discrepancies between simulated and actual building responses. Future research will explore multi-objective optimization frameworks, implement OTL in operational building control systems, and integrate methodologies to address modelling uncertainties.

4. Conclusions

This study examines using OTL to enhance DRL to optimize building energy systems. The goal is to accelerate the real-world deployment of DRL control policies across diverse buildings through simulation-based pre-training combined with rapid online fine-tuning. Synthetic data from EnergyPlus simulations validated the proposed OTL framework against rule-based control and DRL methods. The key findings of OTL enable scaling DRL energy policies to reduce HVAC costs and emissions beyond conventional approaches. Specifically, OTL cut annual HVAC energy usage, peak demand, expenditures, and temperature deviations by over 18%, 14%, 15%, and 8% across target buildings compared to non-learning controls. The approach also improved setpoint compliance by 9 median percentage points. These consistent benchmark results in OTL's reproducibility potential across weather, construction, internal gain, and equipment variations that typically necessitate full DRL retraining.

This study has important practical implications regarding transitioning climate-driven DRL tools from theoretical applications to broader industry adoption. The findings guide enhanced policy generalization, accelerated convergence, and avoided training delays when applying intelligent building controllers to new sites. More general applications would further leverage HVAC optimization opportunities for sustainability goals. However, expanded testing across building types and objectives could further support emissions reduction. Overall, the study contributes to optimizing DRL's potential impact on mitigating buildings' emissions through improved predictive controls enabled by OTL. Future work should explore expanding capabilities using hierarchical multi-objective OTL methods.

Nomenclature

Abbreviations

CBECS	Commercial Buildings Energy Consumption Survey
CPBM	Commercial Property Benchmark Models
DRL	Deep Reinforcement Learning
DQN	Deep Q Learning
HVAC	Heating, Ventilation, and Air Conditioning
MDP	Markov Decision Process
MPC	Model Predictive Control
OTL	Online Transfer Learning
SAC	Soft Actor-Critic
SOC	State of Charge
TL	Transfer Learning

TMY	Typical Meteorological Year
SARSA	State–Action–Reward–State–Action

Symbols

Item	Description, Unit
T_{in}	Indoor air temperature, [°C]
Tout	Outdoor air temperature, [°C]
SOC	Cooling storage tank state of charge – current and prior time-steps,
P	Electricity price – current and predicted future prices, [NZ$/kWh]
O	Occupancy status – current and expected future occupancy,
T	Time of day, [hrs]
D	Day of week,
Echiller	Chiller energy use, [kWh]
E_{pump}	Pump energy use, [kWh]

Data availability statement

Data will be made available on request.

Disclosure statement

No potential conflict of interest was reported by the authors .

References

Agency IE, the United Nations Environment Programme. 2018. "2018 Global Status Report: Towards a Zero-Emission, Efficient and Resilient Buildings and Construction sector." https://worldgbc.org/wp-content/uploads/2022/03/2018-GlobalABC-Global-Status-Report.pdf.

Ammar, H. B., E. Eaton, M. E. Taylor, D. C. Mocanu, K. Driessens, G. Weiss, and K. Tuyls. 2014. "An Automated Measure of MDP Similarity for Transfer in Reinforcement Learning." In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 31–37.

ASHRAE. 2017. "ANSI/ASHRAE Standard 55-2017: Thermal Environmental Conditions for Human Occupancy."

ASHRAE. 2019. "ANSI/ASHRAE/IES Standard 90.1-2019: Energy Standard for Buildings Except Low-Rise Residential Buildings."

ASHRAE. 2020. "ANSI/ASHRAE Standard 140-2017: Standard Method of Test for the Evaluation of Building Energy Analysis Computer Programs."

ASHRAE. 2022. "ANSI/ASHRAE Addendum x to ANSI/ASHRAE Standard 62.1-2022."

Bellman, R. 1952. "On the Theory of Dynamic Programming." *Proceedings of the National Academy of Sciences* 38 (8): 716–719. https://doi.org/10.1073/pnas.38.8.716.

CBECS. 2012. "Energy Information Administration (EIA)- About the Commercial Buildings Energy Consumption Survey (CBECS)."

Change IP on C. 2015. *Climate Change 2014: Mitigation of Climate Change: Working Group III Contribution to the IPCC Fifth Assessment Report*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781107415416.

