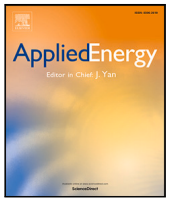
Applied Energy 368 (2024) 123447

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

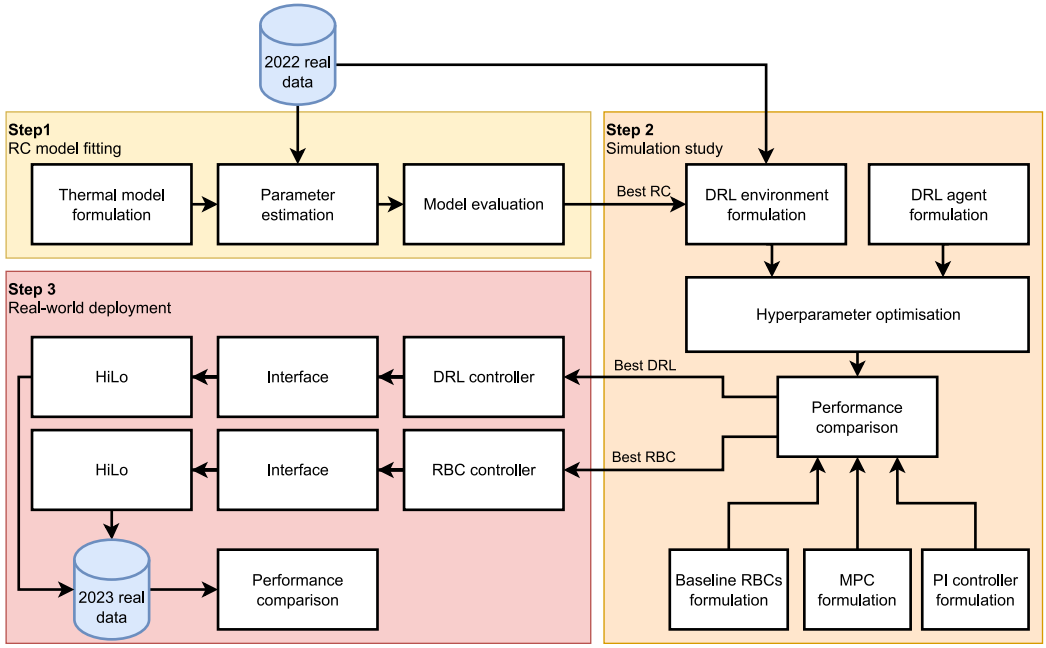


Real building implementation of a deep reinforcement learning controller to enhance energy efficiency and indoor temperature control

Alberto Silvestri [a,*], Davide Coraci [b], Silvio Brandi [b], Alfonso Capozzoli [b], Esther Borkowski [a], Johannes Köhler [c], Duan Wu [d], Melanie N. Zeilinger [c], Arno Schlueter [a]

[a] Architecture and Building Systems, ETH Zurich, Zurich, Switzerland
[b] Politecnico di Torino, Department of Energy, TEBE research group, BAEDA Lab, Torino, Italy
[c] Institute for Dynamic Systems and Control, ETH Zurich, Zurich, Switzerland
[d] Mitsubishi Electric R&D Centre Europe B.V., Livingston, UK

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Deep reinforcement learning
Real implementation
Building energy management
HVAC control
Energy efficiency

ABSTRACT

Deep Reinforcement Learning (DRL) has emerged as a promising approach to address the trade-off between energy efficiency and indoor comfort in buildings, potentially outperforming conventional Rule-Based Controllers (RBC). This paper explores the real-world application of a Soft-Actor Critic (SAC) DRL controller in a building's Thermally Activated Building System (TABS), focusing on optimising energy consumption and maintaining comfortable indoor temperatures. Our approach involves pre-training the DRL agent using a simplified Resistance-Capacitance (RC) model calibrated with real building data. The study first benchmarks the

* Corresponding author.
E-mail address: silvestri@arch.ethz.ch (A. Silvestri).

https://doi.org/10.1016/j.apenergy.2024.123447
Received 20 December 2023; Received in revised form 30 April 2024; Accepted 9 May 2024
Available online 21 May 2024

Temperature control

DRL controller against three RBCs, two Proportional-Integral (PI) controllers and a Model Predictive Controller (MPC) in a simulated environment. In the simulation study, DRL reduces energy consumption by 15% to 50% and decreases temperature violations by 25% compared to RBCs, reducing also energy consumption and temperature violations compared to PI controllers by respectively 23% and 5%. Moreover, DRL achieves comparable performance in terms of temperature control but consuming 29% more energy than an ideal MPC. When implemented in a real building during a two-month cooling season, the DRL controller performances were compared with those of the best-performing RBC, enhancing indoor temperature control by 68% without increasing energy consumption. This research demonstrates an effective strategy for training and deploying DRL controllers in real building energy systems, highlighting the potential of DRL in practical energy management applications.

1. Introduction

In recent years, the interest in the possible improvement of energy efficiency in buildings has increased worldwide, as buildings constitute one of the major contributors to both global energy consumption (i.e., 40%) and emissions of greenhouse gas (i.e., 30%) [1]. In this context, energy management emerges as a viable solution to enhance energy systems operation [2] by reducing energy costs and improving indoor comfort conditions for the occupants [3]. Notably, Heating, Ventilation and Air Conditioning (HVAC) systems account for the highest energy consumption in buildings. Substantial enhancements have been introduced to improve their energy efficiency by implementing more effective system solutions coupled with advanced energy management strategies [4].

Thermally Activated Building System (TABS) has emerged as a promising solution in minimising energy consumption and improving thermal comfort in office buildings. Utilising the thermal inertia of the building structure, TABS stores and releases heat, actively shaping the indoor temperature conditions. Nevertheless, ensuring the efficient management and control of TABS and other HVAC systems is crucial for maximising energy efficiency while preserving optimal thermal comfort conditions.

Currently, HVAC systems are predominantly controlled using *ON-OFF* Rule-Based Controller (RBC) [5]. These methods rely on expert knowledge and predetermined schedules outlined in the ASHRAE Guidelines 36 [6]. Although developed by building control experts, these controllers may behave suboptimally since they are not able to optimise multi-objective control problems, being reactive controllers [7]. Specifically, they cannot dynamically adapt their control policies based on predictions of external factors, such as weather conditions, that impact energy consumption and comfort conditions in buildings [8]. Furthermore, RBC strategies lack optimisation capabilities and are not able to handle multiple and contrasting objective functions [9].

In this context, the increased availability of historical building data, facilitated by the widespread use of Internet of Things (IoT) devices and Information and Communication Technologies (ICT) [10], becomes particularly valuable. This data enables the development of advanced control strategies that estimate and predict current and future building states and energy system dynamics, addressing the limitations of current HVAC control methods [11].

To address the challenges related to the implementation of RBC for HVAC systems, the adoption of advanced control strategies employing predictive and adaptive methods has emerged as a viable solution. These advanced control approaches for HVAC systems primarily fall into two categories: model-based and model-free methods. In model-based methods, the key elements of control systems are a model representing the controlled environment and an optimiser, while model-free methods do not require any model of the environment to be controlled since they learn a near-optimal control policy employing a trial-and-error process with the system to be controlled.

Among model-based control methods, MPC addresses the primary challenges of HVAC system control, including nonlinear and time-varying dynamics and disturbances, by conducting optimisation over a receding control time horizon. MPC utilises a mathematical model of the controlled system and employs an online numerical optimisation to compute optimal control signals over a given time horizon [12,13] also taking into account the possible evolution of system dynamics [14]. MPC has gained attention in building control field for its optimal predictive capabilities [15], showing effectiveness in optimising HVAC system operation [16,17]. Nevertheless, the practical implementation of MPC is limited by its reliance on a model-based approach, requiring a sufficiently detailed characterisation of the specific building and energy system in which the MPC is implemented [18]. This is particularly relevant for HVAC systems, where each building presents a unique entity, making control-oriented modelling of their envelope and energy systems challenging. On top of that, each building is different and subject to various internal and external conditions (i.e., occupancy patterns, outdoor weather), geometries and thermophysical properties. Consequently, despite its robustness and advantages, MPC adoption remains limited in the building industry [19].

As an alternative solution, Reinforcement Learning (RL) emerges among model-free control approaches for its great potential in optimising building energy system control strategies to reduce energy costs related to building operation while enhancing indoor temperature control and comfort conditions for the occupants. RL controllers learn a near-optimal control policy through direct environment interaction, employing a trial and error approach [20]. Due to the complexity and non-linearity of building control problems, Deep Neural Network (DNN) are commonly used in RL algorithms, leading to a DRL approach.

The following subsection provides an overview of the existing literature on the applications of DRL to optimise the operation of HVAC systems in buildings. A more general background on RL theoretical foundations and algorithms is given in Appendix.

1.1. Related works on reinforcement learning applications

DRL has gained significant traction in recent years due to its capability to address the primary challenges of HVAC system control, including nonlinear and time-varying dynamics and disturbances and to handle complex and high-dimensional control problems with conflicting objectives (e.g., the trade-off between energy consumption and occupant comfort), which makes it particularly suitable for managing building HVAC systems, characterised by non-linearity and uncertainty [21–23].

For example, recent studies have exploited DRL in simulation to control the supply water temperature [24,25], the mass flow rate in thermal systems [26,27], the indoor temperature setpoint [28], the operation mode of generation systems [29,30], and both thermal [31] and electrical storage systems [32]. Table 1 reports related works on DRL application developed in simulative way, providing details about objectives, implemented control strategies, energy system and key outcomes.

Despite its potential, there are limited results on the real-world implementation of DRL in HVAC systems [33], particularly in the context of TABS. To address this gap, [34] implemented a DRL algorithm for a radiant heating system in a real office building. The authors employed EnergyPlus to create a physics-based model, calibrated it with measured building data, utilised the model for DRL agent training using the Asynchronous Advantage Actor Critic (A3C) policy gradient method, and deployed the trained agent in the heating system. This

Nomenclature

α	Boltzmann temperature coefficient
$\dot{Q}_{\text{sol}}$	Solar radiation [W/m^2]
$\dot{Q}_{\text{tabs}}$	Cooling power delivered by TABS [W]
η_d	Performance normalisation factor for outdoor temperature and solar radiation
γ	Discount factor
λ	Energy term weight of reward function
$\mathbb{E}$	Expected value
$\mathcal{H}$	Shannon entropy term
μ	Learning rate
$\overline{\dot{Q}}_{\text{tabs,norm},d}$	Normalised daily mean cooling power [W]
$\overline{T_{\text{i}}}$	Upper limit of temperature comfort range [°C]
π	Control policy
π^*	Optimal control policy
τ	DRL controller soft-update coefficient
θ	Reward scaling factor
$\underline{T_{\text{i}}}$	Lower limit of temperature comfort range [°C]
a	Control action at control time step t
b_{occ}	Occupancy boolean variable
E_{tabs}	Energy consumption associated with the TABS operation [kWh]
$Q^{\pi}(s,a)$	Action-value function
$R(s,a)$	Reward function
r_E	Energy term of reward function
r_T	Temperature term of reward function
s	Environment state at control time step t
s'	Environment state at control time step t+1
t_{end}	Occupancy end time
t_{start}	Occupancy start time
T_{i}	Indoor air temperature [°C]
T_{o}	Outdoor air temperature [°C]
$T_{\text{viol,norm},d}$	Normalised daily cumulative temperature violations [°C]
T_{viol}	Temperature violation [°C]
u_t	Percentage opening of the valve
$V^{\pi}(s)$	State-value function

Acronyms

A3C	Asynchronous Advantage Actor Critic
DDPG	Deep Deterministic Policy Gradient
DNN	Deep Neural Network
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
HVAC	Heating, Ventilation and Air Conditioning
HVRF	Hybrid Variable Refrigerant Flow
ICT	Information and Communication Technologies
IoT	Internet of Things
MDP	Markov Decision Process
MSE	Mean Squared Error
ODBC	Open DataBase Connectivity
ODE	Ordinary Differential Equation
PI	Proportional-Integral
PLC	Programmable Logic Controller
PPO	Proximal Policy Optimisation
RBC	Rule-Based Controller
RC	Resistance-Capacitance
RL	Reinforcement Learning
RMSE	Root Mean Squared Error
SAC	Soft Actor-Critic
TABS	Thermally Activated Building System
RES	Renewable Energy Sources
TL	Transfer Learning
TPE	Tree-structured Parzen Estimator

approach, integrating an occupant feedback system via a smartphone app, led to an approximately 18% reduction in heating demand compared to the old rule-based controller over a three-month deployment period. However, several challenges were identified, including low user engagement in the feedback system, inefficiencies in DRL training and difficulties in data handling.

Another study by [35] investigated an offline multi-agent DRL algorithm for a radiant floor heating system. The research involved two tests: a comparative test with a traditional rule-based controller and another with the DRL controller. The benchmarking data were used in the DRL algorithm to train and then deploy it, revealing that the DRL controller demonstrated robustness, performed predictive control-like actions, reduced temperature oscillation by 43%–63%, and decreased overall costs by 14% to 16%.

A study by [38] introduced a control framework, named *DeepValve*, based on the Double Deep Q-learning algorithm. The authors trained the DRL controller in a simulated environment, utilising a surrogate model of a simplified building and considered training data from twenty different buildings. The controller was then tested on three additional buildings, whose dynamics were also represented by a simplified model. Finally, the controller's performance was tested in a real testbed at EPFL, Switzerland, featuring two occupants and a floor and ceiling heating system. The study found a 44% reduction in energy consumption and improved internal comfort compared to the RBC baseline, demonstrating its adaptability to different real-building conditions. However, this work did not provide a benchmark for evaluating DRL performance against the baseline in real-world tests and designed the DRL as high-level controllers without direct interaction with the energy system, relying instead on enabling or disabling the low-level RBC managing the indoor temperature setpoint of the experimental building.

