Deep reinforcement learning-based control of thermal energy storage for university classrooms: Co-Simulation with TRNSYS-Python and transfer learning across operational

Energy Reports 14 (2025) 1349–1367

Contents lists available at ScienceDirect

Energy Reports

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

Research paper

Deep reinforcement learning-based control of thermal energy storage for university classrooms: Co-Simulation with TRNSYS-Python and transfer learning across operational scenarios

Giacomo Buscemi [a],*, Francesco Paolo Cuomo [b], Giuseppe Razzano [a], Francesco Liberato Cappiello [b], Silvio Brandi [a]

[a] Department of Energy "Galileo Ferraris", TEBE Research Group, BAEDA Lab, Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin, 10129, Italy
[b] Department of Industrial Engineering University of Naples Federico II, Via Nuova Agnano, 30-38, Naples, 80125, Italy

ARTICLE INFO

Keywords:
Energy efficiency
AHU systems
Deep reinforcement learning control
HVAC systems
Non-residential building modeling

ABSTRACT

Advanced controllers leveraging predictive and adaptive methods play a crucial role in optimizing building energy management and enhancing flexibility for maximizing the use of on-site renewable energy sources. This study investigates the application of Deep Reinforcement Learning (DRL) algorithms for optimizing the operation of an Air Handling Unit (AHU) system coupled with an inertial Thermal Energy Storage (TES) tank, serving educational spaces at the *Politecnico di Torino* campus. The TES supports both cooling in summer and heating in winter through the AHU's thermal exchange coils, while proportional control mechanisms regulate downstream air conditions to ensure indoor comfort. The building's electric energy demand is partially met by a 47 kW photovoltaic (PV) field and by power grid.

To evaluate the performance of DRL controllers, a co-simulation framework integrating TRNSYS 18 with Python was developed. This setup enables a comparative assessment between a baseline Rule-Based (RB) control approach, and a Soft Actor-Critic (SAC) Reinforcement Learning (RL) controller. The DRL controllers are trained to minimize heat pump (HP) energy consumption, strategically aligning operations with PV availability and electricity price fluctuations, while enforcing safety constraints to prevent temperature violations in the TES.

Simulation results demonstrate that the DRL approach achieved a 4.73 MWh reduction in annual primary energy consumption and a 3.2% decrease in operating costs compared to RB control, with electricity cost savings reaching 5.8%. To evaluate the controller's generalizability, zero-shot transfer learning was employed to deploy the pre-trained DRL agent across different climatic conditions, tariff structures, and system fault scenarios without retraining. In comparison with the RB control, the transfer learning results demonstrated high adaptability, with electricity cost reductions ranging from 11.2% to 24.5% and gas consumption savings between 7.1% and 69.7% across diverse operating conditions.

1. Introduction

Buildings are responsible for a substantial share of global energy use, accounting for about 30% of total consumption and 26% of energy-related emissions (International Energy Agency, 2023). In the European Union this figure reaches roughly 40% (G. u. d. Europa, 2019), underscoring the critical need for improved energy management in the built environment. Consequently, extensive research has explored strategies to reduce building energy demand, ranging from physical retrofits to smarter operational control. Advanced control strategies have demonstrated significant potential in improving energy efficiency, reducing operational costs, and maintaining occupant comfort through more precise and adaptive regulation of building systems. For example, Cappiello (2024) show that the optimization of control strategies and HVAC system management can lead to substantial reductions in both primary energy consumption and operational costs. In a related study, Pavirani et al. (2024) examine the influence of time of use pricing, where energy costs vary across different periods, and highlight the potential of HVAC optimization not only to reduce energy costs but also to improve overall efficiency while providing valuable services to the electrical grid.

Nevertheless, as noted by Wei et al. (2017), most buildings continue to rely on static rule based control strategies for HVAC operation. These traditional approaches are typically designed for fixed reference conditions and lack the adaptability to respond to dynamic factors

* Corresponding author.
E-mail address: giacomo.buscemi@polito.it (G. Buscemi).

https://doi.org/10.1016/j.egyr.2025.07.003
Received 25 February 2025; Received in revised form 16 June 2025; Accepted 9 July 2025
Available online 25 July 2025

such as changing occupancy, weather variability, or the intermittent nature of renewable energy sources. According to Wei et al. (2017), this rigidity often prevents buildings from exploiting their thermal inertia, for instance by preheating or precooling spaces during favorable conditions.

To overcome the limitations of RB control, researchers have further developed more advanced control approaches. Model predictive control (MPC) and other optimization-based strategies use models of building thermal behavior and forecasts of disturbances (weather, occupancy, electricity prices) to determine optimal HVAC setpoints in real time. Such approaches can balance multiple objectives (energy cost, comfort, indoor air quality) simultaneously. For example, Lu et al. (2023) showed that predictive controllers can improve HVAC efficiency by anticipating system dynamics and disturbances, and Jang et al. (2021) successfully combined neural-network-based models with optimization algorithms to minimize energy costs while maintaining practical deployability.

Multi-criteria control algorithms have also been explored to balance competing objectives such as comfort and energy efficiency. For instance, Krinidis et al. (2018) propose a real-time optimization framework that simultaneously considers occupant comfort and energy consumption, demonstrating substantial energy savings without compromising comfort levels.

Despite their potential, model-based approaches such as Model Predictive Control (MPC) present notable implementation challenges. These methods rely on accurate mathematical models of both the building and its environment, which often require significant domain expertise to be developed. Even within a single building, variations in occupancy patterns, internal heat gains, and operational conditions can undermine model accuracy over time. As reported in several studies, the practical deployment of MPC typically involves extensive data collection, high computational demand, and specialized knowledge for model calibration, validation, and ongoing adaptation.

These practical limitations often lead to high implementation costs and constrain the scalability of purely model-based control solutions. As a result, recent research has increasingly focused on machine learning-based approaches, particularly reinforcement learning (RL), as a promising alternative for adaptive HVAC control without the need for manually constructed physical models. RL provides a model-free framework in which an agent learns optimal control policies through trial and error by interacting directly with the environment (Sutton and Barto, 1998).

During this interaction, the RL agent observes the current system state – such as zone temperatures, real-time energy prices, and occupancy levels – and selects control actions accordingly, such as adjusting HVAC setpoints or regulating thermal storage operations. Over time, and across multiple learning episodes, the agent refines its policy to maximize a predefined reward function that typically balances objectives such as energy efficiency, cost reduction, and occupant comfort.

Notably, the advent of Deep Reinforcement Learning (DRL) developed by François-Lavet et al. (2018), combining RL algorithms with deep neural networks, has enabled agents to handle the high-dimensional, non-linear control spaces characteristic of modern building systems. Several studies have demonstrated that DRL-based HVAC controllers can outperform traditional RB control and even model-predictive controllers in terms of reducing energy cost and maintaining occupant comfort. For instance, authors in Razzano et al. (2025), Lu et al. (2023), Fu et al. (2023), Du et al. (2021) have reported significant HVAC energy cost reductions and improved comfort level adherence using deep Q-networks, policy gradient methods, and other DRL algorithms in building climate control tasks. DRL's ability to continuously adapt its strategy based on feedback makes it particularly well-suited to complex, dynamic systems such as HVAC. Significantly, DRL techniques have been extended to integrated thermal energy storage (TES) control, recognizing that thermal storage can substantially increase building energy flexibility. Early studies, such as Henze and Schoenmann (2003), demonstrated the potential of reinforcement learning (RL) for optimizing thermal energy storage (TES) operations in buildings, showing that learning-based controllers can effectively adapt to varying system conditions and contribute to improved energy performance. More recent research, including Wang et al. (2023), has applied deep reinforcement learning (DRL) techniques to more complex TES applications, successfully managing operational decisions in real-world HVAC systems and enhancing demand-side flexibility. Complementary work by Diller et al. (2024) has employed methods such as dynamic programming to address TES control, emphasizing the importance of reduced-order models in achieving computationally efficient optimization.

Collectively, these contributions highlight the growing potential of integrating advanced control techniques with thermal storage systems. In particular, the ability of DRL-based controllers to learn effective charge and discharge strategies enables load shifting in response to variable electricity prices or the availability of on-site renewable energy, such as photovoltaic generation. This integration supports improved energy cost management, greater self-consumption of renewables, and increased operational flexibility within building energy systems.

The development and evaluation of advanced HVAC control strategies increasingly rely on co-simulation platforms that combine detailed building simulation tools with flexible, Python-based algorithmic environments. High-fidelity simulation engines such as EnergyPlus and TRNSYS are used to model the thermal and operational dynamics of buildings with a high degree of realism. These platforms are commonly integrated with Python to enable the implementation of custom control logic and learning algorithms. This co-simulation approach allows researchers to test control strategies extensively under realistic yet controlled conditions prior to real-world deployment.

Several studies have demonstrated the effectiveness of this methodology. For example, Chen et al. (2023) developed a TRNSYS-Python co-simulation framework to train a deep Q-network agent for climate control in office buildings, achieving superior performance compared to a conventional rule-based approach in both energy efficiency and comfort. Integration between TRNSYS and Python is often facilitated by components such as the Type 3157 module, which enables bidirectional data exchange between the simulation engine and external Python scripts (Team, 2023). This setup has been widely adopted in recent research on intelligent HVAC control. For instance, Feng and Nekouei (2024) employed this architecture to implement a privacy-preserving, cloud-based control scheme, while Adesanya et al. (2024) applied a similar framework to optimize PID settings in a greenhouse HVAC system using deep reinforcement learning, leading to improved temperature regulation and energy savings. In this context, this study investigates the application of DRL to optimize the operation of an AHU system coupled with an inertial TES tank, serving educational spaces at the *Politecnico di Torino* campus (Italy). The TES supports seasonal heating and cooling via the AHU's thermal coils, while proportional control maintains indoor comfort. The building electrical demand is partially covered by PV system with a peak power of 47 kW.

To evaluate controller performance, a co-simulation framework integrating TRNSYS 18 with Python was developed, enabling comparison between a RB controller and a DRL agent based on the Soft Actor-Critic (SAC) algorithm. The DRL controller was trained to minimize heat pump energy consumption by aligning system operation with PV availability and electricity price signals, with constraints on TES temperature to ensure comfort and reliability. Eventually, To assess the generalizability of the DRL controller, a zero-shot transfer learning strategy was employed to deploy the pre-trained DRL agent across varying climatic conditions, tariff structures, and system fault scenarios without additional training. The following section presents the research gap, underlying motivations, and key novelties that define the contribution of this work.

Table 1
Main operational parameters of the thermal generation plant.