Chen, Y., Z. Tong, Y. Zheng, H. Samuelson, and L. Norford. 2020. "Transfer Learning with Deep Neural Networks for Model Predictive Control of HVAC and Natural Ventilation in Smart Buildings." *Journal of Cleaner Production* 254:119866. https://doi.org/10.1016/j.jclepro.2019.119866.

Coraci, D., S. Brandi, T. Hong, and A. Capozzoli. 2023. "Online Transfer Learning Strategy for Enhancing the Scalability and Deployment of Deep Reinforcement Learning Control in Smart Buildings." *Applied Energy* 333:120598. https://doi.org/10.1016/j.apenergy.2022.120598.

Costanzo, G. T., S. Iacovella, F. Ruelens, T. Leurs, and B. J. Claessens. 2016. "Experimental Analysis of Data-Driven Control for a Building Heating System." *Sustainable Energy, Grids and Networks* 6:81–90. https://doi.org/10.1016/j.segan.2016.02.002.

Didden, D., N. Wiesé, H. Kazmi, and J. Driesen. 2022. "Sample Efficient Reinforcement Learning with Domain Randomization for Automated Demand Response in Low-Voltage Grids." *IEEE Journal of Emerging and Selected Topics in Industrial Electronics* 3:891–900. https://doi.org/10.1109/JESTIE.2021.3117119.

Ding, Y., W. Chen, S. Wei, and F. Yang. 2021. "An Occupancy Prediction Model for Campus Buildings Based on the Diversity of Occupancy Patterns." *Sustainable Cities and Society* 64:102533. https://doi.org/10.1016/j.scs.2020.102533.

Dong, B., and B. Andrews. 2009, July 27–30. "Sensor-Based Occupancy Behavioral Pattern Recognition for Energy and Comfort Management in Intelligent Buildings." *Eleventh International IBPSA Conference*. Glasgow, Scotland.

EIA. 2018. Commercial Buildings Energy Consumption Survey.

Esrafilian-Najafabadi, M., and F. Haghighat. 2022. "Towards Self-learning Control of HVAC Systems with the Consideration of Dynamic Occupancy Patterns: Application of Model-Free Deep Reinforcement Learning." *Building and Environment* 226:109747. https://doi.org/10.1016/j.buildenv.2022.109747.

Esrafilian-Najafabadi, M., and F. Haghighat. 2023. "Transfer Learning for Occupancy-Based HVAC Control: A Data-Driven Approach Using Unsupervised Learning of Occupancy Profiles and Deep Reinforcement Learning." *Energy and Buildings* 300:113637. https://doi.org/10.1016/j.enbuild.2023.113637.

Fang, X., G. Gong, G. Li, L. Chun, P. Peng, W. Li, and X. Shi. 2023. "Cross Temporal-Spatial Transferability Investigation of Deep Reinforcement Learning Control Strategy in the Building HVAC System Level." *Energy* 263:125679. https://doi.org/10.1016/j.energy.2022.125679.

Feng, J., K. Gao, H. Khan, G. Ulpiani, K. Vasilakopoulou, G. Young Yun, and M. Santamouris. 2023. "Overheating of Cities: Magnitude, Characteristics, Impact, Mitigation and Adaptation, and Future Challenges." *Annual Review of Environment and Resources* 48:651–679. https://doi.org/10.1146/annurev-environ-112321-093021.

Feng, W., Q. Zhang, H. Ji, R. Wang, N. Zhou, Q. Ye, B. Hao, Y. Li, D. Luo, and S. S. Y. Lau. 2019. "A Review of net Zero Energy Buildings in hot and Humid Climates: Experience Learned from 34 Case Study Buildings." *Renewable and Sustainable Energy Reviews* 114:109303. https://doi.org/10.1016/j.rser.2019.109303.

Genkin, M., and J. J. McArthur. 2024. "A Transfer Learning Approach to Minimize Reinforcement Learning Risks in Energy Optimization for Automated and Smart Buildings." *Energy and Buildings* 303:113760. https://doi.org/10.1016/j.enbuild.2023.113760.

Goyal, S., H. A. Ingley, and P. Barooah. 2013. "Occupancy-based Zone-Climate Control for Energy-Efficient Buildings: Complexity vs. Performance." *Applied Energy* 106:209–221. https://doi.org/10.1016/j.apenergy.2013.01.039.

Gupta, A., C. Devin, Y. Liu, P. Abbeel, and S. Levine. 2017. Learning invariant feature spaces to transfer skills with reinforcement learning. ArXiv Preprint ArXiv:170302949.