These studies highlight the potential of DRL in enhancing energy efficiency within TABS control systems. However, real-world implementations still face challenges that should be addressed in future research to harness DRL's potential in real-world buildings [33].

1.2. Research gaps, novelty and contributions

From the literature analysis, it emerges that DRL can be an effective strategy to optimise the operation of energy systems in buildings while ensuring comfortable conditions for occupants. However, to the best of our knowledge, the performance of DRL has been mostly tested by researchers in a simulated environment and only in a few studies in a real experimental testbed. A common approach widely explored in DRL literature evaluated the controller performance using detailed engineering building models such as those created in EnergyPlus [39]. However, this is a time-consuming task that requires both domain expertise and information about the thermal properties of the building, which may not always be readily available [40]. A potential solution to this problem could be the development of a simplified RC model that is used both during the training and testing phases of the controller. Although it retains physical knowledge of the system, it is subject to

Table 1
Summary of simulated DRL applications in building system management.

Ref	Objectives	Control strategies	Energy system	Key outcomes
Zhang et al. [25]	Reduce energy consumption, enhance indoor comfort	A3C	TABS	17% reduction in energy savings, slight increase in PPD
Schreiber et al. [26]	Reduce energy cost, temperature regulation	DQN, DDPG	Chiller	14% reduction in weekly electrical cost
Brandi et al. [30]	Minimise electricity cost in charging/discharging	SAC, MPC	Chiller and TES	Comparable performance with MPC
Wang et al. [31]	Operation management for cost reduction	Deep Q-network	Chiller and TES	8% cost reduction compared to fixed-schedule control
Lei et al. [36]	Reduce cooling energy consumption, improve thermal comfort	Dueling DQN	Hybrid cooling system	14% less cooling energy, 11% improvement in thermal acceptability
Silvestri et al. [37]	Reduce energy costs, decrease temperature violations	PPO, SAC	TABS	PPO: 18% cost reduction, 33% fewer temperature violations; SAC: 14% cost reduction, 64% fewer temperature violations

unavoidable model inaccuracies. Therefore, training a DRL controller with a grey-box model as a proxy for the real building dynamics poses the risk that the control policy may not perform as effectively in real-world scenarios. Furthermore, DRL controllers are commonly employed to function as high-level controllers, as they do not directly interact with the energy system. For instance, as in [38] the DRL was operated to decide whether to enable or disable the low-level RBC managing the indoor temperature setpoint of the experimental building.

Finally, another critical aspect rising from the literature review, is the lack of a benchmark methodology to evaluate the DRL performance against the implemented baseline in the experimental setting.

Thus, the present paper demonstrates the feasibility and effectiveness of DRL in real-world settings, bridging the gap between theoretical research and practical application, discussing the real implementation challenges.

In detail, a DRL controller is pre-trained in a simulation environment employing a RC model as the thermal building surrogate model. During the simulative phase, the performance of the DRL controller was compared with that of an MPC-based controller, two PI controllers and three RBCs.

Afterwards, the DRL controller is deployed in the real testbed to validate its performance on the real building from which the data were extracted for RC model development. A benchmarking procedure is introduced to compare the DRL performance during its real deployment with that of the RBC implemented as benchmark controller. Establishing a performance benchmark for DRL during its real implementation is one of the primary contributions of our research and it is crucial for a comprehensive and meaningful assessment of DRL efficacy in real-world applications.

The DRL controller operates similarly to a thermostatic controller since it is developed to directly manage the optimal percentage opening of the installed in the supply circuit connected to the TABS during a cooling season lasting 2 months (July and August 2023). The objective of DRL agent is to minimise energy consumption while maintaining the indoor temperature between [22, 24] °C.

An overview of the main contributions of this paper is provided as follows:

- A methodological approach for training a Soft Actor-Critic (SAC) controller within a Python-developed simulated environment, utilising a RC model to represent building dynamics has been introduced in this paper. The SAC control agent operates as a low-level controller since it directly manages the percentage opening for the valve installed in the TABS system. Following its training phase, the SAC controller is implemented in a real-world testbed to validate its performance, marking a significant step towards bridging the gap between simulation-based training and real-world application.
- A methodology that exhibits a high degree of interoperability with monitoring and actuation systems has been developed, as it did not require the installation of additional components for effective functionality. Furthermore, it required only a minimal number of sensors to measure indoor and outdoor temperatures, solar radiation and occupant presence.
- A framework for benchmarking the performance of the DRL controller against the RBC in a real-world testbed, has been introduced. This methodological contribution provides a structured approach to evaluate and compare the efficacy of DRL and RBC strategies, offering valuable insights into the practical advantages, challenges and drawbacks of implementing a DRL control systems in real building environments.
- Valuable insights into the practical advantages, challenges and drawbacks of implementing a DRL controllers in real building environments have been extensively discussed to provide guidelines for future implementations in real-world.

The paper is structured as follows: Section 2 provides information regarding the case study and the formulation of the optimal control problem. Section 3 introduces the methodological framework utilised for both training and deploying the DRL controller. Implementation specifics concerning the simulation and deployment phases are outlined in Section 4. Section 5 describes the results, while Sections 6 and 7 discuss the outcomes of this research and provide future directions.

2. Case study and control problem formulation

The NEST building, part of the Swiss Federal Laboratories for Materials Science and Technology (EMPA), is a modular research and innovation facility located in Dübendorf, Switzerland [41]. Opened in 2016, NEST serves as a living lab where partners from academia, industry, and the public sector collaborate. The building features a central backbone and three open platforms, accommodating various research and innovation modules.

Our study focuses on the HiLo (High Performance – Low Emissions) unit, one of the most recent modules within the NEST building [42], shown in Fig. 1. This innovative research environment, designed for testing and developing sustainable building technologies, presents an ideal case study for our controller being a living lab with a large availability of sensors and data.

The HiLo unit encompasses two floors. The lower level houses two office spaces, while the upper floor is allocated for an open-plan area. In our study, the office represented in Fig. 2 and located on the southwest side, covering an area of 22.94 m^2 was employed as case study. This office is equipped with three distinct HVAC systems: a ceiling-mounted integrated TABS, a Hybrid Variable Refrigerant Flow (HVRF) system, and a mechanical ventilation with heat recovery. Due to limitations

Fig. 1. The HiLo unit at EMPA NEST.

in accessing low-level controls, our study focused exclusively on the TABS system. An integrated sensor network monitors relevant physical variables. The data is systematically archived in a dedicated MS-SQL database. For the purpose of this work, only a small portion of the available data points are used, reflecting the data availability limitations encountered in common buildings. Specifically, only the data from the indoor and outdoor temperature sensors, the solar radiation sensor, and the occupancy detector are employed. The office is usually occupied from Monday to Saturday from 7:00 to 21:00 by one or two occupants.

This study focuses on the management of the office's thermal environment via the TABS to reduce the energy consumption associated with its operation while ensuring that, during working hours, the indoor temperature remains within an acceptable range, defined as between 22 and 24 °C. The core aspect involves regulating every 5 min (i.e., corresponding to the control time step value) the opening of the valve that governs the TABS water flow rate and, consequently, the thermal power output. In particular, the valve is of changeover type, and the inlet temperature is fixed at a constant value and cooled down by the chilled water provided by the backbone.

3. Methodology

This section provides details about the methodological framework adopted in this paper, represented in Fig. 3. The framework comprises three key steps: (i) the formulation and fitting of the simplified RC model, (ii) the design and training of the DRL control agent in simulation, and (iii) the description of the real deployment phase for both RBC and DRL controllers.

3.1. RC model fitting

In the field of building physics, modelling is essential for simulating thermal behaviour, which is crucial in the design, sizing, and operational optimisation of buildings [43]. Building thermal models can be categorised into three primary approaches: white-box, black-box, and grey-box [43]. White-box models are founded on physical laws with parameters derived from material properties, providing detailed insights into building dynamics. Conversely, black-box models, detached from physical laws, use data-driven approaches to approximate the effects of inputs on outputs. Grey-box models, a hybrid of these approaches, simplify white-box models by estimating parameters from empirical data.

This study employed a grey-box modelling approach, specifically utilising a RC model to capture the thermal dynamics of the office environment, similarly to [43]. The decision to employ a RC model was driven by the need for a simulation tool that balances speed and transparency. Unlike purely black-box models that offer quick results but lack interpretability, the RC model provides a more comprehensible framework while still ensuring efficient simulation. This allows for a faster modelling process without completely sacrificing the insight into the physical phenomena governing building thermal dynamics. In this approach, the thermal properties of a building are depicted as a network of resistors and capacitors. The resistors represent the thermal resistance of building materials such as walls, windows, and roofs, indicating their capacity to contrast heat flow. Capacitors, in contrast, model the building's thermal mass, reflecting its ability to store and release heat, thus capturing the dynamics of heat transfer and storage within the building.

The design of this model aimed to achieve a balance between simplicity and the complexity required to accurately represent essential thermal characteristics, including heat transfer through the building envelope and the thermal inertia of the TABS. First, various RC network structures are defined, followed by parameter estimation using historical data from the HiLo office.

3.2. Simulation study

The second methodological step considers the design and training of the DRL controller in a simulated environment, being a crucial step before the implementation of the controller in the real testbed. In

Fig. 2. HiLo office zone employed as test facility.

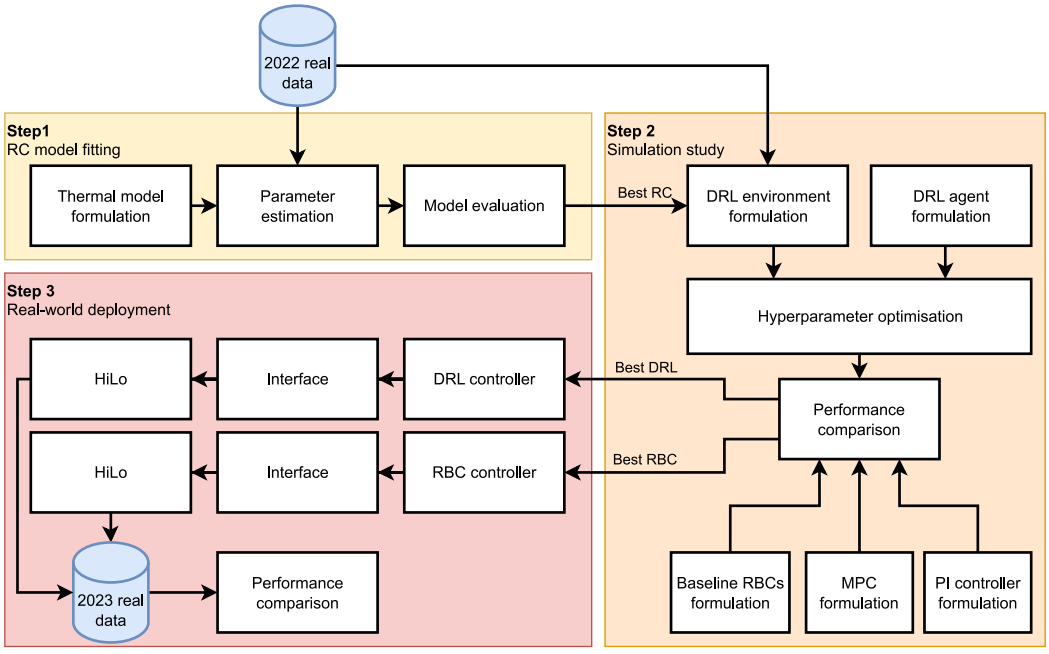


Fig. 3. Methodological framework developed in this work.

this study, the latest version of the SAC algorithm within the Stable-Baselines library [44] is employed. SAC is chosen as DRL algorithm since it is more sample efficient considering that it utilises a replay buffer to store its experiences, enabling it to learn from the same samples multiple times throughout the training process [45]. This aspect is more important in real-world applications where collecting data or samples can be expensive or impractical. By maximising the utilisation of each sample through the replay buffer mechanism, SAC helps in reducing the overall data collection cost, making it more feasible for real-world implementations. Moreover, learning from limited data is a common challenge in real-world scenarios, where access to large datasets may be restricted. SAC's sample efficiency ensures robust

learning even with limited data, which is crucial for applications where gathering extensive training data is impractical or expensive Moreover, SAC provides better performance than other DRL methods in case of large state spaces, as it occurs in real-world control problems [46]. Three rule-based controllers (i.e., $RBC1$, $RBC2$ and $RBC3$), two PI controllers and an MPC are developed to provide a benchmark with the DRL performances. The comparison with RBCs is carried out since they respectively represent the same strategy implemented previously in the building ($RBC1$) and its two improved versions. On the other hand, the comparison with PI is carried out to provide a benchmark with a typical continuous controller. Finally, a comparison with an idealised MPC aims to demonstrate that the DRL controller in simulation is capable of achieving performance comparable to a theoretical near-optimal solution.

Following the formulation of the DRL controller, its training occurred offline in a simulation environment developed in Python by employing the RC model to emulate the building dynamics. Throughout the DRL agent training phase, an automated procedure is executed using Optuna [47] to determine the optimal configuration of hyperparameters for the control algorithm. Given that the performance of DRL controllers is significantly affected by the selection of these variables, this automated process aims to identify the best control agent among the investigated controllers [48]. Specifics of this training approach are described in detail in Section 4.6.

3.3. Real-world deployment

The deployment of the controllers within the office building included on-site implementation and testing, performance benchmarking and incorporating a fail-safe mechanism.

This integration was carried out in line with the infrastructure provided by NEST, ensuring effective communication and control over the building's HVAC systems. As detailed in Section 4.7, software adaptations were critical to harmonise the controllers with the building's existing technological framework.