Parameter	Description	Value
Heating season	Period of the year during which the heating is allowed	15th October–15th April (Parliament, 2013)
Cooling season	Period of the year during which the cooling is allowed	16th April–14th October
$\Delta T_{TK-\text{Heating mode}}$	Design range of water tank top temperature in winter operation	40 °C–50 °C
$\Delta T_{TK-\text{Cooling mode}}$	Design range of water tank bottom temperature in summer operation	9 °C–14 °C
$T_{set,out B}$	Design set-point temperature of the water exiting the gas fired boiler and delivered to the reheat coils in summer operation	45 °C

1.1. Research gaps, motivations and novelties

While DRL-based HVAC control has demonstrated significant potential, important challenges still limit its broader adoption. One of the most critical issues is the limited generalizability of learned control policies beyond the conditions of the training environment. In most studies, DRL controllers are developed and validated for a specific building, climate, and operational scenario, typically relying on a fixed set of weather data and a predefined electricity tariff structure. This narrow focus raises concerns about how well these controllers perform when exposed to different conditions. For example, a policy trained for one climate zone or occupancy pattern may exhibit reduced performance when applied to a substantially different context. To address these limitations, recent research has begun to explore transfer learning (TL) techniques in HVAC control, aiming to leverage knowledge learned in one setting to improve performance in another (Xu et al., 2025; Coraci et al., 2024). For instance, Xu et al. (2020) proposed a modular DRL framework that separates transferable and building-specific components, enabling policy transfer between buildings with minimal retraining. Such transfer learning (TL) approaches have been shown to significantly reduce the training time required for new buildings or configurations by reusing components from previously trained controllers. Similarly, Lissa et al. (2021) demonstrated the direct transfer of DRL policies for heat pump control across comparable residential units within a microgrid, achieving effective control with limited additional training.

However, these methods generally still require some level of fine-tuning or partial retraining in the target environment. Critically, their effectiveness is closely tied to the similarity between the source and target domains. When building characteristics or operational conditions differ substantially, the performance of transferred policies can degrade sharply. To date, truly zero-shot transfer – where a controller trained in one context is deployed in a completely new setting without any adaptation – has not been conclusively demonstrated in the domain of HVAC control.

In addition, practical deployment challenges are often underexplored in the literature. Many studies assess DRL controllers under idealized conditions, without accounting for real-world variations such as sensor errors, equipment faults, or unexpected fluctuations in renewable energy generation. In actual building operations, events such as PV system outages, unpredictable occupancy patterns, or HVAC component failures can pose significant challenges to the reliability of a RL controller. However, there is limited evidence in the literature on how pre-trained DRL agents perform under such off-design conditions, especially without further training or adaptation.

In light of these gaps, this study explores the hypothesis that a DRL-based HVAC control policy, once trained under a specific set of conditions, can outperform a baseline rule-based controller and be applied to different operational scenarios without additional training, while still maintaining satisfactory performance.

To this purpose the study first evaluates the performance of a DRL controller, based on the SAC algorithm, in comparison to a conventional RB reference controller. The controllers are assessed in terms of energy consumption, cost savings, and thermal comfort under a baseline scenario defined by Turin's climate and electricity tariff.

Following this initial evaluation, the study investigates the ability of the pre-trained DRL controller to operate effectively in new scenarios without retraining. These include variations in climate conditions, electricity pricing schemes, and boundary conditions such as reduced PV availability. The goal is to examine both the transferability and the robustness of the DRL controller in realistic deployment settings, where operational conditions often differ from those used during training.

The rest of the paper is organized as follows: Section 2 provides an overview of the considered system layout. In Sections 3 and 4, the employed methodology and the case study are detailed. Section 5 presents the results for the analyzed case studies and the related discussion. Eventually, Section 6 provides conclusions and outlines the next steps in this field of research.

2. System layout

The selected building is a part included into the *Politecnico di Torino* campus that hosts lectures and is designed to host nearly 900 students. Specifically, the building includes four classrooms with a capacity of 220 seats each. The selected building includes five thermal zones: four classrooms and a block corridor - sanitary facilities (C-SF). The classrooms are served by a centralized variable air volume (VAV) system with zone reheat coils (RCs); while a dedicated Rooftop Unit (RTU) serves the C-SF.

The HVAC system layout with advanced control strategies is identical to the one mentioned above, the difference between the two lies only in the control logic, see Section 3.2.

This section provides a detailed description of the classrooms HVAC system. The layout of the HVAC system serving the corridor and the sanitary facility is not presented, because no advanced control strategies were implemented for this system.

2.1. Layout structure

The system layout can be structured into three main parts, Figs. 1(a), 1(b):

- The thermal generation plant (TGP).
- The air handling unit (AHU).
- The VAV terminals.

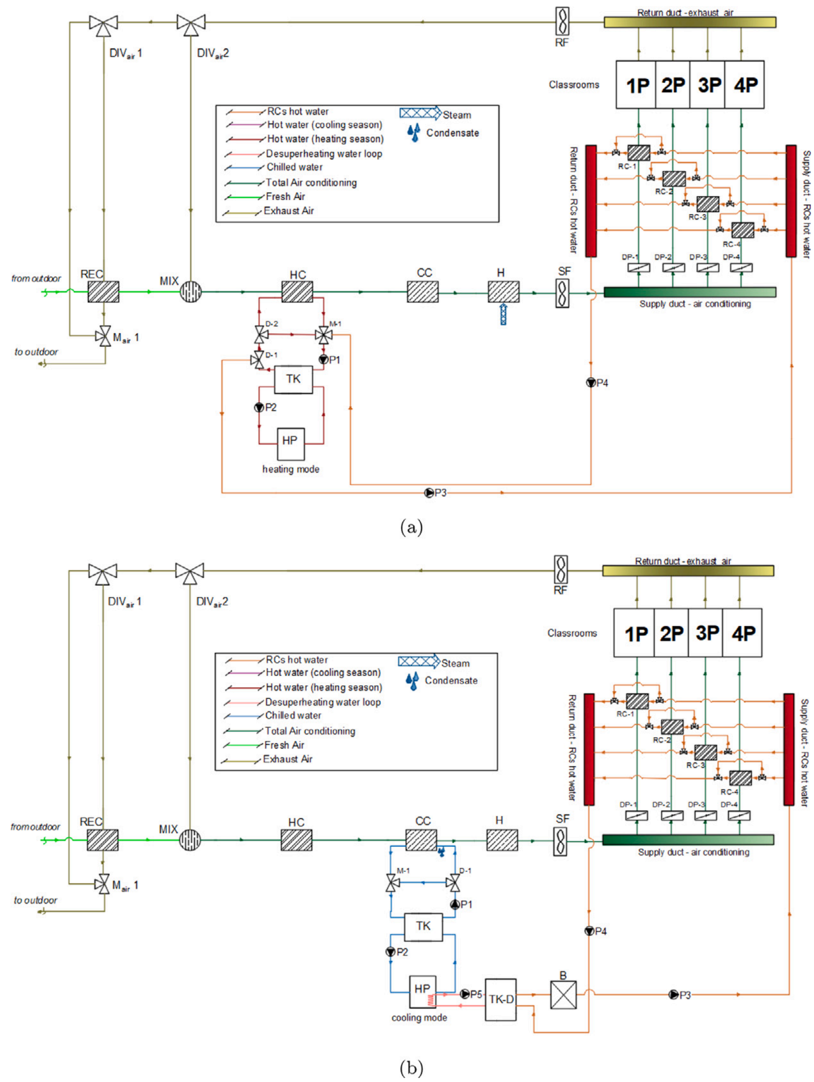


Fig. 1. HVAC system layout serving the classrooms in (a) winter operation and (b) summer operation.

2.1.1. *Thermal generation plant*

The TGP consists of a reversible air-to-water heat pump (HP) Aermec (2025) that drives the AHU and provides thermal energy to the heating coil (HC), cooling coil (CC) and reheat coils (RCs). For sake of clarity, the TGP winter operation and the TGP summer operation will be explained separately.

TGP winter operation. During the heating season (Table 1), the HP operates in heating mode. In the Baseline Control Strategy the HP is controlled by a differential controller monitoring the hot water tank (TK) top temperature ($T_{TK,top}$). This temperature is steered between 40 °C ($T_{TK,top,min}$) and 50 °C ($T_{TK,top,max}$). When the monitored temperature falls below 40 °C, the HP is activated to heat up the hot water stored in the tank, until $T_{TK,top}$ reaches the value of 50.0 °C. The hot water withdrawn from the TK is delivered to the coils to heat up the air conditioning, Fig. 1(a), Table 1.

TGP summer operation. During the cooling season (Table 1), the HP operates in cooling mode. In the Baseline Control Strategy the HP is controlled by a differential controller monitoring the chilled water TK bottom temperature ($T_{TK,bottom}$). This temperature is steered between 14.0 °C ($T_{TK,bottom,max}$) and 9.0 °C ($T_{TK,bottom,min}$). The chilled water withdrawn from the TK is delivered to the CC to cool down and dehumidify the air to be delivered to the classrooms. During summer operation, the waste heat discharged by the HP condenser is recovered for partially matching the RCs thermal energy demand. For this purpose, a heat exchanger is installed on the condenser side of the HP. Through this heat exchanger the refrigerant provides desuperheating heat to a water loop. This desuperheating water loop is connected to a thermal energy storage water tank (TK-D) that serves the RCs. A gas fired boiler (B) is used as an auxiliary system to ensure that the temperature of the water withdrawn from the TK-D to feed the RCs never drops below the set point of 45.0 °C, Fig. 1(b), Table 1. Note that this layout is a partial modification of the one provided in Cappiello (2024).

Table 2
Main operational parameters of the coils.

Parameter	Description	Value
$T_{set,HC}$	Design temperature of the air exiting the heating coil	20 °C
$T_{set,CC}$	Design temperature of the air exiting the cooling coil	15.9 °C

2.1.2. *Air handling unit*

The AHU consists of the following components: air to air sensible recovery heat exchanger (REC); an air flow mixer (MIX); a heating coil (HC); a cooling coil (CC); a steam-driven humidifier (H) and a supply fan (SF). The exhaust air withdrawn from the classrooms is collected

Table 3
Comfort and air quality parameters monitored in classrooms.

Parameter	Description	Value
$T_{indoor,winter}$	Indoor winter allowed range of temperatures for classrooms	20 ± 1 °C (Parliament, 2013)
$\Psi_{indoor,winter}$	Indoor winter allowed range of relative humidity for classrooms	40%–60% (ISO, 2017) (di Unificazione, 1995)
$T_{indoor,summer}$	Indoor summer allowed range of temperatures for classrooms	26 ± 1 °C (Parliament, 2013)
$\Psi_{indoor,summer}$	Indoor summer allowed range of relative humidity for classrooms	40%–60% (ISO, 2017) (di Unificazione, 1995)
CO_2 *concentration*	Level of CO_2 concentration allowed for classrooms	≤ 1000 ppm (ISO, 2017)

by a return fan (RF). For the sake of clarity, the AHU winter operation and the AHU summer operation will be explained separately.

AHU winter operation. During the heating season a controller managing the AHU selects a suitable fresh air mass flow rate to be delivered to the classrooms. Then, this fresh air mass flow rate is delivered to the REC, where it is preheated exploiting the waste heat of the exhaust air leaving the plant. In the MIX the preheated air mass flow rate is mixed with the recirculated air mass flow rate. This mixed stream is heated up to a fixed temperature of 20.0 °C through the HC, Table 2. Then, the preheated air flows through the humidifier, which is managed by a proportional controller. The controller monitors the humidity of the air in the return duct and selects a suitable amount of steam to be delivered to the air stream to maintain the classrooms indoor relative humidity within the design range of 40%–60%, Table 3. Steam is obtained by means of a steam generator. After this, the air stream is delivered to the diverting valve, which delivers equal mass flow rate to each branch. Note that each branch supplies the VAV terminal installed in each classroom. Through the VAV terminal, the air mass flow rate can be further heated up according to the indoor temperature of the specific classroom. However the VAV terminal winter operation will be more in-depth in an appropriate Section 2.1.3.