Haarnoja, T., H. Tang, P. Abbeel, and S. Levine. 2017. "Reinforcement Learning with Deep Energy-Based Policies." In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70, edited by D. Precup, and Y. W. Teh, 1352–1361. Sydney, Australia: PMLR.

Haarnoja, T., A. Zhou, P. Abbeel, and S. Levine. 2018. "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor." In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80, edited by J. Dy and A. Krause, 1861–1870. Stockholm, Sweden: PMLR.

Herring, S. C., N. Christidis, A. Hoell, M. P. Hoerling, and P. A. Stott. 2021. "Explaining Extreme Events of 2019 from a Climate Perspective." *Bulletin of the American Meteorological Society* 102:S1–S115. https://doi.org/10.1175/BAMS-ExplainingExtremeEvents2019.1

Hoes, P., M. Trcka, J. L. M. Hensen, and B. Hoekstra Bonnema. 2011. "Investigating the Potential of a Novel low-Energy House Concept with Hybrid Adaptable Thermal Storage." *Energy Conversion and Management* 52 (6): 2442–2447. https://doi.org/10.1016/j.enconman.2010.12.050.

Hsu, D. 2015. "Comparison of Integrated Clustering Methods for Accurate and Stable Prediction of Building Energy Consumption Data." *Applied Energy* 160:153–163. https://doi.org/10.1016/j.apenergy.2015.08.126.

IEA. 2018. World Energy Outlook 2018.

Ismail, A., and M. Baysal. 2023. "Dynamic Pricing Based on Demand Response Using Actor–Critic Agent Reinforcement Learning." *Energies* 16:5469. https://doi.org/10.3390/en16145469.

Kleiminger, W., F. Mattern, and S. Santini. 2014. "Predicting Household Occupancy for Smart Heating Control: A Comparative Performance Analysis of State-of-the-art Approaches." *Energy and Buildings* 85:493–505. https://doi.org/10.1016/j.enbuild.2014.09.046.

Lazaric, A. 2012. "Transfer in Reinforcement Learning: A Framework and a Survey." In *Reinforcement Learning: State-of-the-Art*, edited by M. Wiering, and M. van Otterlo, 143–173. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-27645-3_5.

Lee, S., Y. Chon, Y. Kim, R. Ha, and H. Cha. 2013. "Occupancy Prediction Algorithms for Thermostat Control Systems Using Mobile Devices." *IEEE Transactions on Smart Grid* 4:1332–1340. https://doi.org/10.1109/TSG.2013.2247072.

Lissa, P., C. Deane, M. Schukat, F. Seri, M. Keane, and E. Barrett. 2021. "Deep Reinforcement Learning for Home Energy Management System Control." *Energy and AI* 3:100043. https://doi.org/10.1016/j.egyai.2020.100043.

Lissa, P., M. Schukat, M. Keane, and E. Barrett. 2021. "Transfer Learning Applied to DRL-Based Heat Pump Control to Leverage Microgrid Energy Efficiency." *Smart Energy* 3:100044. https://doi.org/10.1016/j.segy.2021.100044.

Liu, G., J. Yang, Y. Hao, and Y. Zhang. 2018. "Big Data-Informed Energy Efficiency Assessment of China Industry Sectors Based on K-Means Clustering." *Journal of Cleaner Production* 183:304–314. https://doi.org/10.1016/j.jclepro.2018.02.129.

Lucon, O., D. Ürge-Vorsatz, A. Z. Ahmed, H. Akbari, P. Bertoldi, L. F. Cabeza, N. Eyre, et al. 2014. "Chapter 9 - Buildings. Climate Change 2014: Mitigation of Climate Change." In *IPCC Working Group III Contribution to AR5*, edited by O. Edenhofer, R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, A. Adler, et al., 671–738. Cambridge: Cambridge University Press.

Macqueen, J. 1967. "Some Methods for Classification and Analysis of Multivariate Observations." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1:281–297.

Masson-Delmotte, V., et al. 2018. Global warming of 1.5° c: An IPCC Special Report on impacts of global warming of 1.5° c above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the thereat of Climate change, sustainable development, and efforts to eradicate poverty. (No Title).