Benchmarking the performance of the DRL controller was a key step in evaluating its efficacy. This process involved a comparative analysis against the $RBC3$. The benchmarking focused on assessing energy efficiency and indoor temperature regulation under varying operational conditions.

A critical aspect of the deployment was the implementation of a fail-safe mechanism. In the eventuality that the DRL controller malfunctioned or lost connection with the system, the control logic was automatically reverted to the default control mode developed by the system integrators in NEST. This fail-safe protocol ensured the continuous and stable operation of the building's climate control systems, mitigating potential disruptions from DRL system failures.

4. Implementation

This section provides details about the development of the RC simulation model, RBCs, PI controllers, DRL and MPC strategies. Moreover, it includes information about the training of the DRL control agent and its subsequent real-world deployment.

4.1. Simulation model

Four different RC networks were defined in this work. The first model is a first-order model m_1 described by the following Ordinary Differential Equation (ODE):

$$m_1 : C_i\dot{T}_i = \frac{1}{R_w}(T_o - T_i) + \frac{1}{R_n}(T_n - T_i) - \dot{Q}_{tabs} + \phi_i\dot{Q}_{sol} \tag{1}$$

where T_i, T_o, T_n are the zone temperature, the outdoor ambient temperature and the neighbour room temperature, respectively. C_i represents the zone thermal capacitance. R_w and R_n are the thermal resistances of the building envelope and of the inner wall. The cooling power delivered by the TABS is represented by $\dot{Q}_{tabs}$. The solar radiation $\dot{Q}_{sol}$ is multiplied by a scaling factor $\phi_i \in [0, 1]$.

Two second-order models are formulated as follows:

$$m2 : \begin{cases} C_i\dot{T}_i = \frac{1}{R_w}(T_w - T_i) + \frac{1}{R_n}(T_n - T_i) - \dot{Q}_{tabs} + \phi_i\dot{Q}_{sol} \\ C_w\dot{T}_w = \frac{1}{R_w}(T_i - T_w) + \frac{1}{R_w}(T_o - T_w) + \phi_w\dot{Q}_{sol} \end{cases} \tag{2}$$

and

$$m2s : \begin{cases} C_i\dot{T}_i = \frac{1}{R_w}(T_o - T_i) + \frac{1}{R_n}(T_n - T_i) + \frac{1}{R_s}(T_s - T_i) + \phi_i\dot{Q}_{sol} \\ C_s\dot{T}_s = \frac{1}{R_s}(T_i - T_s) - \dot{Q}_{tabs} + \phi_s\dot{Q}_{sol} \end{cases} \tag{3}$$

The state T_w in Eq. (2) represents the external wall temperature, and C_w is its thermal capacitance. Eq. (3) describes the model m_{2s} where the temperature and the thermal mass of the TABS are modelled as T_s and C_s, respectively. R_s accounts for the thermal resistance between the TABS and the zone. In both models, the solar radiation $\dot{Q}_{sol}$ is considered as an additional disturbance acting on both states and it is scaled by the factors ϕ_i, ϕ_w and $\phi_s \in [0, 1]$. A graphical representation combining the model $m2s$ and the office zone employed as the case study is provided in Fig. 4.

Combining these two models results in the following third-order model, modelling the thermal dynamics of the zone, envelope and TABS:

$$m3 : \begin{cases} C_i\dot{T}_i = \frac{1}{R_w}(T_w - T_i) + \frac{1}{R_n}(T_n - T_i) + \frac{1}{R_s}(T_s - T_i) + \phi_i\dot{Q}_{sol} \\ C_w\dot{T}_w = \frac{1}{R_w}(T_i - T_w) + \frac{1}{R_w}(T_o - T_w) + \phi_w\dot{Q}_{sol} \\ C_s\dot{T}_s = \frac{1}{R_s}(T_i - T_s) - \dot{Q}_{tabs} + \phi_s\dot{Q}_{sol} \end{cases} \tag{4}$$

The data employed for the RC model parameter estimation process have been collected under the baseline control policy, implemented before the experiments conducted in this paper. Empirical data was separated into two distinct periods. The data spanning from July 28, 2022, to August 11, 2022, was used for parameter estimation, while the data from August 19, 2022 to August 26, 2022, served for model validation and testing. MATLAB functions *idgrey* and *greyest*, alongside the optimisation function *fmincon*, were used in this process. The *idgrey* function facilitated the definition of grey-box models in state-space form, while *greyest* with the *fmincon* function was utilised to estimate model parameters by fitting these grey-box models to the training dataset, effectively minimising the Mean Squared Error (MSE) cost function. The MSE measures the average squared difference between actual and predicted values of the indoor temperature T_i.

Finally, the RC model with the lowest Root Mean Squared Error (RMSE) has been integrated into *openAI Gym* [49], a Python toolkit developed to standardise the interface of the environment used to train DRL algorithms.

4.2. Rule-based controllers

During the simulation phase, three different RBCs were considered for benchmarking the performance of the DRL controller. Then, the rule-based strategy that ensured the best performance was employed as the benchmark during the implementation phase in the real testbed. The three RBCs are based on the bang–bang control logic, switching between fully closing (i.e., 0%) or fully opening (i.e., 100%) the valve.

The first rule-based control strategy, $RBC1$, operates throughout the day, and it fully opens the valve when the indoor temperature T_i exceeds the upper limit of the acceptable temperature range $\overline{T_i}$ (i.e., 24 °C) and closes the valve when the temperature is lower than the lower temperature limit $\underline{T_i}$ of 22 °C. $RBC1$ mimics the same control logic of the baseline controller previously implemented in the real

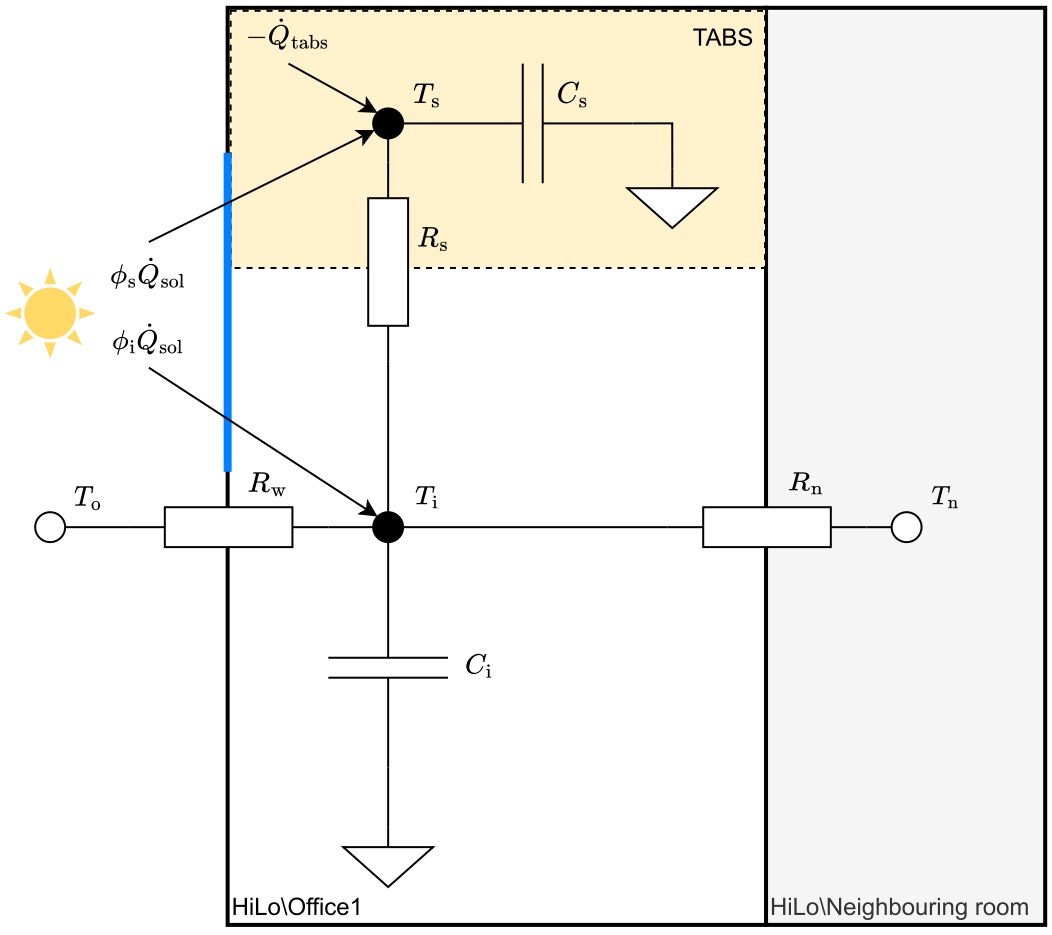


Fig. 4. Graphical representation of *m2s* RC model considered in this study.

building. This controller has been modified to enhance its performance by considering first the occupant schedule (*RBC*2) and then also the pre-cooling (*RBC*3) to improve indoor temperature conditions. Therefore, the second rule-based controller *RBC*2 is similar to *RBC*1 but only operates during the occupied hours, i.e. Monday-Saturday from t_{start} = 7:00 to t_{end} = 21:00.

Fig. 5 shows the last rule-based control strategy, named *RBC*3, which consists of two pre-cooling phases *Pre-Cooling 1* and *Pre-Cooling 2*, and a normal *Cooling* phase. The rules reported for the pre-cooling phases in Fig. 5 resulted from a sensitivity analysis where different combinations of start time and indoor temperature thresholds were tested to minimise temperature violations at the start of the occupancy period. The *Pre-Cooling 1* phase lasts from $t_{\text{start}} - 4h$ = 3:00 to $t_{\text{start}} - 3h$ = 4:00 and activate the TABS when $T_{\text{i}} \geq \overline{T_{\text{i}}} + 1$ °C. Next, the *Pre-Cooling 2* mode is activated until $t_{\text{start}} - 2h$ = 5:00, when $T_{\text{i}} \geq \overline{T_{\text{i}}} + 0.5$ °C. In both cases, the valve is closed if the indoor temperature is lower than the lower limit $\underline{T_{\text{i}}}$. During the pre-cooling phases, no temperature violations are taken into account since no occupant is present, as defined in Eq. (12). Following the two pre-cooling phases at t_{start} = 7:00, *RBC*3 manages the system as per *RBC*1. This RBC strategy remains active until occupants leave the building at t_{end} = 21:00 and ensures that the valve is fully closed on Sundays to save energy since the office zone is not occupied.

4.3. PI controllers design

In a continuous-time PI controller, the control output $u(t)$, is determined as:

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau)\, d\tau \tag{5}$$

where, $e(t)$ is the error signal, defined as the difference between the setpoint and the measured indoor temperature T_{i}. The proportional gain K_p scales the error directly, allowing the controller to respond proportionally to the magnitude of the error. The integral gain K_i scales the integral of the error over time, addressing any accumulated error that persists over time.

In this study, the controller's gains K_p and K_i have been tuned using MATLAB's *PID Tuner* set on *balanced* performance. Moreover, the control action u is saturated in the range [0,1], as for the other controllers. To avoid integral windup, an anti-windup scheme has been included in the controller [50].

This study includes two PI controllers: PI_{23} and PI_{24}. PI_{23} is configured to follow an indoor temperature setpoint of 23 °C, which falls in the centre of the temperature range [22,24] °C. PI_{24} is set to maintain the indoor temperature at the upper limit of 24 °C. Finally, both controllers only operate from Monday to Saturday from 3:00 to 21:00, similarly to the RBC described in Section 4.2.

4.4. MPC strategy design

The MPC approach involves predicting the future states of a controlled system over a finite time horizon using a dynamic model f of the process. The optimisation problem consists of computing, at each time-step t, the control sequence $u_t, u_{t+1}, \ldots, u_{t+N-1}$, that minimise the cumulative cost over N stages, where $N \in \mathbb{N}_0$ is the optimisation horizon. Once the optimal input trajectory has been computed, only the first control input is applied to the system until the next time step, when the horizon is shifted, and the optimisation process is repeated in a receding horizon fashion [15]. Thus, in our formulation, the

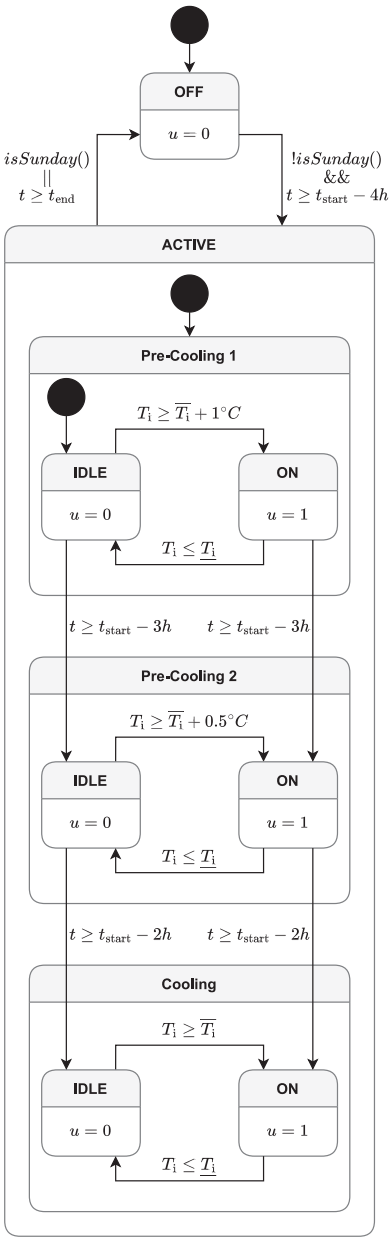


Fig. 5. *RBC*3 control logic represented as a finite state machine.

constrained optimal problem can be written as follows:

$$\min_{u_t,\ldots,u_{t+N-1}} -\sum_{t=0}^{N-1} r(s_t, u_t) \tag{6}$$

$$\text{s.t.} \quad s_{t+1} = f(s_t, u_t, d_t) \tag{7}$$

$$u_t \in [0, 1] \tag{8}$$

where r is the reward formulated in Eq. (10), and d_t is the vector of measurable disturbances (e.g., outdoor temperature T_{o}). In this paper, the MPC assumes full knowledge of the future disturbances d_t, with a prediction horizon of 24 h. The controller was provided with the current measurements and the predictions of the input variables of the RC model m_{2s} defined in Eq. (3). Therefore, it is not a realistic implementation, but an approximation for the optimal controller that could be achieved considering no modelling error and no additional noise. This ideal MPC has access to perfect knowledge of the system dynamics.