AHU summer operation. During the cooling season a controller managing the AHU selects a suitable fresh air mass flow rate to be delivered to the classrooms. Then this fresh air mass flow rate is delivered to the REC, where it is precooled exploiting the waste cooling energy of the exhaust air leaving the plant. In the MIX the precooled air mass flow rate is mixed with the recirculated air mass flow rate. This mixed stream is cooled down to a fixed temperature of 15.9 °C through the CC, Table 2. Therefore, the air flow downstream of the CC is cooled down and dehumidified; this allows, among other things, to reduce the specific humidity of the air stream and to maintain the classrooms indoor relative humidity within the design range of 40%–60%, Table 3. Following the CC and through the air diverting valve the supply air flow rate is divided into four parts (not necessarily equal, unlike the previous winter operation) each sent to the corresponding VAV terminal and then to the corresponding classroom. In the cooling season, the design classrooms indoor air temperature is 26.0 ± 1.0 °C Table 3, this indoor temperature is controlled by means of VAV terminals. The VAV terminal summer operation will be more in-depth in an appropriate Section 2.1.3.

2.1.3. VAV terminals

The VAV terminal consists of a VAV box containing a damper (DP) and a reheat coil (RC). Thanks to the VAV terminal it is possible to control the delivered thermal energy according to the specific classroom demand. For the sake of clarity, the VAV terminal winter operation and the VAV terminal summer operation will be explained separately.

VAV terminal winter operation. During the heating season the supply air flow rate is divided into four equal parts, each sent to the corresponding VAV terminal. Therefore, during the winter operation, the air conditioning system works at constant flow rate and the building space heating demand of each classroom is tracked by the zonal RC. In particular, a specific thermostat monitors the indoor temperature of each classroom, this signal is handled by a proportional controller, managing the RC outlet air temperature. This means that the proportional controller manages the hot water flow rate delivered to the RC. In other words the RC adjusts the hot water flow to match the heating load, in order to steer the indoor air temperature of the classroom at 20.0 ± 1.0 °C. Note that this layout is developed according to the state of the art in the matter of VAV systems, Cammarata (2016).

VAV terminal summer operation. During the cooling season the supply air flow rate is divided into four parts, each sent to the corresponding VAV terminal. Specifically, each air mass flow rate is changed in order to balance the classroom building space cooling load. Thus, unlike the winter operation, the system works at variable air mass flow rate. In particular, a thermostat installed in each classroom reads the indoor temperature. This information is used by a controller managing the operation of the AHU. In particular, a proportional controller manages the damper opening, i.e., the air mass flow rate is regulated according to the cooling load to be matched, for steering the indoor air temperature of the classroom at 26.0 ± 1.0 °C. Note that the air mass flow rate can be reheated by the RC. This layout is developed according to the state of the art in the matter of VAV systems, Cammarata (2016).

2.2. Air flows management and free-cooling mode

The external air withdrawn by the VAV system is selected by a suitable controller monitoring the occupation of the classrooms. According to *UNI 10339* (di Unificazione, 1995), at least a fresh air flow rate of $25.2\,\mathrm{m}^3/\mathrm{hr}$ for each person must be guaranteed in each classroom. However, an additional fresh air flow rate can be withdrawn based on the measured levels of CO_2 concentration. Specifically, according to the standard *ISO 17772-1:2017* (ISO, 2017), the CO_2 concentration must not exceed the maximum threshold of 1000 ppm.

The air mass flow rate to be delivered to each classroom is selected by a proportional controller changing the damper opening in the zonal VAV box, steering the classroom indoor air temperature to the selected set point. There are two controls: one sets the external air flow rate; the other sets the air flow rate in each specific classroom and thus the total air flow rate. During the summer operation the VAV system can also operate in free-cooling mode, i.e. when the external air temperature is sufficiently lower than the exhaust air in the return duct, the ratio of fresh air flow rate and recirculated air flow rate is properly selected by the controllers: the air mass flow rate is directly delivered to the classrooms without activating the HP.

3. Methodology

This section provides a description of the methodological framework (Fig. 2) that led to the definition of the models and control strategies developed to manage the HVAC system serving the classrooms. The classroom model is implemented within the TRNSYS environment (Klein et al., 2017), while the advanced control strategies are developed using Python. Section 3.1 introduces the TRNSYS software

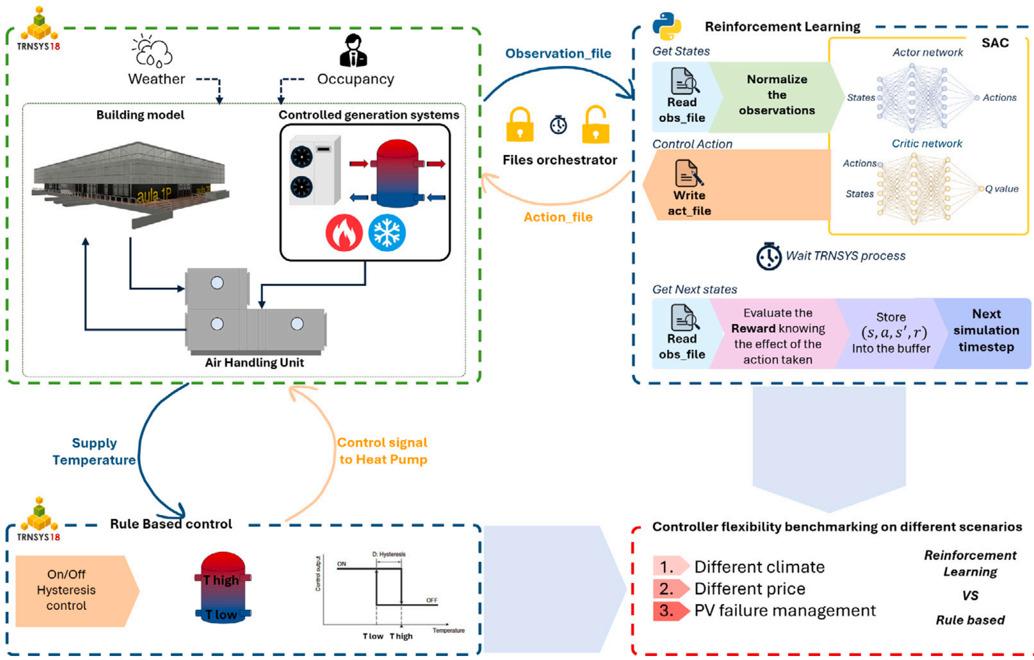


Fig. 2. Methodological framework.

and the main libraries used in the model, including: the building model, the heat pump model, and the thermal Energy Storage tank model. In Section 3.2, the basic control strategy is initially summarized, followed by an in-depth analysis of the control strategy based on Deep Reinforcement Learning, with a specific focus on the Soft Actor-Critic algorithm. Finally, the methodology adopted for training the DRL agent is described in detail. Section 3.3 discusses the co-simulation environment, outlining its architecture and different phases of execution. To verify the scalability and flexibility of the DRL controller, the trained agent was evaluated and compared with the RB controller under different operational scenarios, using transfer learning techniques. The description and rationale behind these scenarios are given in Section 3.4. Finally, Section 3.5 presents the Thermoeconomic Key Performance Indicators based on which the comparison between the results of the RB control system and RL control system is performed.

3.1. TRNSYS model

The models of the proposed HVAC systems and of the selected classrooms are developed in TRNSYS 18, a well-known and reliable tool widely adopted in academic and commercial applications. This software comes with an extensive library of components and enables real-time operative simulations of complex renewable energy-based plants and systems, Refs. Cacabelos et al. (2015), Cappiello and Erhart (2021), Buonomano et al. (2013).

3.1.1. Building model

The geometric model of the selected building was defined using Google SketchUp and TRNSYS 3D. Then, it is imported into the TRNSYS 18 environment, using type 56. This library allows an accurate modeling of the thermophysical properties of the building while also allowing the user to subdivide the building into different thermal zones. In addition, this type can simulate the dynamic energy performance of the building by considering internal thermal gains, solar gain, building orientation, while considering shading effects from objects such as overhangs or nearby buildings that shield the building from direct solar radiation. More details on this library are provided in Calise et al. (2023a), Rashad et al. (2022).

3.1.2. Heat pump model

Type 941 models the performance of a reversible single stage air-to-water heat pump adopting the normalized catalogue data look-up principle. Therefore, Type 941 performs the mass and energy balances considering the data provided by the manufacturer's operating map. Further information on this subject can be found in Cappiello (2024), Kropas et al. (2022).

3.1.3. Thermal energy storage tank model

Type 4 models the performance of a fluid-filled sensible energy storage tank. The fluid in the tank is subject to a thermal stratification by assuming that the tank consists of N fully-mixed equal volume segments. The Multi-port Design of this type allows the entry and exit of fluids at different levels of stratification; the thermal cycling of this type allows, among other things, a more efficient operation of the HP with which it is coupled. More details on this type are given in Klein et al. (2017).

3.2. Control strategies

3.2.1. Baseline control strategy

The Baseline Control Strategy employs a traditional hysteresis-based approach, aimed at maintaining the thermal energy storages' temperatures within predefined thresholds:

$$40\ ^{\circ}\mathrm{C} \leq T_{TK,\mathrm{top},Heating\ mode} \leq 50\ ^{\circ}\mathrm{C} \tag{1}$$

$$9\ ^{\circ}\mathrm{C} \leq T_{TK,\mathrm{bottom},Cooling\ mode} \leq 14\ ^{\circ}\mathrm{C} \tag{2}$$

The Baseline Control Strategy is well explained in Section 2.1.1 for both winter operation and summer operation. This rule-based approach is simple, reliable, and widely used in practical applications. However, it does not account for external factors such as energy prices, occupancy patterns, and weather conditions. As a result, its performance can be suboptimal in dynamic or highly variable environments, where these external factors significantly affect the efficiency and cost of the system.

Table 4
State and action variables for heat pump energy system control optimization, showing operational ranges and units for normalization.

State Variable	Min	Max	Unit
$j_{el,fromGRID}$	0.09	1.0	€/kWh
$j_{el,fromGRID\ +1h}$	0.09	1.0	€/kWh
$j_{el,fromGRID\ +3h}$	0.09	1.0	€/kWh
$j_{el,fromGRID\ +6h}$	0.09	1.0	€/kWh
$P_{el,\ HP}$	0	69.44	kW
$\Delta T_{HP\,TK}$	−25	25	°C
j_{NG}	0.8	1.3	€/Sm^3
V_{NG}	0	15	Sm^3
$P_{el,PV}$	0	120	kW
$P_{el,fromGRID}$	0	120	kW
COP	0	10	–
$T_{outdoor}$	5	35	°C
$T_{outdoor\ +1h}$	5	35	°C
$T_{outdoor\ +3h}$	5	35	°C
$T_{outdoor\ +6h}$	5	35	°C
Occupancy	0	1	–
Timestep Next Occupancy	0	169	h
Timestep Occupancy	0	41	h
Heat transfer to air, AHU	0	250	kW
Heat transfer to water, HP	0	250	kW
Action Variable	**Min**	**Max**	**Unit**
Heat pump operation control signal	0 (off)	1 (on)	binary

3.2.2. Deep reinforcement learning-based control strategy

Deep Reinforcement Learning is a machine learning paradigm in which an agent learns to make decisions by interacting with its environment to achieve a specific objective. The agent performs actions, receives feedback through rewards or penalties, and uses this feedback to improve its decision-making process. The ultimate goal is to derive a policy, i.e., a sequence of actions that maximizes the cumulative rewards received over time, Barto (1997). The foundation of DRL lies in the Markov Decision Process (MDP) (Puterman, 1990), a mathematical framework for modeling sequential decision-making problems. At the core of an MDP are several key components that guide the agent's learning and adaptation:

- **States (S)**: A set of observable variables, $S = s_1, s_2, \ldots, s_n$, representing all possible situations the agent might encounter.
- **Actions (A)**: A set of possible decisions, $A = a_1, a_2, \ldots, a_n$, from which the agent can choose at each step.
- **Transition Probabilities (T)**: These define the likelihood of transitioning from one state to another, given a specific action, represented as $T = (s, a, s')$.
- **Rewards (R)**: A numerical value assigned to specific state–action pairs, $R = (s, a, s')$, which provides feedback to the agent on the quality of its decisions.
- **Discount Factor** (γ): A parameter ($0 \leq \gamma \leq 1$) that balances the importance of immediate rewards against future rewards, enabling the agent to prioritize long-term gains.
- **Policy** (π): A strategy or decision-making rule that determines the agent's actions. The goal is to discover an optimal policy, denoted as π^*, which maximizes the expected cumulative reward.