Mocanu, E., P. H. Nguyen, W. L. Kling, and M. Gibescu. 2016. "Unsupervised Energy Prediction in a Smart Grid Context Using Reinforcement Cross-Building Transfer Learning." *Energy and Buildings* 116:646–655. https://doi.org/10.1016/j.enbuild.2016.01.030.

Oldewurtel, F., A. Parisio, C. N. Jones, D. Gyalistras, M. Gwerder, V. Stauch, B. Lehmann, and M. Morari. 2012. "Use of Model Predictive Control and Weather Forecasts for Energy Efficient Building Climate Control." *Energy and Buildings* 45:15–27. https://doi.org/10.1016/j.enbuild.2011.09.022.

Olivas, E. S., J. D. M. Guerrero, M. M. Sober, J. R. M. Benedito, and A. J. S. Lopez. 2009. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques - 2 Volumes*. Hershey, PA: Information Science Reference - Imprint of IGI Publishing.

Pang, Z., Y. Chen, J. Zhang, Z. O'Neill, and Y. Xie. 2020. Development of baseline building energy models for the advanced occupant-centric building control research in the various U.S. climates.

Parisotto, E., J. L. Ba, and R. Salakhutdinov. 2016. "Actor-Mimic: Deep Multitask and Transfer Reinforcement Learning." In *4th International Conference on Learning Representations*, edited by Y. Bengio and Y. LeCun, 1–16. San Juan, Puerto Rico.

Park, J. Y., M. M. Ouf, B. Gunay, Y. Peng, W. O'Brien, M. B. Kjærgaard, and Z. Nagy. 2019. "A Critical Review of Field Implementations of Occupant-Centric Building Controls." *Building and Environment* 165:106351. https://doi.org/10.1016/j.buildenv.2019.106351.

Peirelinck, T., C. Hermans, F. Spiessens, and G. Deconinck. 2021. "Domain Randomization for Demand Response of an Electric Water Heater." *IEEE Transactions on Smart Grid* 12:1370–1379. https://doi.org/10.1109/TSG.2020.3024656.

Pérez-Lombard, L., J. Ortiz, and C. Pout. 2008. "A Review on Buildings Energy Consumption Information." *Energy and Buildings* 40 (3): 394–398. https://doi.org/10.1016/j.enbuild.2007.03.007.

Philip, S. Y., S. F. Kew, G. J. Van Oldenborgh, F. S. Anslow, S. I. Seneviratne, R. Vautard, D. Coumou, et al. 2022. "Rapid Attribution Analysis of the Extraordinary Heat Wave on the Pacific Coast of the US and Canada in June 2021." *Earth System Dynamics* 13:1689–1713. https://doi.org/10.5194/esd-13-1689-2022.

Pippia, T., J. Lago, R. De Coninck, J. Sijs, and B. De Schutter. 2019. "Scenario-based Model Predictive Control Approach for Heating Systems in an Office Building." In *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, 1243–1248. Vancouver, British Columbia, Canada: IEEE Press. https://doi.org/10.1109/COASE.2019.8842846.

Pörtner, H.-O., D. C. Roberts, M. Tignor, E. Poloczanska, K. Mintenbeck, A. Alegr'ia, M. Craig, S. Langsdorf, S. Löschke, V. Möller, et al. 2022. IPPC 2022: Climate Change 2022: impacts, adaptation and vulnerability: working group II contribution to the sixth assessment report of the intergovernmental panel on climate change.

Prívara, S., J. Široký, L. Ferkl, and J. Cigler. 2011. "Model Predictive Control of a Building Heating System: The First Experience." *Energy and Buildings* 43 (2-3): 564–572. https://doi.org/10.1016/j.enbuild.2010.10.022.

Prototype Building Models Building Energy Codes Program. n.d. https://www.energycodes.gov/prototype-building-models (accessed April 29, 2025).

Quang, T Van, and N. L. Phuong. 2024. "Using Deep Learning to Optimize HVAC Systems in Residential Buildings." *Journal of Green Building* 19 (1): 29–50. https://doi.org/10.3992/jgb.19.1.29.

Rhinoceros. n.d. ® User's Guide for Windows.

Riahi, K., E. Kriegler, N. Johnson, C. Bertram, M. den Elzen, J. Eom, M. Schaeffer, et al. 2015. "Locked into Copenhagen Pledges — Implications of Short-Term Emission Targets for the Cost and Feasibility of Long-Term Climate Goals." *Technological Forecasting and Social Change* 90:8–23. https://doi.org/10.1016/j.techfore.2013.09.016.