4.5. DRL controller design

This section discusses the design of the DRL controller, defining its main components: action-space, state-space, and reward function.

The action-space A consists of all possible actions that can be performed by the agent. In this paper, the action space A is continuous as required by SAC algorithm, and defined as:

$$A : u_t \in [0, 1] \tag{9}$$

At each control time step t (i.e., every 5 min), the agent selects the percentage of valve opening u_t, which is directly proportional to the fraction of the nominal cooling power $\dot{Q}_{\text{tabs}}$ supplied by the TABS. The control actions u_t chosen by the DRL are set to 0 if their value is less than 0.1. This threshold value was determined according to the operational characteristics of the TABS.

The state-space includes a set of observations presented as input to the agent, which are described in Table 2, along with the reference control time step and their respective lower and upper boundaries. These boundaries are employed for re-scaling the state space via *min–max* normalisation before providing the variables as input to the DRL controller.

Outdoor air temperature T_{o} and *Solar Radiation* $\dot{Q}_{\text{sol}}$ are included in the state-space since they are the main exogenous drivers that impact building cooling energy consumption and indoor temperature. The observations of outdoor temperature include both historical measurements and predictions. The agent is provided with two-hourly lagged values and the hourly predictions for the following six hours. Information regarding *Indoor temperature* has been assessed not by considering its value itself but by defining the temperature difference relative to the two temperature limits specified by the acceptable temperature range, thereby ensuring an adaptive definition as further defined in the temperature-related term of reward function in Eq. (11). By combining these two variables, the DRL agent knows the indoor temperature status with respect to the temperature acceptability range. The information related to the indoor air temperature is integrated into the state-space at the current control time step t and for 2 lagged values (15-min and 30-min before) to assess the temperature progression in the building over time and accounting for the influence of building thermal dynamics [51]. Furthermore, to assess the influence of neighbouring spaces on the analysed thermal zone, the temperature of the adjacent room T_{n} has been included as one of the states. To conclude, *Time to occupancy start* and *Time to occupancy end* are the two state-space variables combining in one metric the information related to the time of the day and occupancy status of the thermal zone [52]. When the building is unoccupied, time to occupancy start represents the number of control timesteps remaining before occupants' scheduled arrival time, while during the occupied period time to occupancy end represents the number of timesteps until occupants' departure time. These variables are set to zero, respectively during occupied and non-occupied periods.

The reward function provided to the agent after the action selection must be defined according to the control objectives, finding a trade-off between the opposing terms, i.e., the energy-related and temperature-related terms, weighted by employing the reward factor $\lambda > 0$.

The reward formulation r is as follows:

$$r = -\theta(\lambda \cdot r_{\text{E}} + r_{\text{T}}) \tag{10}$$

where θ is a reward scaling factor. Employing a scaling factor in the reward function serves as an effective strategy to maintain training stability, particularly in cases where rewards exhibit high variance [45]. The value of $\lambda = 2.1$ has been designed so that the agent receives the same penalty for using 0.48 kWh and for being 1 °C outside of the comfort bounds, similarly to [37,53].

The energy-related term r_{E} is equal to E_{tabs} [kWh], the energy consumption associated with the TABS operation and proportional to

Table 2
Variables included in the state-space.

Variable	Min value	Max value	Unit	Timestep
T_o	12	38	°C	t − 2 h, t − 1 h, t, t + 1 h, …, t + 6 h
$\dot{Q}_{sol}$	0	1300	W/m^2	t
$T_i - \overline{T_i}$	−10	10	°C	t, t − 15 min, t − 30 min
$T_i - \underline{T_i}$	−10	10	°C	t, t − 15 min, t − 30 min
T_n	15	30	°C	t
Time to occupancy start	0	407	–	t
Time to occupancy end	0	169	–	t

the control action u_t. The temperature-related term r_T has two different formulations according to the presence of occupants:

$$r_T = \begin{cases} 0 & \text{if } b_{occ} = 0 \\ \max(0, T_i - \overline{T_i})^2 + \max(0, \underline{T_i} - T_i)^2 & \text{if } b_{occ} = 1 \end{cases} \quad (11)$$

where b_{occ} is a Boolean variable being 1 during working hours and 0 otherwise.

4.6. Training of DRL controller

This section discusses the training phase of DRL controller carried out by employing the RC model to simulate the building dynamics. The simulation environment, the RC model and the DRL controller are all designed and implemented in Python. During the training phase, the open-source Python library Optuna [47] was used to automatically determine the optimal configuration of hyperparameters, which significantly influence the effectiveness of DRL control agents. Specifically, the Tree-structured Parzen Estimator (TPE) algorithm [54] was selected as the sampling method within Optuna for this optimisation process. In this study, the optimisation of hyperparameters is executed to identify the optimal configuration that provides the most favourable balance between reducing energy consumption and enhancing indoor temperature control for the DRL agent, benchmarking the achieved performance against that of RBC. The hyperparameters subject to optimisation are the learning rate μ, the reward scaling factor θ, the number of hidden layers and the number of neurons per layer. The hyperparameters optimisation is carried out by evaluating the minimisation of energy consumption E_{tabs} and cumulative sum of temperature violations T_{viol} obtained during the testing period $t \in [0, t_N]$ in simulation at the end of the DRL training per each set of hyperparameters. T_{viol} is measured in °C and defined as follows:

$$T_{viol} = \sum_{t=0}^{t_N} b_{occ,t} \cdot T_{viol,t} \quad (12)$$

Given the multi-objective nature of hyperparameter optimisation, it results in multiple Pareto-optimal solutions [55], then a criterion for selecting the best solution among these optimal choices should be established. The criterion adopted in this work refers to the minimum distance from the ideal point [56], which represents the point with coordinates corresponding to the minimum values of both objective function terms. In this context, the distance is computed between the points representing solutions on the Pareto front and the ideal point within a plane defined by coordinates $[E_{tabs}, T_{viol}]$.

A temperature violation $T_{viol,t}$ is computed as the absolute temperature difference between the indoor temperature T_i and the upper $\overline{T_i}$ or lower $\underline{T_i}$ bounds of the temperature acceptability range of [22, 24] °C. This calculation occurs when the indoor temperature falls outside this range during occupancy, represented by $b_{occ,t}$, a Boolean variable equal to 1 when the thermal zone is occupied. $T_{viol,t}$ is computed as defined in the following equation:

$$T_{viol,t} = \begin{cases} \underline{T_i} - T_i & \text{if } T_i < \underline{T_i} \\ 0 & \text{if } \underline{T_i} \le T_i \le \overline{T_i} \\ T_i - \overline{T_i} & \text{if } T_i > \overline{T_i} \end{cases} \quad (13)$$

Table 3
Values and range of DRL controller hyperparameters.

Hyperparameters	Value	Step
Learning rate μ	$[3 \cdot 10^{-4}, 1 \cdot 10^{-3}]$	$1 \cdot 10^{-4}$
Reward scaling factor θ	[3, 15]	2
# Hidden layers	[2, 4]	2
# Neurons per hidden layer	[64, 128]	64
Discount factor γ	0.99	–
Soft-update coefficient τ	$5 \cdot 10^{-3}$	–
Batch size	128	–
Training episodes	30	–

Thirty agents are trained for 30 episodes during the hyperparameters optimisation procedure. Each training episode corresponds to a cooling season and consists of 91 days, from June 1 to August 30, 2022. The Euclidean distance between the DRL performance of each trial at the end of the training phase and the ideal point was computed. Thus, the solution with the minimum distance and the best performance compared to the three developed RBCs in terms of total energy consumption and cumulative sum of temperature violations was selected as the optimal one. Table 3 includes in the first five rows the optimised hyperparameters (i.e., μ, θ, *Number of hidden layers*, *Number of neurons per hidden layer*) with the corresponding range and step value, while the other rows of Table 3 indicates the values of hyperparameters that remain fixed for computational constraints: *Discount factor* γ, *Soft-update coefficient* τ, *Batch size* and *Training episodes*.

4.7. Real-world deployment

This section describes the practical deployment of the control systems in a real-world setting, following the methodology described in Section 3.3.

The controllers considered in this work were implemented in a remote desktop PC specifically configured for this purpose. The desktop PC has a 4-core CPU running at 3.40 GHz and 16 GB of RAM, serving as the control logic's central hub. The controllers operated within a Python virtual environment.

The communication infrastructure between the remote PC and the NEST HiLo unit is depicted in Fig. 6. On the lower level of the infrastructure, the Programmable Logic Controller (PLC) located in HiLo takes care of collecting data from the sensors and sending control signals to the actuators using several protocols, such as Modbus RTU RS485 and standard analog/digital signals (0–10 V, 4–20 mA, PT1000, DI/DO). The PLC communicates via the multiplatform, open-source OPC-UA protocol with a gateway located on a virtual machine. The gateway sends the data collected from the lower level to a MS-SQL historical database using the Open DataBase Connectivity (ODBC). The database is located in the NEST cloud on a virtual machine that can be accessed remotely by a REST API integrated into Python. Real-time data and control signals are exchanged by the remote client to the gateway server using the OPC-UA protocol. Due to the specific control architecture of the NEST units, the control signal needs to be sent with an overhead, including a signal requesting the remote controllability of the system and a square wave watchdog signal that needs to alternate between a true and false state every thirty seconds to maintain the remote control of the system.

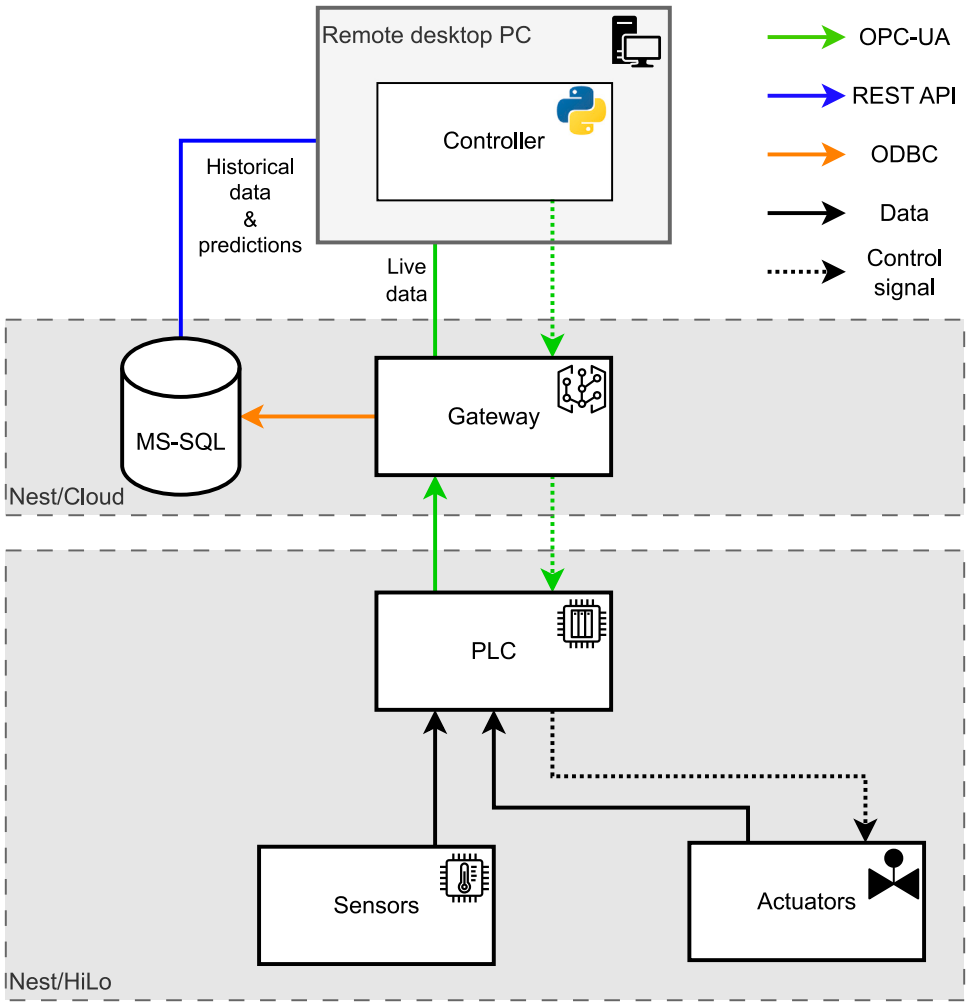


Fig. 6. Physical layout of the control system. The dotted lines represent the control signals, while the solid lines represent the data. The blue colour shows the communication with the OPCUA protocol, and the green depicts the REST API communication.

During the deployment, the RBC ran from June 24, 2023 to June 29, 2023 with the standard operations ventilation control. Next, different RL controllers were deployed, each time after being pre-trained. From July 1, 2023 to July 19, 2023 a DRL controller with the action space and observation space described in Section 3.2 was deployed. After these periods, the supply water temperature of the TABS system has been lowered from 19 °C to 18 °C due to operational adjustments in the system. The decrease in the supply water temperature resulted in an increased cooling power from around 600 W to 750 W. After this change, from July 23, 2023 a newly trained RL controller was implemented in HiLo until the end of the experiments on August 21, 2023. During this phase, the RL controller for TABS had identical specifications to the previous iterations, except the neighbouring room temperature was omitted from the observation space.