The DRL agent learns the dynamics of the environment through Replay Buffers, which store experiences in the form of tuples (s, a, r, s', d) where s is the state, a is the action, r is the reward, s' is the next state, and d is the done signal indicating episode termination. The replay buffer facilitates the off-policy learning process of SAC by allowing the agent to learn from past experiences rather than relying solely on the latest interaction.

When neural networks are employed to approximate the policy and value functions, the approach is classified as DRL. This extension has proven highly reliable for systems with complex state and action spaces.

For this case study, a *Soft Actor Critic* (Haarnoja et al., 2018) controller was developed based on a DRL algorithm optimized for continuous control tasks. The controller leverages an actor-critic framework, where the *actor* is responsible for learning the policy function, mapping observations of the environment to actions, and the *critic* evaluates the quality of these actions relative to the current state.

A key innovation of the algorithm is its use of entropy regularization, which balances exploration and exploitation by encouraging diverse action selection. By maximizing both expected rewards and policy entropy during training, the agent learns strategies that are both effective and adaptable to changing conditions.

For the current case study, the continuous *action* taken by the agent within the range $[0, 1]$ was adapted to meet the requirements of the heat pump model in the TRNSYS environment, which operates with an On/Off control logic. To achieve this, the continuous action space was discretized into two distinct intervals corresponding to the binary states of the heat pump: On and Off.

The DRL control strategy dynamically adjusts system operations based on a *reward function* that incorporates multiple objectives. In this study, the reward function is formulated to promote energy efficiency by minimizing both electrical and natural gas consumption, while ensuring the operating temperatures of the heat pump and air handling unit remain within a safe and efficient range. The reward R_t at each timestep is defined as:

$$R_t = -\beta_1 \cdot \phi_{\text{operative costs}} - \beta_2 \cdot \phi_{\text{temperature}} \tag{3}$$

where (β_1 and β_2) are the weighting coefficients for the operative costs and temperature penalties, respectively. These coefficients were determined through iterative adjustments to achieve a balanced contribution from both penalty terms to the overall reward function. This balanced approach prevents either objective from dominating the learning process, allowing the agent to simultaneously optimize energy costs while maintaining safe operating temperatures.

The penalties related to the operative costs $\phi_{\text{operative costs}}$ and to temperature $\phi_{\text{temperature}}$ are defined as follows:

$$\phi_{\text{operative costs}} = \alpha_{el} \cdot (P_{el,\ HP} - P_{el,PV}) \cdot j_{el,fromGRID} + \alpha_{NG} \cdot V_{NG} \cdot j_{NG} \tag{4}$$

$$\phi_{\text{temperature}} = (\Delta T_{HP\,TK})^2 \cdot Occupancy \tag{5}$$

- $\phi_{\text{operative costs}}$ is defined as the sum of the cost of net and instantaneous electric demand, computed as the difference between the HP load ($P_{el,\ HP}$) and on-site photovoltaic generation ($P_{el,PV}$), multiplied by the current electricity price ($j_{el,fromGRID}$), and the cost of natural gas, calculated as the volume of the consumed natural gas (V_{NG}) multiplied by its corresponding price (j_{NG}). These cost terms are appropriately weighted through α factors to reflect their economic impact.

- $\phi_{\text{temperature}}$ is a penalty reflecting the deviations from the desired tank temperature range, weighted more heavily during occupied periods by means of the variable *Occupancy*, Table 4 . In particular, $\Delta T_{HP\ TK}$ represents the temperature difference between the heat pump's water supply temperature and the temperature of the thermal storage tank: during the winter period, it reflects the difference between the minimum value of the top of the tank ($T_{TK,top,min}$ = 40 °C, Table 1) and the heat pump's output hot water temperature. In the summer period, $\Delta T_{HP\ TK}$ is the difference between the heat pump's output chilled water temperature and the maximum value of the bottom of the tank ($T_{TK,bottom,max}$ = 14 °C, Table 1), see Eq. (6). This penalty formulation has been designed to ensure a fair comparison between the DRL and RB controller, which operates based on a hysteresis strategy. The RB maintains the thermal energy storage temperature within predefined limits, specifically between 40 °C and 50 °C in winter and 9 °C and 14 °C in summer. To allow for a meaningful evaluation, the RL approach is penalized for deviations from these reference temperature ranges, ensuring that any performance improvements are not merely a result of relaxed temperature constraints but rather stem from enhanced operational efficiency.

$$\begin{aligned}\Delta T_{\text{HP TK, Heating mode}} &= T_{TK,top,min} - T_{\text{hot water out HP}}\\ \Delta T_{\text{HP TK, Cooling mode}} &= T_{\text{chilled water out HP}} - T_{TK,bottom,max}\end{aligned} \tag{6}$$

The selected *observation* variables reported in Table 4 provide a comprehensive representation of the system's state, crucial for the DRL agent to make informed decisions. Each variable captures essential aspects of the building's energy dynamics, external conditions, and operational performance. It is important to note that the minimum and maximum values listed in Table 4 do not represent the absolute extremes observed in the system but rather serve as normalization bounds for training the DRL agent. This normalization process is crucial, as it ensures numerical stability, prevents large variations from dominating the learning process, and facilitates efficient convergence of the DRL model. By scaling input values within a predefined range, the agent can learn more effectively and generalize its policy across different operational scenarios. The observation variables selected for the DRL agent encompass key factors critical to optimizing energy use and minimizing costs. Variables such as electricity and natural gas prices enable cost-effective decision-making. Power and heat related variables allow monitoring the energy consumptions, aiding in the effective use of renewable energy sources.

For the training phase two dedicated DRL controllers were developed independently: a DRL controller for heating operations and a DRL controller for cooling operations. The training period for the first controller was from January 1st to mid-March, allowing the agent to focus on optimizing heating operations during the colder months. Similarly, the other controller was trained from mid-May to the end of July, allowing it to specialize in cooling operations under warmer seasonal conditions (in August the building is closed). By narrowing the scope of each training period to these targeted intervals, the computational demands of the training process were significantly reduced while maintaining the effectiveness of the controllers. Following the training phase, the controllers were deployed to simulate their performance over an entire year.

Meteorological data for the city of Turin and 2023 electricity and natural gas prices data were used to reflect real-world conditions. During deployment, the winter-trained controller was applied during the heating season, which in Turin extends from 15th October to 15th April (Parliament, 2013). The summer-trained controller was deployed during the rest of the year. This seasonal deployment strategy ensured that each controller operated within the period most aligned with its training conditions, thereby maximizing system performance. Furthermore, the division of the training process into two targeted periods allowed for both computational efficiency and the development of controllers that were fine-tuned to their respective seasonal requirements.

3.3. Co-simulation environment

A key role of this study is the co-simulation framework that combines the dynamic simulation capabilities of TRNSYS with a Python-based SAC agent to optimize building energy management. The library Type 3157 is used to integrate the TRNSYS model and the advanced controller built in Python.

The co-simulation operates iteratively, synchronizing the TRNSYS simulation environment with the decision-making process of the DRL agent in Python. During initialization, TRNSYS configures key parameters, including the simulation time range, timestep, and dimensions of the observation and action spaces. The simulation's state is captured in an observation file, while the DRL agent's chosen actions are recorded in an action file. To ensure seamless synchronization, two signal files are employed to coordinate the exchange of data, guaranteeing that TRNSYS progresses to the next simulation step only after receiving the updated actions from the DRL agent.

At each timestep, TRNSYS evaluates the current state of the simulation environment based on the actions applied during the previous step. These observations are saved to the observation file, and a signal file notifies the DRL agent that the data is ready. The DRL agent reads the observations and computes the optimal action using the SAC algorithm. The computed action is stored in the action file, and a corresponding signal file informs TRNSYS that the action is ready to be implemented in the next simulation step. This process repeats until the simulation reaches its termination condition, Fig. 2 .

During training, the SAC agent alternates between data collection from TRNSYS and parameter updates. Observations are normalized within the range $[0,1]$ to improve stability. In addition, detailed orchestration mechanisms ensure data integrity during the file-based communication process. Python's threading and file locking mechanisms prevent race conditions between read/write operations.

This co-simulation framework enables the integration of TRNSYS, a high-fidelity simulator, with SAC, a state-of-the-art DRL algorithm. Using this combination, the integrated model achieves dynamic optimization of building energy management while ensuring scalability and adaptability to complex real-world scenarios. A detailed description of the co-simulation phases is given in the Algorithm 1.

3.4. Benchmarking scenarios and transfer learning strategy

To evaluate the scalability and generalizability of the DRL controller, the trained agent was compared to the RB controller across multiple scenarios designed to simulate complex, real-world operational challenges. Specifically, the benchmarking tests involve the following varying scenarios:

- Different climatic conditions: The controller was tested in several Italian climate zones to assess performance under different weather conditions. These included Napoli, representing a moderately warmer climate than Torino; Lampedusa, with a significantly hotter and more humid climate; and Bolzano, characterized by a colder alpine climate. Typical meteorological year (TMY) data for each location were generated using Meteonorm (Meteonorm, 2024).

Algorithm 1 Co-Simulation Process Between TRNSYS and Python SAC Agent

Require: Observation file obs_file, Action file act_file, Signal files $signal\_obs$, $signal\_act$, DRL agent π_θ, simulation time T, TRNSYS environment
Ensure: Trained DRL policy π_θ
Initialize simulation time $t \leftarrow 0$, action $a_t \leftarrow 0$, episode counter $n \leftarrow 0$
Write initial observations s_0 to obs_file (zeros or system-defined values)
while Simulation not terminated ($t < T$) **do**
  Observation Phase
  TRNSYS writes system state s_t to obs_file
  TRNSYS signals completion by creating $signal\_obs$
  Wait for TRNSYS confirmation: $signal\_obs \neq \emptyset$
  Read o_t from obs_file
  Agent Action Phase
  Normalize o_t using min-max scaling: $\tilde{s}_t = \frac{s_t - s_{\min}}{s_{\max} - s_{\min}}$
  Select action $a_t = \pi_\theta(\tilde{s}_t)$
  Discretize action $a_t \rightarrow \{0, 1\}$ (if binary control is used)
  Write a_t to act_file
  Signal TRNSYS by creating $signal\_act$
  TRNSYS Execution Phase
  Wait for TRNSYS to process action: $signal\_act \neq \emptyset$
  TRNSYS processes a_t and advances simulation, returning o_{t+1}
  TRNSYS writes new observation o_{t+1} to obs_file
  TRNSYS deletes $signal\_act$ to signal the agent
  Reward Computation Phase
  Compute reward r_t
  Agent Update Phase
  Store (s_t, a_t, r_t, s_{t+1}) in replay buffer $\mathcal{D}$
  if $t >$ learning_start **then**
    Sample batch $\mathcal{B} \sim \mathcal{D}$
    Update actor and critic parameters using SAC:
  end if
  Increment time step: $t \leftarrow t + 1$
end while
Save the trained policy π_θ for deployment

Table 5
Scenarios characterization.