Rogelj, J., D. Shindell, K. Jiang, S. Fifita, P. Forster, V. Ginzburg, C. Handa, et al. 2018. "Mitigation Pathways Compatible with 1.5°C in the Context of Sustainable Development." *Global Warming of 1.5C*, 93–174.

Ruelens, F., B. J. Claessens, S. Vandael, B. De Schutter, R. Babuška, and R. Belmans. 2017. "Residential Demand Response of Thermostatically Controlled Loads Using Batch Reinforcement Learning." *IEEE Transactions on Smart Grid* 8:2149–2159. https://doi.org/10.1109/TSG.2016.2517211.

Shteingart, H., and Y. Loewenstein. 2014. "Reinforcement Learning and Human Behavior." *Current Opinion in Neurobiology* 25:93–98. https://doi.org/10.1016/j.conb.2013.12.004.

Started G. 2022. EnergyPlusTM Version 22.2.0 Documentation.

Sutton, R. S., and A. G. Barto. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Tao, Y., J. Qiu, and S. Lai. 2022. "A Hybrid Cloud and Edge Control Strategy for Demand Responses Using Deep Reinforcement Learning and Transfer Learning." *IEEE Transactions on Cloud Computing* 10:56–71. https://doi.org/10.1109/TCC.2021.3117580.

Taylor, M. E., and P. Stone. 2009. "Transfer Learning for Reinforcement Learning Domains: A Survey." *Journal of Machine Learning Research* 10:1633–1685.

Thornton, B. A., M. I. Rosenberg, E. E. Richman, W. Wang, Y. Xie, J. Zhang, H. Cho, V. Mendon V, R. A. Athalye, and B. Liu. 2011. Achieving the 30% Goal: Energy and Cost Savings Analysis of ASHRAE Standard 90.1-2010. United States. https://doi.org/10.2172/1015277.

Turley, C., M. Jacoby, G. Pavlak, and G. Henze. 2020. "Development and Evaluation of Occupancy-Aware HVAC Control for Residential Building Energy Efficiency and Occupant Comfort." *Energies* 13:5396. https://doi.org/10.3390/en13205396.

Ürge-Vorsatz, D., L. F. Cabeza, S. Serrano, C. Barreneche, and K. Petrichenko. 2015. "Heating and Cooling Energy Trends and Drivers in Buildings." *Renewable and Sustainable Energy Reviews* 41:85–98. https://doi.org/10.1016/j.rser.2014.08.039.

Ürge-Vorsatz, D., C. Rosenzweig, R. J. Dawson, R. Sanchez Rodriguez, X. Bai, A. S. Barau, K. C. Seto, and S. Dhakal. 2018. "Locking in Positive Climate Responses in Cities." *Nature Climate Change* 8 (3): 174–177. https://doi.org/10.1038/s41558-018-0100-6.

van Otterlo, M., and M. Wiering. 2012. "Reinforcement Learning and Markov Decision Processes." In *Reinforcement Learning: State-of-the-Art*, edited by M. Wiering, and M. van Otterlo, 3–42. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-27645-3_1.

Vautard, R., M. Van Aalst, O. Boucher, A. Drouin, K. Haustein, F. Kreienkamp, G. J. Van Oldenborgh, et al. 2020. "Human Contribution to the Record-Breaking June and July 2019 Heatwaves in Western Europe." *Environmental Research Letters* 15:094077. https://doi.org/10.1088/1748-9326/aba3d4.

Wang, Z., and T. Hong. 2020. "Reinforcement Learning for Building Controls: The Opportunities and Challenges." *Applied Energy* 269:115036. https://doi.org/10.1016/j.apenergy.2020.115036.

Xu, S., Y. Wang, Y. Wang, Z. O'Neill, and Q. Zhu. 2020. "One for Many: Transfer Learning for Building HVAC Control." In *BuildSys 2020 - Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 230–239. New York, NY, USA: Association for Computing Machinery, Inc. https://doi.org/10.1145/3408308.3427617.

Yu, L., W. Xie, D. Xie, Y. Zou, D. Zhang, Z. Sun, L. Zhang, Y. Zhang, and T. Jiang. 2020. "Deep Reinforcement Learning for Smart Home Energy Management." *IEEE Internet of Things Journal* 7:2751–2762. https://doi.org/10.1109/JIOT.2019.2957289.