The performance of the controllers has been compared by looking at two metrics: the normalised daily cumulative temperature violations $T_{\text{viol,norm},d}$ and the normalised daily mean cooling power $\overline{\dot{Q}}_{\text{tabs,norm},d}$, computed as follows:

$$T_{\text{viol,norm},d} = \eta_d T_{\text{viol},d} \tag{14}$$

$$\overline{\dot{Q}}_{\text{tabs,norm},d} = \eta_d \overline{\dot{Q}}_{\text{tabs},d} \tag{15}$$

where, $T_{\text{viol},d}$ is the daily cumulative sum of temperature violations and $T_{\text{viol},d}$ is daily mean cooling power. Both metrics include a normalisation factor $\eta_d > 0$ to consider the influence of the boundary conditions [57], such as the different outdoor temperatures and solar levels of radiation occurring each day:

$$\eta_d = \frac{\overline{T_{\text{o},d}}}{T_{\text{o},d}} \cdot \frac{\overline{\dot{Q}_{\text{sol},d}}}{\dot{Q}_{\text{sol},d}} \tag{16}$$

where, $\overline{T_{\text{o},d}}$ and $\overline{\dot{Q}_{\text{sol},d}}$ are the average values of the daily outdoor temperatures $T_{\text{o},d}$ and daily solar radiation $\dot{Q}_{\text{sol},d}$, during the considered periods. This is important because these external factors can significantly influence the performance metrics, and not normalising the data might lead to misleading conclusions about the controllers' effectiveness. For example, during a warm summer day d with high solar radiation, the normalisation factor η_d will be smaller than one, so the resulting normalised daily cumulative temperature violations $T_{\text{viol,norm},d}$ and normalised daily mean cooling power $\overline{\dot{Q}}_{\text{tabs,norm},d}$ are reduced. This reduction reflects the fact that the higher outdoor temperature and solar radiation would naturally lead to increased cooling requirements and possibly greater temperature violations, which might not be a direct result of controller performance.

Additionally, the deployment included a live visualisation feature using Grafana. This tool provided real-time monitoring of the system's performance, offering insights into metrics such as temperature trends and energy consumption. The implementation of such real-time data visualisation was crucial for ongoing system evaluation and management, enabling prompt identification and addressing of system irregularities.

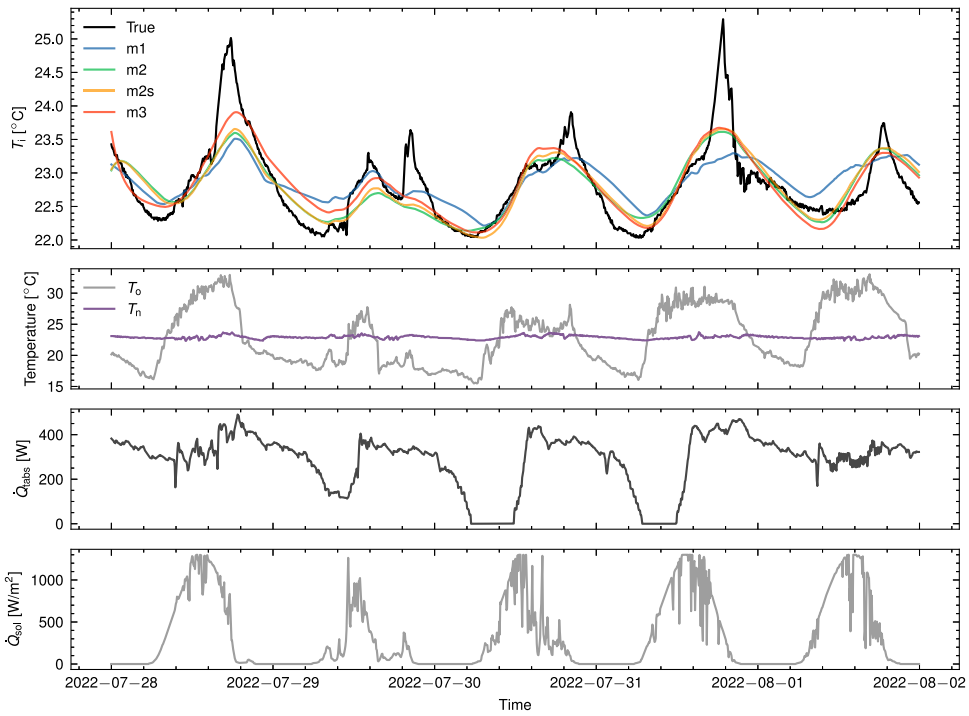


Fig. 7. RC models results after the estimation of the parameters.

Table 4
RC models RMSE in the training and testing datasets.

Model	RMSE train [°C]	RMSE test [°C]
*m*1	0.42	0.72
*m*2	0.35	0.66
*m*2*s*	0.33	0.59
*m*3	0.30	0.62

Table 5
Optimal values of the hyperparameters subject to optimisation for the DRL controller.

Hyperparameter	Value
Learning rate μ	0.0009
Reward scaling factor θ	15
# Hidden layers	4
# Neurons per hidden layer	64

5. Results

This section describes the results following the step indicated in the methodological framework as described in Section 3. The first subsection introduces the results retrieved during the simulation phase for the RC model of the building and the DRL controller. Then, the second subsection includes the outcomings from the real-world implementation of the DRL agent.

5.1. Simulation

Fig. 7 shows the results of the models in a period of the test dataset and the corresponding model inputs as those indicated in Eq. (3). All the models are able to catch the essential dynamics of the office room. Despite the models matching the low-frequency components of the dynamics, they are not able to correctly describe the high-frequency variations of the indoor temperature. In detail, the indoor temperature shows considerable peaks during late afternoons, that are hardly explained by the input data, and they are likely caused by unmeasured disturbances, such as internal heat gains (e.g., appliances and lights).

Table 4 shows the model RMSE during the training and validation period. The *m*2*s* model resulted in the lowest RMSE in the testing dataset and, therefore, it was used in the remaining of our study to train the DRL controller and to perform the simulation to compare the performance of different controllers.

As described in Section 4, during the training phase of the DRL controller, an automated optimisation approach was performed using Optuna [47], employing the criteria of minimum distance from the ideal point. The optimal hyperparameter configuration, identified through this procedure and detailed in Table 5, ensures the best performance in terms of energy consumption and cumulative sum of temperature violations.

Then, the performances of the DRL agent with optimised hyperparameters were compared with those of MPC and of the two PI controllers (i.e., PI_{23}, PI_{24}) and the three RBCs (i.e., $RBC1$, $RBC2$, $RBC3$). As shown in Fig. 8, the DRL controller outperforms all three RBCs by minimising energy consumption and temperature violations throughout the entire cooling season. Specifically, DRL reduces energy consumption by 51%, 12%, and 15% when compared to the three rule-based controllers (from $RBC1$ to $RBC3$), also decreasing the cumulative sum of temperature violations by 20% to 26%. Moreover, DRL consumes 68% more energy than PI_{24}, which however still turns out to be the benchmark controller with the highest amount of temperature violations compared to others (i.e., +67% compared to DRL). On the other hand, the DRL controller manages to improve performance overall by 23% in terms of E_{tabs} and 5% in terms of T_{viol} compared to PI_{23}. However, MPC ensures the best performance as it saves approximately 29% of energy compared to the DRL controller while ensuring comparable indoor temperature control performance (i.e., T_{viol} is just 5% less than DRL).

Furthermore, from the analysis of the results for the three RBCs, it emerges that $RBC3$ ensures the best trade-off between energy consumption and comfortable indoor temperature conditions. Therefore, it was selected as the benchmark for the real implementation phase of the DRL controller. Specifically, $RBC1$ provides superior indoor temperature control compared to the other RBCs but consumes the highest amount of energy, as it can be activated on any day of the week and

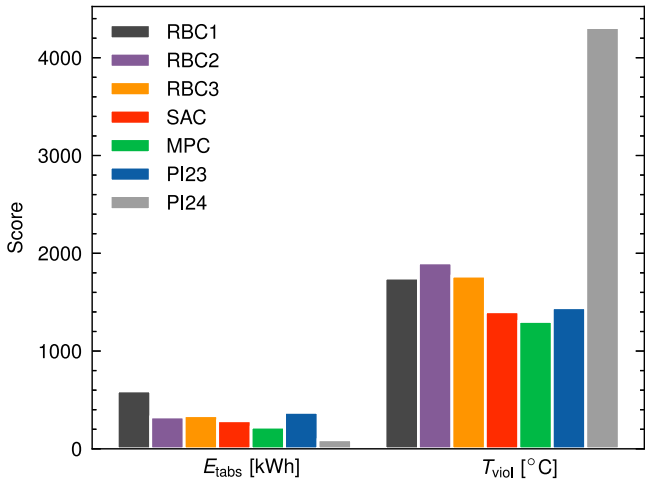


Fig. 8. Comparison of overall performances achieved by the three RBCs, two PI, MPC and DRL controllers during simulation.

at any time. On the other hand, $RBC2$ consumes the least amount of energy but records the highest cumulative sum of temperature violation value since its activation period is limited to occupancy hours (i.e., from 7:00 to 21:00).

Fig. 9 presents a comparative analysis of indoor temperature profiles and energy consumption of $RBC3$, PI_{23}, MPC and DRL controllers. The DRL controller exhibited superior performance in terms of indoor temperature management and reduction of energy consumption by implementing a more balanced energy system management strategy compared to $RBC3$ and PI_{23}. Moreover, although the ideal MPC ensured the best performance, the DRL controller can effectively track the indoor temperature trend obtained by employing the near-optimal policy from the MPC. In detail, the DRL agent controls the valve opening that regulates the flow of thermal fluid to the TABS. This process serves two main purposes, following a similar control policy as MPC: pre-cooling the environment during nighttime and providing cooling energy when needed, especially when occupants are present. The agent responds to increases in indoor temperature, which various external factors like the presence of occupants, high external temperatures, or intense solar radiation can trigger. It ensures efficient cooling even when the indoor temperature is within the acceptable range. In this way, the DRL controller can proactively prevent, or in the most adverse scenario, delay and reduce the peak value of indoor temperature compared to that achieved by the $RBC3$ and PI_{23}. Moreover, the DRL exploits the office thermal inertia through pre-cooling to avoid temperature peaks that especially occur in the late afternoon, while minimising energy consumption, resulting in a more refined control policy than the RBC and PI_{23}. Although on certain days the indoor temperature exceeds the upper-temperature limit $\overline{T_i}$, the DRL effectively minimises the cumulative sum of temperature violations compared to the $RBC3$. To conclude, from the third sub-plot in Fig. 9 emerges the advantage of MPC in terms of energy consumption. In detail, the MPC minimises energy consumption in the pre-cooling phases, only opening the valve when needed during occupancy periods. This different behaviour in the control policy from DRL may be related to the MPC having perfect knowledge of the building's dynamics since the controller model is equivalent to that of emulating the dynamics of the controlled building.

5.2. Deployment

Fig. 10 depicts the connection loss in hours across the experimental period, from June 24, 2023, to August 21, 2023. The data illustrate the temporal distribution and extent of connectivity interruptions encountered by the building management system while implementing various control strategies. The connection losses are due to external factors and are not related to the deployment of the controllers. During the initial deployment phase, labelled $RBCv$, representing the RBC variant, minor connection losses were observed, with only a few instances surpassing the 2.5-hour/day threshold denoted by the dashed line. This threshold represents a significant level of connection loss that may impact the system's operational reliability. The 2.5 h threshold has been chosen after careful consideration of the trade-off between the quality and quantity of the experimental results. Having a bigger threshold increased the number of considered days but lowered the quality of the results (since the control policy reverted to the default RBC for longer times) while lowering the threshold too much limited the considered days. The value of 2.5 h, (i.e. ~10% of 24 h) resulted in the best trade-off.

The $RL600Wnv$ controller shows a similar connectivity pattern, with infrequent connection losses that rarely exceed the 2.5-h threshold. Notably, the $RL750Wn$ controller is marked but was excluded from the analysis. The reason for this is that there was an issue occurred in the ventilation system, which prevented it from working normally. As detailed in Section 4.7, this period encountered an issue with the ventilation system being switched off, making the collected data during this phase non-representative of standard operating conditions. The final period, with the $RL750Wv$ controller, showed a pronounced increase in connection losses, with several instances significantly exceeding the threshold.

Fig. 11 illustrates the normalised daily indoor temperature violations on the x-axis against the normalised daily TABS cooling rate on the y-axis, where the days with too many connection losses (≥2.5 h threshold) have been excluded.