Scenario	Climate zone	Reference year for electricity tariffs	PV fault
Reference	Torino	2023	No
Scenario 1	Lampedusa	2023	No
Scenario 2	Bolzano	2023	No
Scenario 3	Napoli	2023	No
Scenario 4	Torino	2022	No
Scenario 5	Torino	2023	Yes

- Different electricity tariffs: The controller's ability to optimize cost savings under varying market conditions was assessed by testing its performance against real electricity price data from 2022 and 2023. These data were obtained from publicly available historical records provided by the Italian energy market operator (Gestore dei Mercati Energetici, GME) (Gestore dei Mercati Energetici, 2024). The two years reflect distinct pricing dynamics, with 2022 characterized by significantly higher electricity costs.
- System fault scenario: A fault condition was introduced by simulating the complete failure of the PV system. This scenario was designed to evaluate the DRL controller's ability to adapt its operation and maintain stable performance in the presence of unexpected disruptions.

In total, six distinct scenarios were evaluated, summarized in Table 5.

The evaluation focused on periods during the summer (15th June to 15th July) and winter (15th November to 15th December), selected for their representativeness of the cooling and heating operational modes of the system.

The SAC based DRL controller was initially trained under a baseline scenario (Torino, 2023) and then directly deployed in all other test scenarios using a zero shot transfer learning approach. The trained neural network parameters, including the weights of the policy (actor) and value (critic) networks, were transferred without any retraining or fine tuning.

Zero shot transfer is appropriate as the underlying Markov Decision Process (MDP) – including system configuration, state and action definitions, and the reward function – remains unchanged across scenarios. Only exogenous inputs, such as weather forecasts, electricity prices, and PV availability, differ. The agent uses a forecast informed state representation that includes multi horizon predictions of outdoor temperature and electricity prices (Table 4). To ensure stable performance and avoid distributional issues, normalization bounds were adjusted for each scenario based on expected input ranges.

This benchmarking and transfer setup demonstrates the DRL controller's ability to scale, adapt, and perform reliably across varied operating conditions.

3.5. Thermoeconomic key performance indicators

The comparison between the RB and DRL systems is carried out using thermoeconomic Key Performance Indicators (KPIs), designed to assess both energy and economic performance.

Energy KPIs are: the Self-Consumption Ratio (SCR); the Self-Sufficiency Ratio (SSR) and the Renewable Energy Ratio (R_{renew}) index.

The Self-Consumption Ratio assesses the amount of the total PV energy production that is locally self-consumed by the building. It is defined as the ratio of electricity not exported to the grid to total PV production ($E_{el,PV}$). It can be expressed as follows.

$$SCR = \frac{E_{\text{el,PV}} - E_{\text{el,toGRID}}}{E_{\text{el,PV}}} \cdot 100 \tag{7}$$

Where $E_{el,toGRID}$ represents the excess electricity delivered to the power grid.

The Self-Sufficiency Ratio (SSR) measures the degree of independence of the building from the power grid. It is defined as the ratio of self-consumed energy to the user total load ($E_{el,LOAD}$). It can be evaluated as follows.

$$SSR = \frac{E_{\text{el,LOAD}} - E_{\text{el,fromGRID}}}{E_{\text{el,LOAD}}} \cdot 100 \tag{8}$$

Where $E_{el,fromGRID}$ is the electricity withdrawn from the power grid.

The R_{renew} index estimates how much primary energy (PE) is saved in the transition from the RB system to the DRL system.

$$R_{renew}[\%] = \frac{PE_{RB} - PE_{RL}}{PE_{RB}} \cdot 100 \tag{9}$$

$$PE = \sum_t \left[\frac{E_{el,\text{from GRID}}}{\eta_{el}} + \frac{E_{th,\text{B}}}{\eta_B} + \frac{E_{th,\text{SG}}}{\eta_{SG}} \right]_t \tag{10}$$

The primary energy demands of the two compared systems are evaluated according to Eq. (10). The values of the conventional thermoelectric power plant (η_{el}), the boiler (η_B) and the steam generator (η_{SG}) efficiencies are depicted in Table 6.

The building electricity demand includes the electricity for driving the HVAC system and the electricity for zones lighting.

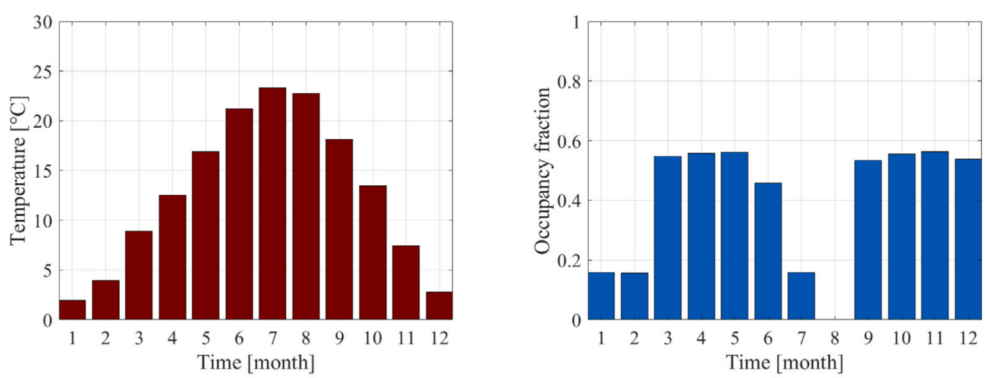
From an economic point of view, the KPI is the percentage change in operative costs (C_{op}) between the RB and the DRL systems:

$$\Delta Cop[\%] = \frac{C_{op,RB} - C_{op,RL}}{C_{op,RB}} \cdot 100 \tag{11}$$

$$C_{op} = \sum_t \Big[j_{\text{el,from GRID}} E_{\text{el,from GRID}} - j_{\text{el,to GRID}} E_{\text{el,to GRID}}$$

Table 6
Design and operating parameters.

Parameter	Description	Value	Unit
$j_{el,from\ GRID}$	Electricity purchasing cost (hourly variable) (Gestore dei Mercati Energetici, 2024)	0.091 [min] 0.448 [max] 0.244 [average]	€/kWh
$j_{el,to\ GRID}$	Electricity exporting cost (yearly constant) (Calise et al., 2023b)	0.06	€/kWh
j_{NG}	Natural-Gas purchasing price (monthly variable) (Gestore dei Mercati Energetici, 2024)	0.806 [min] 1.289 [max] 0.948 [average]	€/Sm^3
LHV_{NG}	Natural gas lower heating value	9.59	kWh/Sm^3
η_{el}	thermo-electric power plant efficiency	0.46	–
η_B	Boiler efficiency	0.75	–
η_{SG}	Steam generator efficiency	0.98	–



(a) (b)

Fig. 3. Monthly average temperatures in Turin (a) and monthly average occupancy fraction of classrooms (b).

$$+ j_{\mathrm{NG}}\left(\frac{E_{\mathrm{th,\ B}}}{\eta_B \cdot LHV_{NG}} + \frac{E_{\mathrm{th,\ SG}}}{\eta_{SG} \cdot LHV_{NG}}\right)\Bigg]_t \tag{12}$$

The operative costs of the two compared systems are evaluated according to Eq. (12). Table 6 summarizes the main thermoeconomic assumptions.

4. Case study

The studied educational building is located in Turin, within Italy's climate zone E (Della Repubblica, 1993), characterized by cold winters and moderately warm summers, Fig. 3(a). The building includes four identical classrooms (with a capacity of 220 seats each), a central distribution corridor, and two sanitary facilities. Each classroom has a floor area of 218.09 m^2, resulting in a total building floor area of 1378.28 m^2. The height of each room is 5.10 m, where the volume of the single classroom is 1112.26 m^3, while the volume of the whole structure is 7029.23 m^3. All the main geometric features are provided in the Appendix. Fig. 4 displays the 3D model of the selected building with different views. Note that, a lightweight, semi-transparent metal envelope surrounds the building. More information about the envelope features can be found in the Appendix. The main geometric and thermophysical features of this building were evaluated through various inspections and the study of technical documentation related to the building and HVAC system, further information was also found in the work of Berta et al. (2017). The average number of students in the classrooms varies according to the month under consideration: during the months when university courses are delivered a larger presence of students occurs. In this context, Fig. 3(b) shows the classrooms monthly average occupancy compared to total classroom capacity. It is assumed that the building is open from 8:30 a.m. to 6:30 p.m. during weekdays and from 8:30 a.m to 2:00 p.m. on Saturday. These assumptions are consistent with the information available on the Politecnico di Torino website (Politecnico of Turin, 2024) and with a briefly survey performed among the students using such classrooms. For the analyzed case study the heat gains due to students and lights Appendix are selected according to of Heating et al. (2009) and Cammarata (2003). In addition, the heat gains related with electrical equipment were neglected because they have little impact, either by number or heat loss, in relation to the volume of the thermal zone, to the definition of the indoor thermal load. In Appendix the main features of the HVAC system are also reported. Finally, on the rooftop of the building a PV field of 47 kW is installed.

5. Results & discussion

This section presents the results achieved by the DRL controller in comparison with the RB controller. The analysis first examines the outcomes of a full-year simulation under the *Reference Scenario* conditions (Table 5), providing an overall assessment of the DRL controller's performance. Subsequently, the evaluation will extend to different study scenarios, examining how the DRL controller adapts to varying operational and environmental conditions.

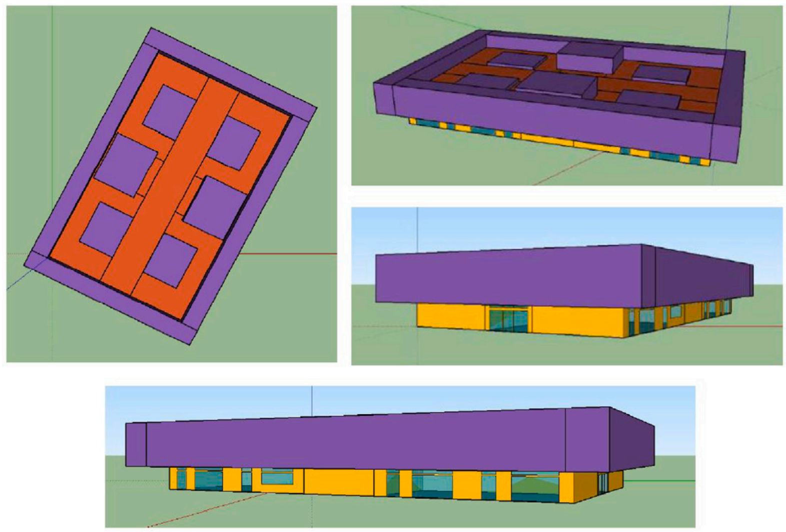


Fig. 4. 3D model of the studied classrooms P building.

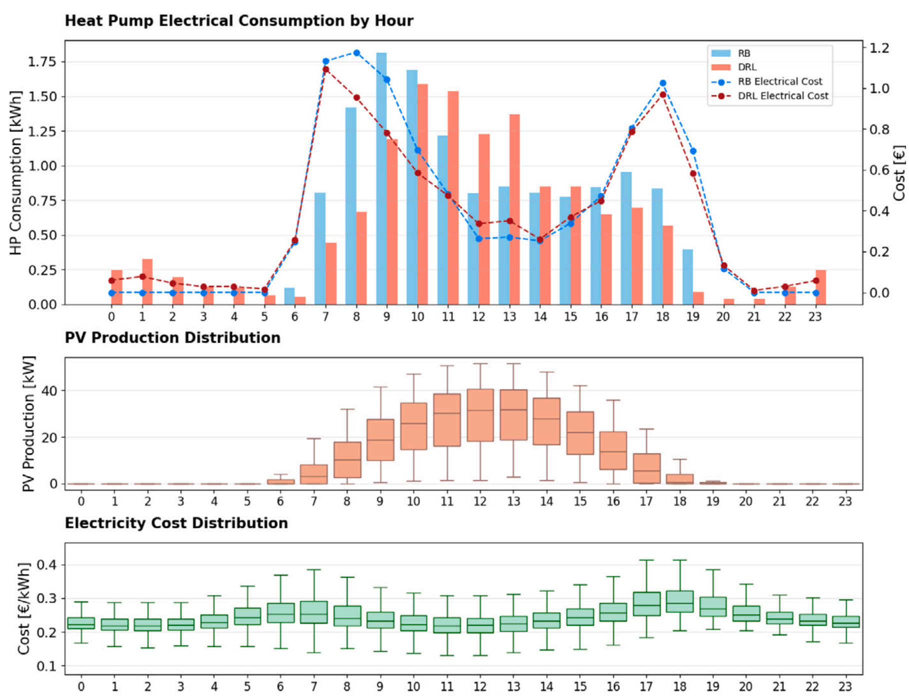
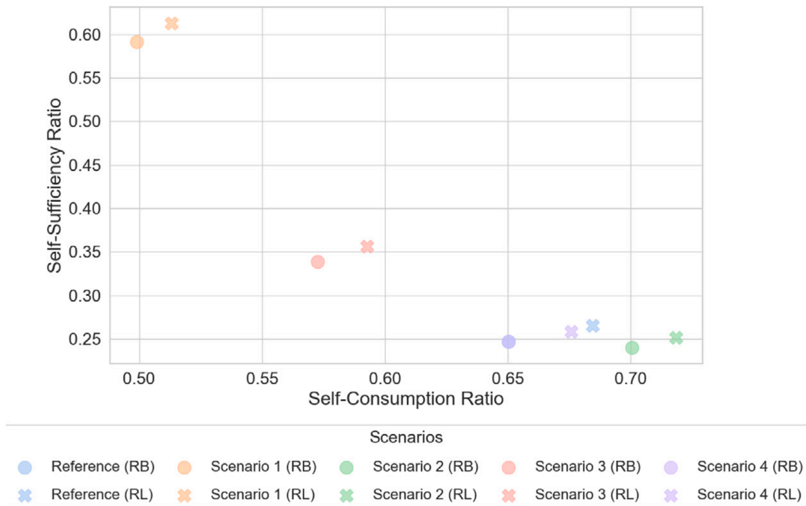


Fig. 5. Average hourly Heat Pump consumption, PV production, and electricity cost distribution, evaluated over the entire year.

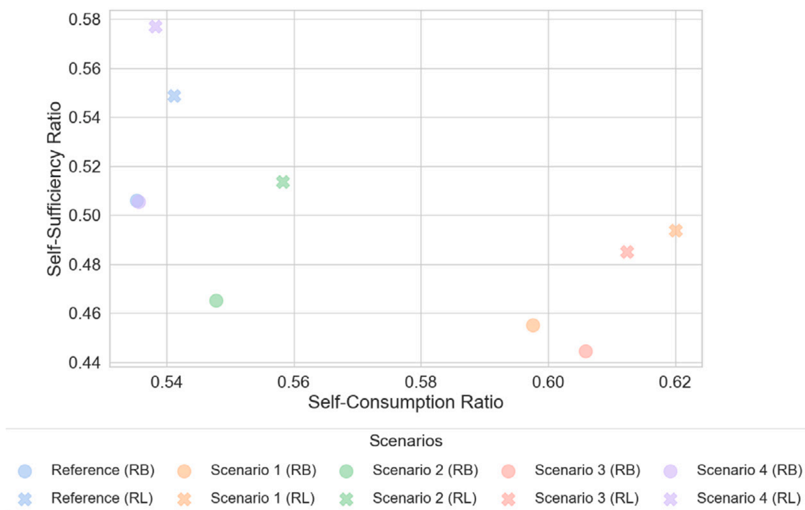
5.1. Reference scenario results

To evaluate the baseline performance of the DRL controller, this section analyses its operation under the *Reference Scenario*, comparing it to the RB controller over a full year. A key aspect of this comparison is understanding how the DRL controller optimally schedules heat pump activation in response to photovoltaic production and electricity cost variations. Fig. 5 illustrates the hourly distribution of the heat pump's average electrical consumption, along with the corresponding average electricity cost, photovoltaic power production, and electricity prices, evaluated over the entire year. This visualization highlights how the DRL strategy optimally manages energy resources compared to RB approach. The analysis of heat pump activation patterns indicates that the DRL controller exhibits greater flexibility, dynamically responding to photovoltaic generation and electricity price fluctuations to capitalize on favorable conditions. In contrast, the RB controller follows a predefined, rigid operational schedule, lacking the ability to adapt to changing external conditions.

Although the DRL-controlled heat pump shows increased electricity consumption at specific times, particularly around midday, the dotted lines in the first subplot, representing electricity costs, indicate that the total electricity expenditure under DRL remains lower overall than that of the RB approach. This is due to the adaptive nature of the DRL controller, which dynamically responds to external conditions. In fact, it is able to exploit the renewable PV production, turning on the heat pump in the central part of the day, when the maximum renewable power production occurs. In addition, DRL also activates the HP during the night when the price of the electricity is lower. Thus, such controller



(a)



(b)

Fig. 6. SCR and SSR across different scenarios in (a) winter period and (b) summer period.

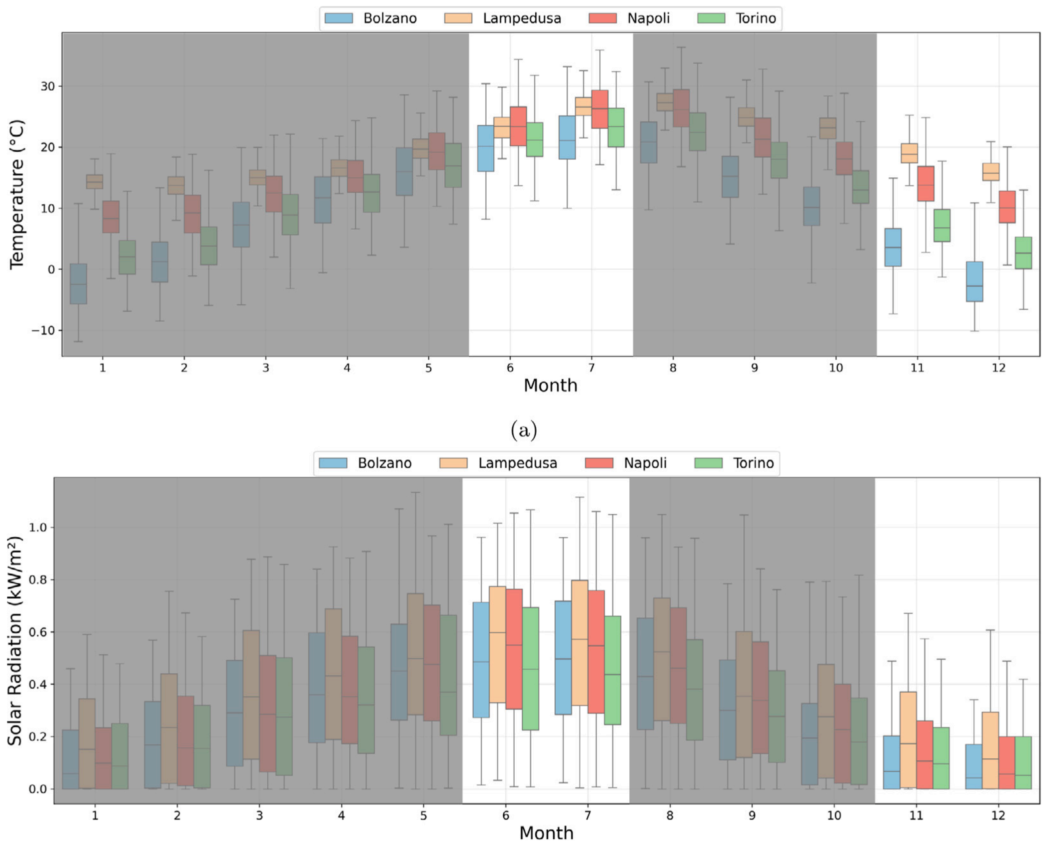
Table 7
Comparison of RB and DRL performance metrics for the entire year.

Metric	RB	DRL	Δ	Unit
Energy Metrics				
$E_{el,PV}$	79.79	79.79	0	MWh
$E_{el,LOAD}$	86.49	86.64	−0.15	MWh
$E_{el,HP}$	19.45	19.43	0.02	MWh
$E_{el,fromGRID}$	48.49	46.96	1.53	MWh
$E_{el,toGRID}$	40.51	39.21	1.30	MWh
PE	157.80	153.07	4.73	MWh
R_{renew}		3.0		%
SCR	49.22	50.85	−1.63	%
SSR	43.94	45.80	−1.86	%
Cost Metrics				
C_{op}	15 916	15 402	514	€
C_{el}	13 159	12 644	515	€
C_{NG}	5 187	5 102	85	€

achieves the dual results of maximizing the renewable electricity self-consumption and exploiting the lower electricity prices. Note that, this results is related to the fact that the DRL exploits the thermal energy storage in order to store the heat during the hours of higher renewable production or the hours of lower electricity price. This means that the stored thermal energy is released during less favorable conditions.

Therefore, DRL leads to an energy-shifting mechanism enhancing the overall system efficiency and its economic outcomes. Indeed, the DRL control reduces dependency on grid electricity (both SCR and SSR indexes increase with DRL) and that implies a 4.73 MWh/year reduction of primary energy consumption, i.e. 3%. Additionally, user operative costs are reduced by 514 € (3.23%) compared with the RB system, Table 7 .