The dashed lines represent the mean values for each variable, effectively dividing the plot into four quadrants, with the intersection representing the mean performance. The quadrant in the bottom left corner is where both temperature violations and TABS heat rate are below average, indicating high performance in temperature control with lower energy consumption. The quadrant in the top right corner is the region where both temperature violations and TABS heat rate are above their respective averages. This indicates a less desirable performance where the temperature is not maintained effectively and more energy is consumed. Controllers in the top left quadrant are characterised by lower-than-average temperature violations but higher-than-average TABS heat rates, suggesting that while they maintain comfortable indoor temperature conditions more effectively, they do so at the cost of higher energy use. The quadrant in the bottom right represents conditions where the temperature violations are above average, but the TABS heat rate is below average. Controllers in this quadrant are less effective in maintaining temperature but do so with less energy usage. The DRL controllers have several points in the bottom left quadrant, which is indicative of a balance between maintaining indoor temperature control and energy efficiency. They achieve better performance with fewer temperature violations and lower TABS heat rates. On the other hand, the RBC controller shows a spread across the quadrants on the right-hand side, indicating less consistency with a more pronounced tendency towards temperature violations, while the energy consumption remains comparable. Specifically, the $RL600W$ controller showed a 69% reduction in daily temperature violations and 9% reduction in the daily TABS heat rate compared to the RBC. The $RL750W$ controller reduced the daily temperature violations by 68% but increased the average daily TABS heat rate. This outcome is likely a consequence of the RL750W controller leveraging the capacity to use a more powerful 750 W system, which would naturally consume more power. Still, the increased energy usage was offset by the controller's ability to maintain temperature conditions more effectively. Both RL controllers significantly enhanced the maintenance of the indoor temperature within the desired temperature bounds, and

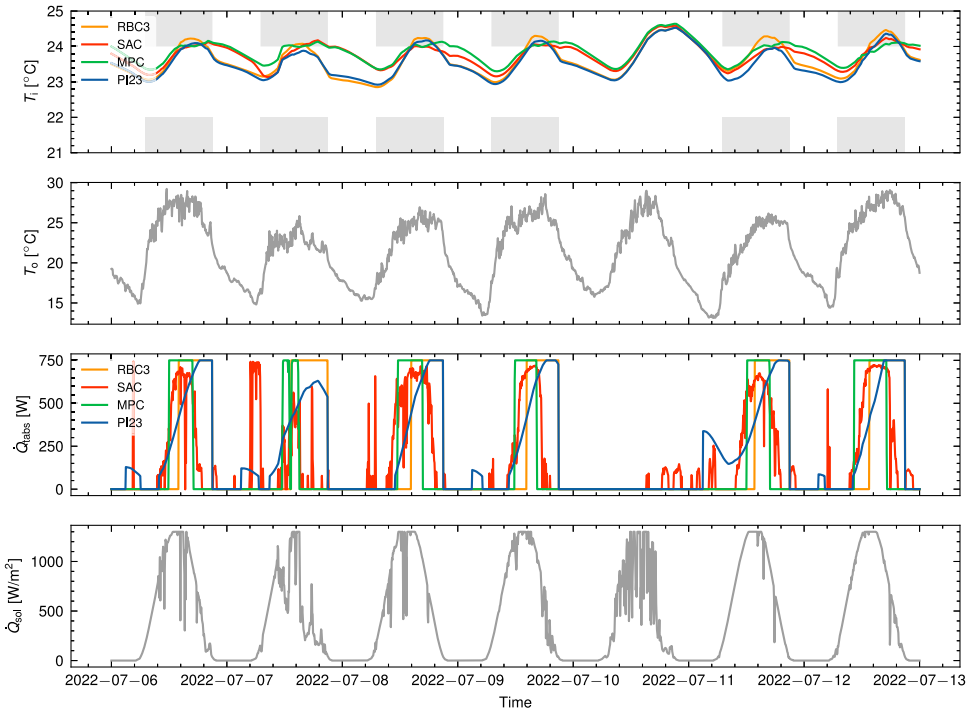


Fig. 9. Simulation comparison of indoor temperature and cooling energy consumption related to the operation of $RBC3$, PI_{23}, MPC and DRL controllers.

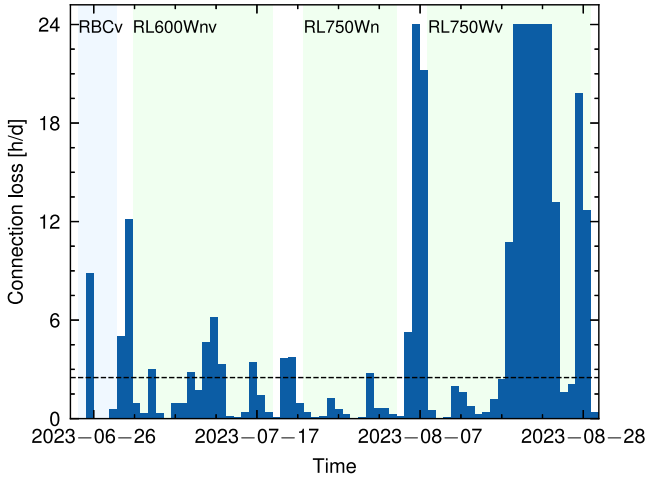


Fig. 10. Connection issues occurred during the deployment period of $RBCv$, $RL600Wnv$, $RL750Wn$ and $RL750Wv$ controllers. v represents that the ventilation system is running and n indicates that T_n is included in the agent observation space.

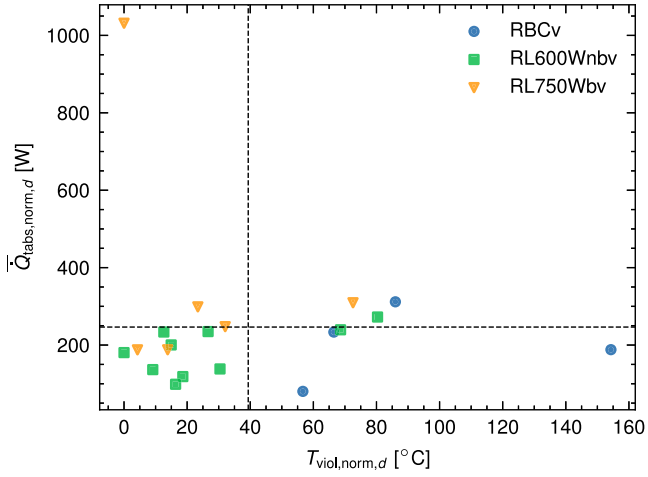


Fig. 11. Performance benchmarking of RBC3 and DRL controllers during the deployment phase. Each point denotes the normalised daily indoor temperature violations against the normalised daily mean cooling rate.

they present a trade-off when it comes to energy consumption. The $RL750W$ controller, using a more powerful system, was less energy efficient, increasing the average heat rate due to its higher capacity. In contrast, the $RL600W$ controller managed to reduce both temperature violations and energy use, pointing to an overall better balance between maintaining comfortable indoor temperature conditions and energy efficiency.

To conclude, Fig. 12 shows the profiles of indoor and outdoor temperatures, real measured cooling power provided by TABS, the corresponding action u_t taken by the agent, and solar radiation measured over three days during the real deployment period of the DRL (from August 15, 2023 to August 17, 2023) where the connection losses were much lower than the threshold (i.e. <1 h/day). Additionally, the upper subplot provides details of the occupied periods (7:00–21:00) and the acceptable temperature range ([22,24] °C). The indoor temperature profile over the analysed days remains around 24 °C (i.e., the upper bound temperature of the acceptability range), aiming to minimise both TABS energy consumption and temperature violations. Towards the end of the occupancy period, the indoor temperature rises by approximately 0.5 °C above the acceptability range. The coexistence of extreme outdoor conditions measured throughout the day (high values for outdoor temperature and solar radiation) and the limited cooling capacity of TABS still led to a slight increase in indoor temperature.

Nevertheless, the DRL effectively controlled the TABS to manage indoor temperatures. It achieved this by keeping the valve almost fully open to ensure the required power to bring the indoor temperature back within the acceptable range. This control behaviour is depicted in the middle subplot, where the green dashed line represents u_t. According to this result, Fig. 12 reveals that in the second half of the day on August 17, when solar radiation and outdoor temperature decrease,

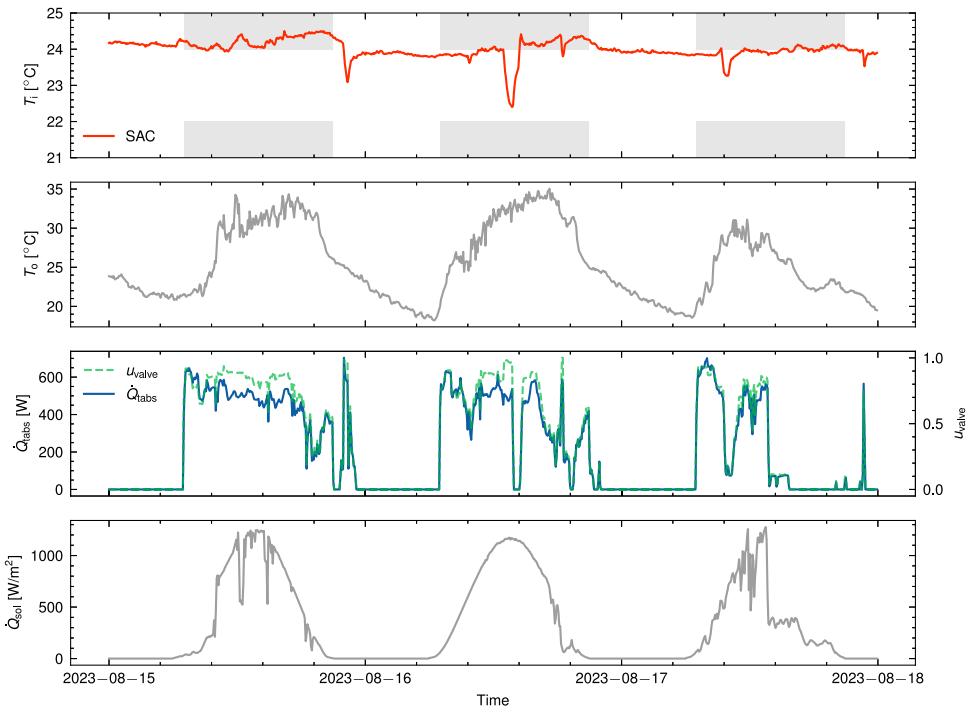


Fig. 12. Indoor and outdoor temperatures, TABS cooling power and solar radiation measured during three summer days (from August 15 to August 17, 2023) of the DRL real-world deployment.

the DRL closes the valve to avoid the cooling energy supply to the office employing TABS. In conclusion, around the middle of the occupancy period on August 16, the indoor temperature dropped sharply as the occupants opened the door to the neighbouring room where the measured temperature was approximately 22 °C. In this case, the DRL was able to adapt to the change in indoor temperature conditions by closing the valve and not supplying unnecessary cooling energy to the office. Therefore, despite the limitations associated with the sizing of TABS and its proper functioning under cooling conditions, the DRL demonstrated excellent capability in optimising the operation of the energy system under these constraints.

6. Discussion

This study explores the real-world deployment of a DRL controller to minimise energy consumption and enhance indoor temperature control by optimising the operation of a TABS in an existing office building.

A novelty of the proposed approach lies in the practical implementation of a DRL-based control agent in an office used as a living lab. The occupants' active interaction with the environment, such as opening doors/windows, introduces real-world variability in indoor temperature conditions.

The DRL controller is easily implementable in the real building, leveraging a minimum number of the existing sensors in the HiLo facility's monitoring infrastructure. Indoor temperature sensors monitor the analysed office zone and adjacent areas, and the outdoor sensor tracks the external conditions, including external temperature and solar radiation. The use of a simplified observation space for the DRL control agent ensures experiment replicability, as the selected variables align with common building monitoring measurements. The DRL agent's control action, implemented on the TABS supply valve through an existing actuator, further enhances interoperability without the need for additional invasive measures.

Although the monitoring infrastructure includes sensors for detecting the presence of occupants, the office is used by different people with different habits, so to ensure the indoor temperature is within the boundaries, only the information on the working hours is used. Therefore, based on historical occupancy data collected, it emerges that the building is occupied predominantly from Monday to Saturday, 8:30–21:00, although in some cases, occupants may arrive early (i.e., 7:00). In this study, the TABS is integrated within an extremely lightweight concrete structure, a novel approach that improves its thermal response compared to traditional TABS [58]. This integration results in faster temperature adjustments than typical TABS due to the reduced thermal mass. However, it is important to note that, while enhanced, this system's responsiveness remains slower compared to more immediate systems like air terminal units (e.g., fan coils). In this context, a conservative approach is employed since the office is considered occupied from Monday to Saturday 7:00–21:00. Furthermore, the formulation of the reward function plays a crucial role in the operation of the DRL agent. Given that our case study refers to an office building, it is chosen not to prioritise the occupants' comfort requirements as more stringent as could be required in other buildings (e.g., hospitals). In future applications, greater importance could be assigned to the temperature term in evaluating how the DRL-based controller adapts its policy to simultaneously minimise energy consumption.

Possible equipment damage or extreme indoor environment conditions limit the widespread DRL controller implementation in real buildings [33]. A fallback safety system is introduced in HiLo to ensure continuous energy system operation, preventing abrupt interruptions in HVAC functionality due to DRL controller failures associated with connection issues. This system is designed to reintroduce the default controller implemented in HiLo, providing a safety barrier for system operation.

Moreover, potential initial instability of the DRL control policy may lead to unacceptable initial performance [33]. Therefore, the direct implementation of the DRL controller in real buildings is avoided. The approach explored in this work involved the use of a simplified RC model during the pre-training phase of the DRL controller. The RC model's simple design facilitated efficient computation, maintaining the accuracy necessary for realistic simulations and enabling rapid and effective training of DRL agents. The parameters of the RC model were derived from real data monitored in the thermal zone used as a case

study. However, the quality and quantity of data influence the physical response of the RC model used as a surrogate model of the building.

The potential use of the RC model as a building simulator for comparing the performance achieved in the building between RBC and DRL controllers could emerge as a limitation to be addressed. This simplified model may not guarantee a physically similar response compared to what would be obtained if the same controller were implemented in reality. Nevertheless, the DRL controller demonstrated robustness to boundary conditions encountered during the real-world deployment phase, ensuring good performance compared to RBC, despite being trained on a simplified model. However, the development of a more accurate building model could be considered in the future to achieve a more straightforward performance benchmark.