Overall, the DRL strategy enhances coordination and scheduling, effectively reducing both energy consumption and costs within the existing system constraints. The pricing schedule plays a crucial role in determining the cost savings reported above. The DRL controller can leverage temporal variations and fluctuations in electricity prices to optimize energy costs. Depending on the reward formulation, the DRL strategy has the potential to achieve greater percentage savings, though possibly at the expense of increased thermal storage temperature violations. However, it is essential to assess whether these temperature deviations impact the final performance of the AHU, particularly in maintaining supply air temperature within operational limits. This aspect is further examined in the scenario analysis presented in the following sections, where the trade-offs between economic performance and thermal comfort constraints are systematically evaluated.



(a)

(b)

Fig. 7. Monthly distribution of (a) temperature and (b) solar radiation, highlighting the period of scenario evaluation.

5.2. Comparison of results across different scenarios

To evaluate the scalability and adaptability of the DRL controller, it was deployed not only in the *Reference Scenario* on which it was trained, but it was also tested in other scenarios, as outlined in Section 3.4. Specifically two specialized SAC agents – one for heating and one for cooling – were deployed in all benchmark scenarios without additional training. The actor and critic networks retained their original weights and structure, as the core control task, including system configuration, state and action spaces, and reward function, remained consistent across scenarios. Only external factors such as weather, electricity prices, and PV availability changed.

In this regard, Fig. 6 presents the values of the *SSR* and *SCR* indexes for each of the scenarios analyzed. Note that only *Scenario 5* (i.e. PV fault scenario) was not represented because in this case the two indexes *SSR* and *SCR* are misleading. In particular, Fig. 6(a) refers to the selected winter period (15th Nov–15th Dec), while Fig. 6(b) illustrates the indexes for the examined summer period (15th Jun–15th Jul). Fig. 6 shows that, in all analyzed scenarios and both in winter and summer periods, the adoption of the DRL controller results in a simultaneous increase in *SCR* and *SSR* indexes.

These results stem from the DRL controller's ability to optimize renewable power utilization and thermal energy storage. In fact, the DRL activates the heat pumps in the hours of higher renewable production storing the cooling/heating energy. This enables stored energy to be utilized when renewable production is unavailable. As expected, this trend is more pronounced in summer operation (Fig. 6(b)), due to higher PV generation (Fig. 7(b)). Scenario 1 (Lampedusa) and Scenario 3 (Napoli) show increases of 2.3% and 0.7% in SCR, with SSR improvements of 3.8% and 4.1%, respectively. Despite Lampedusa's higher renewable generation, its greater cooling demand (Fig. 7(a)) limits the efficiency gains, as the DRL model was trained under Torino's climatic conditions.

In Scenario 4 (2022), which features higher electricity prices (Frilingou et al., 2023), the DRL controller prioritizes minimizing energy costs by predominantly activating the heat pump during lower-priced hours. As illustrated in Fig. 8, this strategy achieves a larger overall cost reduction. However, one motivation behind this outcome is that, in a high-price environment, the DRL places a stronger emphasis on cost minimization, overshadowing the objective of maximizing renewable self-consumption. Since these cheaper hours do not necessarily overlap with peak solar production, the potential for on-site renewable energy self-consumption is diminished, thereby compromising primary energy reduction.

During winter, in Fig. 6(a), the impact of DRL on SCR and SSR is more moderate, primarily due to lower solar radiation (Fig. 7(b)). The baseline scenario sees a 3.4% increase in SCR, while Bolzano experiences a 1.8% improvement. SSR improvements are limited (1.9% and 1.2%, respectively) because the reduced winter solar potential, limits how much of the total load can be covered even under optimal control.

Overall, the DRL controller enhances SCR and SSR in both seasons by aligning heat pump operation with PV generation peaks and utilizing thermal energy storage to shift energy use to more favorable conditions. This strategy reduces grid reliance and improves self-sufficiency, particularly in summer when higher renewable availability allows for more effective energy shifting.

Fig. 8 illustrates a comparison of energy costs between the DRL and RB systems across different scenarios. As explained in the preceding subsection, DRL uses thermal energy storage to activate the heat pump during periods of renewable energy production and low electricity prices, resulting in tangible economic benefits. Then, In the Reference

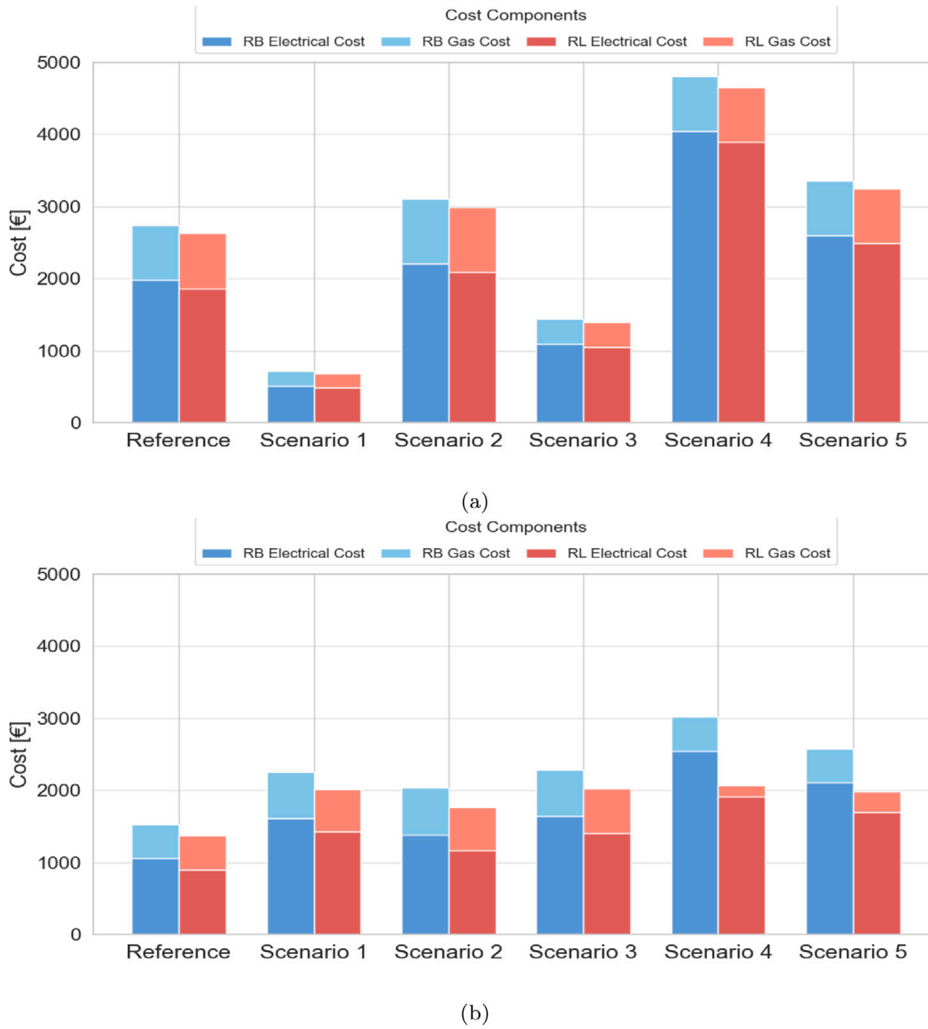


Fig. 8. Energy cost comparison between DRL and RB controller across scenarios in (a) winter period and (b) summer period.

Scenario, DRL achieves a 5.8% reduction in electricity costs (€1,980 to €1,866), while gas costs remain almost constant at €759 (0.03% increase), as shown in Fig. 8(a). Similar trends are observed in other locations:

- Lampedusa (Scenario 3): 5.3% lower electricity costs (€513 to €486).
- Bolzano (Scenario 2): 5.4% reduction (€2,212 to €2,093).
- 2022 high-price scenario (Scenario 4): Despite higher overall costs, DRL still yields a 3.9% electricity cost reduction (€4,052 to €3,893).

During summer (Fig. 8(b)), cost savings are more pronounced due to higher renewable availability. In the Reference Scenario, DRL achieves a 14. 4% reduction in electricity costs (€1,059 to €903 and gas costs remain stable at around €471 (0.05% increase). The most substantial improvements are seen in:

- Lampedusa (Scenario 1): 11.2% lower electricity costs and an 8.4% reduction in gas consumption.
- Bolzano (Scenario 2): 15.8% electricity cost reduction and 7.1% lower gas consumption.
- 2022 high-price scenario (Scenario 4): 24.5% reduction in electricity costs, with a 69.7% drop in gas costs.
- PV fault scenario (Scenario 5): 19.5% electricity savings, with gas costs reduced by 39.7%, demonstrating the resilience of DRL even under renewable generation failures.

Then, despite the fact that the DRL controller was trained under Torino conditions, it successfully achieved significant economic performance even when deployed in substantially different condition. The primary reason why gas costs remain nearly constant across all scenarios while electricity costs show significant reductions under the DRL control strategy lies in the different operational flexibility between the heat pump (electricity-driven) and the backup boiler (gas-driven). The DRL controller optimizes heat pump scheduling, leveraging thermal energy storage to shift electric demand to periods of higher renewable availability or lower electricity prices. However, gas consumption remains relatively stable because the backup boiler primarily fulfills two fixed functions: providing heat for the steam generator, which supplies the humidification system, and supplementing a secondary Tank (TK-D) used for post-heating in the summer period. The detailed description is provided in Section 2. In winter, the thermal storage tank supplies both pre-heating and post-heating, but gas consumption remains largely independent of HP operation. This is because the backup boiler is mainly responsible for generating steam for humidification, which is required regardless of the heat pump's operation. During summer, the gas boiler acts as a thermal booster, ensuring that the secondary heating tank (TK-D) is maintained within a fixed temperature range (e.g., 45 °C). This secondary tank is primarily heated by waste heat recovery from the desuperheating water loop of the chiller, offering an energy-efficient approach to temperature maintenance. The gas boiler is activated only when necessary to compensate for any shortfall in heat recovery, thereby ensuring that the outgoing water from TK-D consistently reaches the required setpoint temperature. The dependency on