Another limitation is associated with the controllers evaluated as a benchmark during the real-world implementation of DRL. Contrary to the simulation study, the performance comparison between the DRL controller, PI controllers and MPC is not carried out due to the limited duration of the real-world investigation, during which some connection issues further reduced the time window for DRL implementation. Additionally, the proposed version of MPC would not have performed as well as in simulation due to unavoidable modelling errors and not perfect predictions. Nevertheless, the objective of this work is to provide insights about the applicability and reliability of DRL in real-world scenarios, despite the performance comparison carried out in the simulation study indicating that DRL controller achieves similar performance as MPC.

In conclusion, the obtained results mirror the potential demonstrated by DRL controllers in simulated building applications, extending their applicability to real-world implementations. While safety adjustments are essential, the developed approach demonstrates the seamless integration of DRL controllers with existing energy and monitoring systems, minimising the need for invasive technical operations. This study contributes valuable insights into the practical implementation of DRL controllers in building environments, bridging the gap between simulation and reality.

7. Conclusion

This paper has presented a comprehensive study of the application of DRL for controlling the thermal dynamics of an office building used as a living lab equipped with TABS. First, a simulation model based on the RC network structure has been formulated and identified on real data collected on a real office building located in Switzerland. The simulation model has been used as the training environment for the SAC controller and to benchmark its performance against different variations of RBC strategies. The simulation study showed that the DRL controller achieved a reduction in energy consumption between 15% and 50% with a 25% decrease in temperature violation compared to RBCs, while ensuring a reduction in energy consumption from TABS of 23% and in temperature violation of 5% compared to a PI controller considering 23 °C as temperature setpoint (i.e., the average value of the [22, 24] °C acceptability range).

Moreover, DRL reach the same performance level in terms of indoor temperature control as an ideal MPC but consuming 29% more energy. Based on these metrics, the best RBC and SAC controller has been implemented in the real office building throughout the cooling season lasting two months. The real-world deployment results revealed that the DRL controller decreased the temperature violation by 68% while, on average, maintaining the same energy consumption as the RBC.

Future work will focus on:

- The development of a more detailed building surrogate model that better emulates the real dynamics of the building compared to the RC model. In this context, a surrogate model developed using Energyplus or Modelica could be employed both to pre-train the DRL controller and used as a high-fidelity environment during the performance benchmarking phase [52].
- An extension of the real-world comparison of the DRL agent performance beyond just RBC. A PI controller and an MPC can be used to evaluate the DRL agent performance in the real testbed.
- The comparison of SAC controller performances with other state-of-the-art DRL algorithms, as Proximal Policy Optimisation (PPO) and Deep Deterministic Policy Gradient (DDPG).
- The evaluation of the proposed methodology for controlling both the TABS and the other systems installed in the office, developing a multi-action DRL controller. This would allow overcoming limitations associated with the use of TABS during the cooling season and ensure faster response to maintain appropriate indoor building temperature conditions.
- The implementation of Transfer Learning (TL) to adapt the DRL controller implemented in the analysed office zone to the other thermal zones included in the whole real building.

CRediT authorship contribution statement

Alberto Silvestri: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Davide Coraci:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Silvio Brandi:** Writing – review & editing, Methodology, Investigation, Formal analysis, Conceptualization. **Alfonso Capozzoli:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **Esther Borkowski:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Johannes Köhler:** Writing – review & editing, Validation, Methodology, Formal analysis. **Duan Wu:** Writing – review & editing, Validation, Methodology. **Melanie N. Zeilinger:** Writing – review & editing, Validation. **Arno Schlueter:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Alberto Silvestri reports financial support was provided by Mitsubishi Electric R&D Centre Europe BV. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgements

The data used in this study were collected and made available by Empa with the support of the Swiss Federal Office of Energy, Switzerland and the Swiss National Science Foundation, Switzerland. The experiments in HiLo were carried out with the support of Sascha Stoller and Reto Fricker.

Funding

This study was partly financed by Mitsubishi Electric R&D Centre Europe B.V. [contract number ETH ID No 16234].

The work of Silvio Brandi is funded by the project NODES which has received funding from the MUR — M4C2 1.5 of PNRR funded by the European Union — NextGenerationEU (Grant agreement no. ECS00000036).

Appendix. Background on reinforcement learning

In the field of RL, the framework of Markov Decision Process (MDP) provides a fundamental and mathematically rigorous approach for modelling decision-making problems under uncertainty. An MDP is defined by the tuple (S, A, P, R, γ) [20]:

1. **States** (S): A finite set of states, denoting all possible situations or configurations in which the system might find itself.
2. **Actions** (A): A finite set of actions available to the decision-maker or agent, determining the possible moves or decisions that can be made in each state.
3. **Transition Probabilities** (P): The state transition probability matrix, where $P(s'|s,a)$ represents the probability of moving from state s after taking action a.
4. **Rewards** (R): The reward function, $R(s,a)$, specifies the immediate reward received after transitioning from state s to state s' due to action a. This function quantifies the benefit (or cost) associated with each action in each state.
5. **Discount Factor** (γ): A discount factor $\gamma \in [0,1]$, used to balance the importance of immediate versus future rewards. It determines the present value of future rewards, with lower values placing more emphasis on immediate rewards.

The goal within an MDP framework is to identify a policy $\pi : S \rightarrow A$ that maximises the expected cumulative reward over time. This entails computing the expected sum of discounted rewards, expressed as $\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]$.

In the context of RL, two fundamental concepts are the state value function and the action value function. The state value function, denoted as $V^{\pi}(s)$, represents the expected cumulative reward for being in a state s and following a particular policy π [20]. It is mathematically expressed as:

$$V^{\pi}(s) = \sum_{a \in A} \pi(a|s) \sum_{s' \in S} P(s'|s,a)\left[R(s,a) + \gamma V^{\pi}(s')\right] \tag{A.1}$$

On the other hand, the action value function, denoted as $Q^{\pi}(s,a)$, estimates the expected cumulative reward of taking an action a in state s and then following the policy π [20]. This is given by:

$$Q^{\pi}(s,a) = \sum_{s' \in S} P(s'|s,a)\left[R(s,a) + \gamma \sum_{a' \in A} \pi(a'|s') Q^{\pi}(s',a')\right] \tag{A.2}$$

In the domain of RL, Q-learning emerges as an algorithm for solving MDPs without requiring a model of the environment. Central to Q-learning is the approximation of the optimal Q-function, $Q^*(s,a)$, which represents the expected cumulative reward starting from state s, taking action s, and thereafter following the optimal policy π^*. The Q-function in Q-learning is iteratively updated using a sample-based approach rather than a predefined policy π. The update rule, is given by [59]:

$$Q(s,a) \leftarrow Q(s,a) + \mu\left[R(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a)\right] \tag{A.3}$$

where the term $\max_{a'} Q(s',a')$ represents the maximum expected future reward obtainable from the next state s'.

In this equation, $\mu \in [0,1]$ is the learning rate, a factor determining the extent to which new information overrides old information. An RL agent implementing a learning rate μ equal to 0 does not learn anything and does not use new knowledge to update the control policy. Conversely, setting the learning rate μ to 1 in a Q-learning algorithm significantly alters how new information is incorporated. This means that the algorithm puts full weight on the most recent information, updating the Q-value based solely on the latest reward and the estimated maximum future Q-value. This approach can be effective in certain scenarios, particularly where adapting quickly to new information is crucial. However, it may also lead to instability or failure to converge if the environment is noisy or the latest information is not always reliable. Typically, a balance is sought where the learning rate is set to a value that allows the algorithm to learn from new information while retaining some of the previous knowledge. This process of iteratively updating the Q-values based on the Bellman equation refines the strategy towards the optimal policy.

However, a significant limitation arises in Q-learning when dealing with environments with large state or action spaces. In such scenarios, the tabular representation of the Q-function becomes impractical due to the exponential growth in the number of state–action pairs, leading to issues of scalability and memory requirements [60]. This challenge is particularly pronounced in real-world problems, where states can be continuous or high-dimensional. To address this, the concept of function approximation is introduced, where a parameterised function, often a neural network, is used to estimate the Q-values. The integration of neural networks in Q-learning culminates in the development of Deep Q-Network (DQN) [20]. DQN leverages deep learning to approximate the Q-function, enabling the handling of high-dimensional state spaces that are infeasible for tabular methods. This is expressed as:

$$Q(s,a;\theta) \approx Q^*(s,a) \tag{A.4}$$

In this work, it was implemented the SAC, an advanced DRL algorithm. SAC [61] is an off-policy actor-critic algorithm that optimises a stochastic policy in an entropy-regularised reinforcement learning framework. The key feature of SAC is its objective to maximise both the expected return and the entropy, which is a measure of randomness in the policy [62]. This dual objective is formulated as:

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t,a_t) + \alpha \mathcal{H}(\pi(\cdot|s_t))\right)\right] \tag{A.5}$$

where $\mathcal{H}(\pi(\cdot|s))$ denotes the entropy of the policy π at state s, and α is a temperature parameter that determines the relative importance of the entropy term against the reward. In conventional reinforcement learning algorithms, α is set equal to 0. The Shannon entropy term $\mathcal{H}$ quantifies the degree of uncertainty or randomness in the agent's action selection process. It reflects how much the agent explores different actions in a given state, with higher values of $\mathcal{H}$ indicating a greater propensity for exploration through random action choices. Essentially, Shannon entropy serves as a measure of the unpredictability or variability in the agent's behaviour, promoting exploration in the learning process. This definition of the target function guarantees the appropriate trade-off between exploitation and exploration: it avoids the agent returning sub-optimal control policies and it ensures that the agent is explicitly pushed towards the exploration of new policies.

SAC demonstrates several advantages over DQN, particularly in handling continuous action spaces and in its sample efficiency. While DQN is well-suited for discrete action spaces, SAC's ability to operate in continuous domains makes it a more versatile choice for a broader range of applications.

The architecture of SAC is characterised by its actor-critic approach. It maintains two separate networks:

- The *Actor Network*, which proposes a policy, represented as $\pi(a|s;\xi)$, mapping states to actions. Here, ξ are the parameters of the actor network.
- The *Critic Network*, which evaluates the proposed actions by estimating the action-value function $Q(s,a;\theta)$, with θ being the critic network parameters.

The critic network is trained to minimise the Bellman error [59], and the actor network is updated to maximise the expected return plus entropy, ensuring a balance between exploration and exploitation. This actor-critic framework enables SAC to learn effectively in complex environments, making it a robust choice for tasks that require decision-making under uncertainty.

References

[1] Nweye K, Liu B, Stone P, Nagy Z. Real-world challenges for multi-agent reinforcement learning in grid-interactive buildings. Energy AI 2022;10:100202. http://dx.doi.org/10.1016/j.egyai.2022.100202, URL https://www.sciencedirect.com/science/article/pii/S2666546822000489.

[2] Piscitelli MS, Brandi S, Capozzoli A, Xiao F. A data analytics-based tool for the detection and diagnosis of anomalous daily energy patterns in buildings. Build Simul 2021;14(1):131–47. http://dx.doi.org/10.1007/s12273-020-0650-1.

[3] Coraci D, Brandi S, Piscitelli MS, Capozzoli A. Online implementation of a soft actor-critic agent to enhance indoor temperature control and energy efficiency in buildings. Energies 2021;14(4). http://dx.doi.org/10.3390/en14040997.

[4] Martinopoulos G, Papakostas KT, Papadopoulos AM. A comparative review of heating systems in EU countries, based on efficiency and fuel cost. Renew Sustain Energy Rev 2018;90:687–99. http://dx.doi.org/10.1016/j.rser.2018.03.060.

[5] Dorokhova M, Ballif C, Wyrsch N. Rule-based scheduling of air conditioning using occupancy forecasting. Energy AI 2020;2:100022. http://dx.doi.org/10.1016/j.egyai.2020.100022, URL https://www.sciencedirect.com/science/article/pii/S2666546820300227.

[6] ASHRAE G. 36: High performance sequences of operation for HVAC systems. Atlanta: American Society of Heating, Refrigerating and Air-Conditioning Engineers; 2021, URL https://www.ashrae.org/news/ashraejournal/guideline-36-2021-what-s-new-and-why-it-s-important.

[7] Finck C, Beagon P, Clauß J, Péan T, Vogler-Finck P, Zhang K, Kazmi H. Review of applied and tested control possibilities for energy flexibility in buildings. Technical report from IEA EBC annex 67 - energy flexible buildings, 2017, p. 1–59. http://dx.doi.org/10.13140/RG.2.2.28740.73609.

[8] Salsbury TI. A survey of control technologies in the building automation industry. IFAC Proc Vol 2005;38(1):90–100. http://dx.doi.org/10.3182/20050703-6-CZ-1902.01397, 16th IFAC World Congress.

[9] Wang Z, Hong T. Reinforcement learning for building controls: The opportunities and challenges. Appl Energy 2020;269:115036. http://dx.doi.org/10.1016/j.apenergy.2020.115036.

[10] Fan C, Xiao F, Yan C. A framework for knowledge discovery in massive building automation data and its application in building diagnostics. Autom Constr 2015;50:81–90. http://dx.doi.org/10.1016/j.autcon.2014.12.006.

[11] Miller C, Nagy Z, Schlueter A. Automated daily pattern filtering of measured building performance data. Autom Constr 2015;49:1–17. http://dx.doi.org/10.1016/j.autcon.2014.09.004.

[12] Naidu DS, Rieger CG. Advanced control strategies for heating, ventilation, air-conditioning, and refrigeration systems—An overview: Part I: Hard control. HVAC&R Res 2011;17(1):2–21. http://dx.doi.org/10.1080/10789669.2011.540942.

[13] Stoffel P, Henkel P, Rätz M, Kümpel A, Müller D. Safe operation of online learning data driven model predictive control of building energy systems. Energy AI 2023;14:100296. http://dx.doi.org/10.1016/j.egyai.2023.100296, URL https://www.sciencedirect.com/science/article/pii/S266654682300068X.

[14] Serale G, Fiorentini M, Capozzoli A, Bernardini D, Bemporad A. Model predictive control (MPC) for enhancing building and HVAC system energy efficiency: Problem formulation, applications and opportunities. Energies 2018;11(3). http://dx.doi.org/10.3390/en11030631.

[15] Drgoňa J, Arroyo J, Cupeiro Figueroa I, Blum D, Arendt K, Kim D, Ollé EP, Oravec J, Wetter M, Vrabie DL, Helsen L. All you need to know about model predictive control for buildings. Annu Rev Control 2020. http://dx.doi.org/10.1016/j.arcontrol.2020.09.001.

[16] Cho S, Zaheer-uddin M. Predictive control of intermittently operated radiant floor heating systems. Energy Convers Manage 2003;44(8):1333–42. http://dx.doi.org/10.1016/S0196-8904(02)00116-4.

[17] Wang H, Bo S, Zhu C, Hua P, Xie Z, Xu C, Wang T, Li X, Wang H, Lahdelma R, Granlund K, Teppo E. A zoned group control of indoor temperature based on MPC for a space heating building. Energy Convers Manage 2023;290:117196. http://dx.doi.org/10.1016/j.enconman.2023.117196, URL https://www.sciencedirect.com/science/article/pii/S0196890423005423.

[18] Prívara S, Široký J, Ferkl L, Cigler J. Model predictive control of a building heating system: The first experience. Energy Build 2011;43(2):564–72. http://dx.doi.org/10.1016/j.enbuild.2010.10.022.

[19] Kontes GD, Giannakis GI, Sánchez V, De Agustin-Camacho P, Romero-Amorrortu A, Panagiotidou N, Rovas DV, Steiger S, Mutschler C, Gruen G. Simulation-based evaluation and optimization of control strategies in buildings. Energies 2018;11(12). http://dx.doi.org/10.3390/en11123376.

[20] Sutton RS, Barto AG. Reinforcement learning: an introduction. 2nd ed.. The MIT Press; 2018, URL http://incompleteideas.net/book/the-book-2nd.html.

[21] Brandi S, Piscitelli MS, Martellacci M, Capozzoli A. Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings. Energy Build 2020;224:110225. http://dx.doi.org/10.1016/j.enbuild.2020.110225.

[22] Yang L, Nagy Z, Goffin P, Schlueter A. Reinforcement learning for optimal control of low exergy buildings. Appl Energy 2015;156:577–86. http://dx.doi.org/10.1016/j.apenergy.2015.07.050.

[23] 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. http://dx.doi.org/10.1016/j.applthermaleng.2023.120430, URL https://www.sciencedirect.com/science/article/pii/S1359431123004593.

[24] Kathirgamanathan A, Mangina E, Finn DP. Development of a soft actor critic deep reinforcement learning approach for harnessing energy flexibility in a large office building. Energy AI 2021;5:100101. http://dx.doi.org/10.1016/j.egyai.2021.100101, URL https://www.sciencedirect.com/science/article/pii/S2666546821000537.

[25] Zhang Z, Chong A, Pan Y, Zhang C, Lam KP. Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning. Energy Build 2019;199:472–90. http://dx.doi.org/10.1016/j.enbuild.2019.07.029.

[26] Schreiber T, Eschweiler S, Baranski M, Müller D. Application of two promising reinforcement learning algorithms for load shifting in a cooling supply system. Energy Build 2020;229:110490. http://dx.doi.org/10.1016/j.enbuild.2020.110490.

[27] Du Y, Zandi H, Kotevska O, Kurte K, Munk J, Amasyali K, Mckee E, Li F. Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning. Appl Energy 2021;281:116117. http://dx.doi.org/10.1016/j.apenergy.2020.116117.

[28] Vázquez-Canteli JR, Ulyanin S, Kämpf J, Nagy Z. Fusing TensorFlow with building energy simulation for intelligent energy management in smart cities. Sustainable Cities Soc 2019;45:243–57. http://dx.doi.org/10.1016/j.scs.2018.11.021.

[29] Coraci D, Brandi S, Hong T, Capozzoli A. An innovative heterogeneous transfer learning framework to enhance the scalability of deep reinforcement learning controllers in buildings with integrated energy systems. Build Simul 2024;1–32. http://dx.doi.org/10.1007/s12273-024-1109-6.

[30] Brandi S, Fiorentini M, Capozzoli A. Comparison of online and offline deep reinforcement learning with model predictive control for thermal energy management. Autom Constr 2022;135:104128. http://dx.doi.org/10.1016/j.autcon.2022.104128.

[31] Wang X, Kang X, An J, Chen H, Yan D. Reinforcement learning approach for optimal control of ice-based thermal energy storage (TES) systems in commercial buildings. Energy Build 2023;301:113696. http://dx.doi.org/10.1016/j.enbuild.2023.113696, URL https://www.sciencedirect.com/science/article/pii/S037877882300926X.

[32] Hussain A, Musilek P. Energy management of buildings with energy storage and solar photovoltaic: A diversity in experience approach for deep reinforcement learning agents. Energy AI 2024;15:100313. http://dx.doi.org/10.1016/j.egyai.2023.100313, URL https://www.sciencedirect.com/science/article/pii/S266654682300085X.

[33] Nagy Z, Henze G, Dey S, Arroyo J, Helsen L, Zhang X, Chen B, Amasyali K, Kurte K, Zamzam A, Zandi H, Drgoňa J, Quintana M, McCullogh S, Park JY, Li H, Hong T, Brandi S, Pinto G, Capozzoli A, Vrabie D, Bergés M, Nweye K, Marzullo T, Bernstein A. Ten questions concerning reinforcement learning for building energy management. Build Environ 2023;241:110435. http://dx.doi.org/10.1016/j.buildenv.2023.110435, URL https://www.sciencedirect.com/science/article/pii/S0360132323004626.

[34] Zhang Z, Lam KP. Practical implementation and evaluation of deep reinforcement learning control for a radiant heating system. In: Proceedings of the 5th conference on systems for built environments. BuildSys '18, New York, NY, USA: Association for Computing Machinery; 2018, p. 148–57. http://dx.doi.org/10.1145/3276774.3276775.

[35] Blad C, Bøgh S, Kallesøe C, Raftery P. A laboratory test of an offline-trained multi-agent reinforcement learning algorithm for heating systems. Appl Energy 2023;337:120807. http://dx.doi.org/10.1016/j.apenergy.2023.120807, URL https://www.sciencedirect.com/science/article/pii/S030626192300171X.

[36] Lei Y, Zhan S, Ono E, Peng Y, Zhang Z, Hasama T, Chong A. A practical deep reinforcement learning framework for multivariate occupant-centric control in buildings. Appl Energy 2022;324:119742. http://dx.doi.org/10.1016/j.apenergy.2022.119742, URL https://www.sciencedirect.com/science/article/pii/S0306261922010297.

[37] Silvestri A, Coraci D, Wu D, Borkowski E, Schlueter A. Comparison of two deep reinforcement learning algorithms towards an optimal policy for smart building thermal control. J Phys Conf Ser 2023;2600(7):072011. http://dx.doi.org/10.1088/1742-6596/2600/7/072011.

[38] Heidari A, Khovalyg D. DeepValve: Development and experimental testing of a reinforcement learning control framework for occupant-centric heating in offices. Eng Appl Artif Intell 2023;123:106310. http://dx.doi.org/10.1016/j.engappai.2023.106310, URL https://www.sciencedirect.com/science/article/pii/S0952197623004943.

[39] Crawley DB, Lawrie LK, Winkelmann FC, Buhl W, Huang Y, Pedersen CO, Strand RK, Liesen RJ, Fisher DE, Witte MJ, Glazer J. EnergyPlus: creating a new-generation building energy simulation program. Energy Build 2001;33(4):319–31. http://dx.doi.org/10.1016/S0378-7788(00)00114-6, Special Issue: BUILDING SIMULATION'99.

[40] Di Natale L, Svetozarevic B, Heer P, Jones C. Physically consistent neural networks for building thermal modeling: Theory and analysis. Appl Energy 2022;325:119806. http://dx.doi.org/10.1016/j.apenergy.2022.119806, URL https://www.sciencedirect.com/science/article/pii/S0306261922010819.

[41] Richner P, Heer P, Largo R, Marchesi E, Zimmermann M. NEST – A platform for the acceleration of innovation in buildings. Inf Constr 2018;69(548):222. http://dx.doi.org/10.3989/id.55380, URL http://informesdelaconstruccion.revistas.csic.es/index.php/informesdelaconstruccion/article/view/5879.

[42] Block P, Schlueter A, Veenendaal D, Bakker J, Begle M, Hischier I, Hofer J, Jayathissa P, Maxwell I, Echenagucia TM, Nagy Z, Pigram D, Svetozarevic B, Torsing R, Verbeek J, Willmann A, Lydon GP. NEST HiLo: Investigating lightweight construction and adaptive energy systems. J Build Eng 2017;12:332–41. http://dx.doi.org/10.1016/j.jobe.2017.06.013, URL https://www.sciencedirect.com/science/article/pii/S235271021730342X.

[43] Amara F, Agbossou K, Cardenas A, Dubé Y, Kelouwani S. Comparison and simulation of building thermal models for effective energy management. Smart Grid Renew Energy 2015;06(04):95–112. http://dx.doi.org/10.4236/sgre.2015.64009, URL http://www.scirp.org/journal/doi.aspx?DOI=10.4236/sgre.2015.64009.

[44] Raffin A, Hill A, Gleave A, Kanervisto A, Ernestus M, Dormann N. Stable-Baselines3: Reliable reinforcement learning implementations. J Mach Learn Res 2021;22(268):1–8, URL http://jmlr.org/papers/v22/20-1364.html.

[45] Haarnoja T, Zhou A, Hartikainen K, Tucker G, Ha S, Tan J, Kumar V, Zhu H, Gupta A, Abbeel P, Levine S. Soft actor-critic algorithms and applications. 2019, arXiv:1812.05905.

[46] Mock JW, Muknahallipatna SS. A comparison of PPO, TD3 and SAC reinforcement algorithms for quadruped walking gait generation. J Intell Learn Syst Appl 2023;15:36–56. http://dx.doi.org/10.4236/jilsa.2023.151003.

[47] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. KDD '19, New York, NY, USA: Association for Computing Machinery; 2019, p. 2623–31. http://dx.doi.org/10.1145/3292500.3330701.

[48] Coraci D, Brandi S, Hong T, Capozzoli A. Online transfer learning strategy for enhancing the scalability and deployment of deep reinforcement learning control in smart buildings. Appl Energy 2023;333:120598. http://dx.doi.org/10.1016/j.apenergy.2022.120598.

[49] Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, Zaremba W. OpenAI gym. 2016, arXiv:1606.01540.

[50] Visioli A. Modified anti-windup scheme for PID controllers. Control Theory Appl IEE Proc 2003;150:49–54. http://dx.doi.org/10.1049/ip-cta:20020769.

[51] Brandi S, Gallo A, Capozzoli A. A predictive and adaptive control strategy to optimize the management of integrated energy systems in buildings. Energy Rep 2022;8:1550–67. http://dx.doi.org/10.1016/j.egyr.2021.12.058.

[52] Coraci D, Brandi S, Capozzoli A. Effective pre-training of a deep reinforcement learning agent by means of long short-term memory models for thermal energy management in buildings. Energy Convers Manage 2023;291:117303. http://dx.doi.org/10.1016/j.enconman.2023.117303, URL https://www.sciencedirect.com/science/article/pii/S0196890423006490.

[53] Di Natale L, Svetozarevic B, Heer P, Jones CN. Near-optimal deep reinforcement learning policies from data for zone temperature control. In: 2022 IEEE 17th international conference on control & automation. ICCA, IEEE; 2022, p. 698–703. http://dx.doi.org/10.1109/ICCA54724.2022.9831914.

[54] Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. In: Proceedings of the 24th international conference on neural information processing systems. NIPS '11, Red Hook, NY, USA: Curran Associates Inc.; 2011, p. 2546–54, URL https://proceedings.neurips.cc/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

[55] Xin Q. 3 - optimization techniques in diesel engine system design. In: Xin Q, editor. Diesel engine system design. Woodhead Publishing; 2013, p. 203–96. http://dx.doi.org/10.1533/9780857090836.1.203.

[56] Zelany M. A concept of compromise solutions and the method of the displaced ideal. Comput Oper Res 1974;1(3):479–96. http://dx.doi.org/10.1016/0305-0548(74)90064-1.

[57] Coraci D, Brandi S, Capozzoli A. Effective pre-training of a DRL agent by means of LSTM models for thermal energy management in buildings. In: Proceedings of 17th conference on sustainable development of energy, water and environment systems (SDEWES) - paphos (Cyprus). 2022.

[58] Lydon G, Schlueter A. Small-scale experiments on the operational performance of a lightweight thermally active building system. J Build Eng 2023;78:107372. http://dx.doi.org/10.1016/j.jobe.2023.107372, URL https://www.sciencedirect.com/science/article/pii/S2352710223015528.

[59] Bellman R. Dynamic programming. Science 1966;153(3731):34–7. http://dx.doi.org/10.1126/science.153.3731.34, arXiv:https://science.sciencemag.org/content/153/3731/34.full.pdf.

[60] Pinto G, Deltetto D, Capozzoli A. Data-driven district energy management with surrogate models and deep reinforcement learning. Appl Energy 2021;304:117642. http://dx.doi.org/10.1016/j.apenergy.2021.117642.

[61] Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. 2018, http://dx.doi.org/10.48550/ARXIV.1801.01290, URL https://arxiv.org/abs/1801.01290.

[62] Pinto G, Piscitelli MS, Vázquez-Canteli JR, Nagy Z, Capozzoli A. Coordinated energy management for a cluster of buildings through deep reinforcement learning. Energy 2021;229:120725. http://dx.doi.org/10.1016/j.energy.2021.120725.