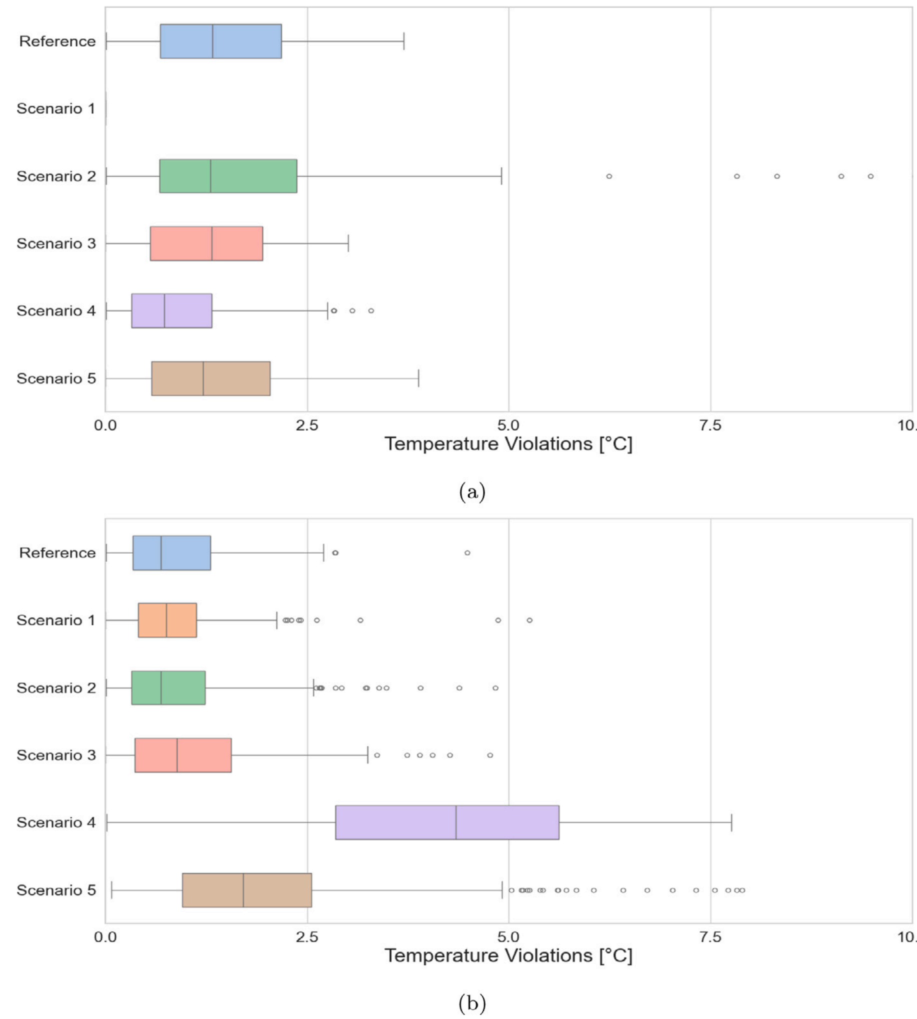


Fig. 9. Comparison of Tank Temperature violations across (a) winter scenarios and (b) summer scenarios.

gas consumption is directly influenced by the temperature of the water entering TK-D. When the incoming water temperature is already close to 45 °C, the demand for gas remains minimal. Conversely, if the water temperature is lower, the system requires additional gas input to compensate and achieve the desired setpoint temperature. Specially this last dynamics highlight the interplay between DRL-controlled thermal storage management and gas consumption, demonstrating how optimized tank temperature regulation can minimize reliance on gas-based heating while maintaining system performance.

This interaction between thermal storage, electricity and gas usage is further reflected in the distribution of tank temperature violations across different scenarios, as illustrated in Fig. 9.

In the winter scenarios (Fig. 9(a)), temperature violations, measured relative to the lower temperature limit, are generally well-contained, with most scenarios exhibiting violations within the range of 0–2.5 °C. Among them, Scenario 2 (Bolzano) displays the widest interquartile range, indicating greater variability in temperature deviations, whereas Scenario 4 (higher electrical price) shows the most compact distribution, suggesting more stable and consistent temperature control. Notably, Scenario 1, corresponding to the Lampedusa weather conditions, does not exhibit any temperature violations, indicating effective thermal regulation in this instance. This outcome is attributable, in part, to the reduced activation frequency of the heat pump for heating purposes, owing to the elevated outdoor temperatures. In contrast, the summer scenarios (Fig. 9(b)) present a significantly different pattern, with larger and more variable temperature violations, measured relative to the upper temperature limit. Although the median temperature violation remains below 1 °C in most cases, certain scenarios exhibit considerably larger deviations. Scenario 4 (higher electricity price) experiences the highest temperature violations, with a broad distribution extending beyond 6 °C. Similarly, Scenario 5 (PV fault) displays a wider range of violations compared to its winter counterpart, underscoring the increased difficulty in maintaining temperature constraints under these conditions. A key factor contributing to the cost savings observed in Scenarios 4 and 5 (Fig. 8(b)) could be the higher distribution of tank temperature violations (Fig. 9(b)). This pattern suggests that, on average, the tank temperature remains slightly higher than in other scenarios, which in turn reduces the frequency of on-off cycling for the heat pump, leading to lower electrical consumption.

This trend suggests that while the DRL controller enhances energy efficiency and cost savings during summer, it does so at the expense of more frequent and severe temperature constraint violations. Conversely, the Reference Scenario and Scenarios 1, 2 and 3 maintain relatively stable temperature control in both seasons, indicating that the more extreme conditions in Scenarios 4 and 5 pose particular challenges for effective thermal management during summer operations.

Table 8 presents the differences in energy performance between RB and DRL control strategies across various winter and summer scenarios. Positive values indicate reductions achieved by the DRL controller. The results highlight distinct seasonal trends in energy optimization. During winter, DRL effectively reduces grid electricity imports, primarily through intelligent temporal shifting rather than merely reducing

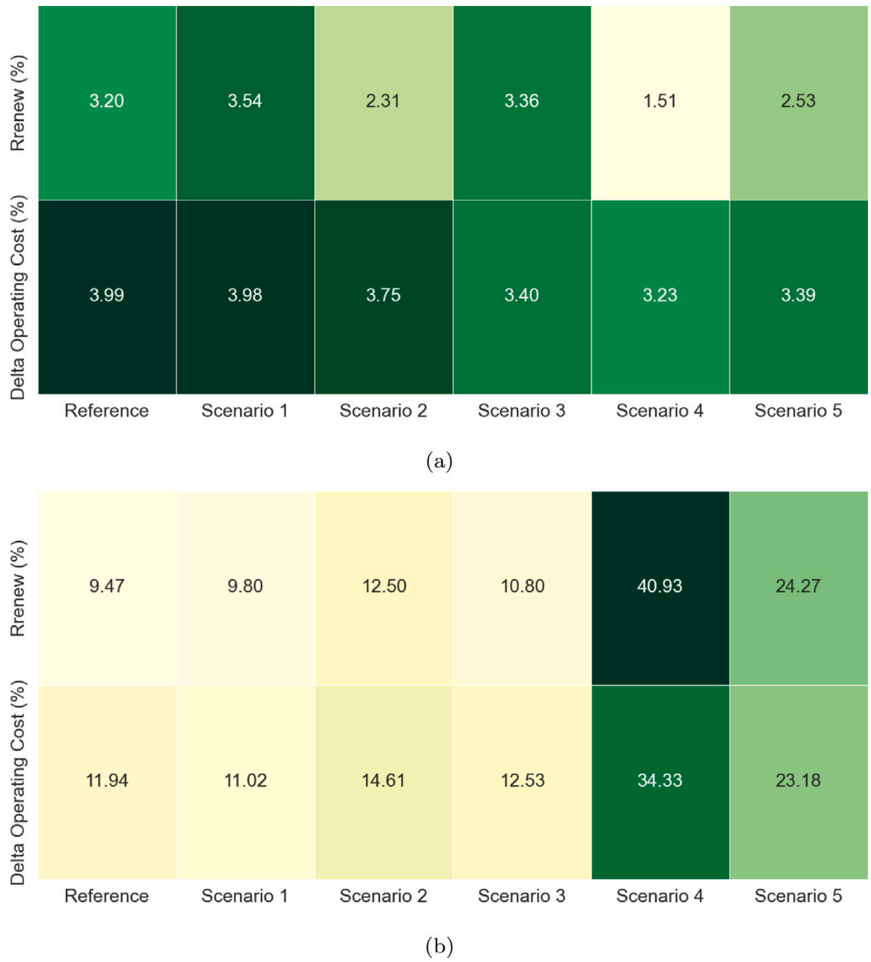


Fig. 10. Comparison of scenarios' delta metrics between RB and DRL in (a) winter period and (b) summer period.

Table 8
Percentage difference between RB and DRL in terms of Electrical Consumption, Energy From Grid, and Energy To Grid across various scenarios in (a) winter and (b) summer.

Scenario	$\Delta E_{el,LOAD}$	$\Delta E_{el,fromGRID}$	$\Delta E_{el,toGRID}$
Reference	2.28%	4.77%	9.68%
Scenario 1	0.57%	5.56%	2.76%
Scenario 2	2.07%	3.57%	6.12%
Scenario 3	1.45%	4.13%	4.89%
Scenario 4	0.72%	2.26%	7.25%
Scenario 5	3.37%	3.38%	0.00%

(a) Winter Period

Scenario	$\Delta E_{el,LOAD}$	$\Delta E_{el,fromGRID}$	$\Delta E_{el,toGRID}$
Reference	6.80%	14.88%	1.27%
Scenario 1	3.83%	10.59%	5.59%
Scenario 2	7.52%	15.78%	2.35%
Scenario 3	7.15%	13.96%	1.79%
Scenario 4	11.88%	24.52%	0.60%
Scenario 5	19.97%	19.93%	0.00%

(b) Summer Period

overall demand. This suggests that the algorithm prioritizes aligning energy consumption with photovoltaic availability to enhance grid independence.

In summer, the DRL strategy demonstrates even greater reductions in grid imports across all scenarios. The increased solar energy availability allows for more effective coordination between cooling operations and PV generation peaks, maximizing self-consumption and minimizing grid reliance. The ability of DRL to schedule heat pump operation in alignment with peak renewable production underscores its effectiveness in optimizing energy management throughout the year.

Fig. 10 presents a heatmap visualization comparing two key metrics across different scenarios: the R_{renew} coefficient and Daily Operating Cost between RB and DRL approaches for both winter and summer periods. The difference between the seasons is evident: during the winter period DRL exhibits relatively limited improvements across all scenarios, Fig. 10(a), with R_{renew} values ranging from 1.51% to 3.54% and operational costs fluctuating between 3.23% and 3.99%.

Conversely, during the summer period DRL exhibits notably improvement, with R_{renew} values reaching as high as 40.93% in Scenario 4 and 24.27% in Scenario 5, compared to baseline values around 9%–12% in other scenarios. Similarly, the daily operating costs show more substantial reductions in summer, particularly in Scenarios 4 and 5, where improvements of 34.33% and 23.18% respectively were achieved. This marked seasonal difference suggests that the DRL controller's advantages are particularly significant during summer operations: when it is better able to exploit the available renewable electricity, 10(b).

6. Conclusion

This paper presented a zero-shot transfer learning strategy for Soft Actor-Critic-based Deep Reinforcement Learning controllers managing an Air Handling Unit coupled with a Thermal Energy Storage tank in a university classroom building. A co-simulation framework integrating TRNSYS and Python was developed to evaluate the controller's performance. In the reference scenario, the DRL controller reduced primary energy consumption by 4.73 MWh/year and lowered annual operating costs by 3.2% compared to a rule-based baseline.

The main contribution of this work lies in demonstrating the zero-shot transferability of the developed DRL controllers. Without any retraining, the same policy was successfully deployed across a range of operational scenarios—spanning different climatic conditions, electricity tariff structures, and PV system failure cases. In these diverse contexts, the controller continued to perform effectively, with electricity cost reductions reaching 24.5% under high-price summer conditions and 15.8% in colder climates.

The controllers demonstrated strong adaptive capabilities, consistently aligning energy demand with external conditions such as electricity pricing and photovoltaic availability. In particular, they effectively leveraged the thermal inertia of the TES to shift HVAC operation to periods with lower electricity costs or higher on-site renewable generation. This behavior reflects advanced temporal optimization, enabling proactive and strategic scheduling of energy use that goes beyond the reactive nature of traditional rule-based control strategies.

While the controller demonstrated consistently strong performance, the study revealed a trade-off between cost optimization and TES temperature violations. In summer, the controller effectively prioritized PV utilization and reduced operational costs. In winter, with reduced solar availability, gains in self-consumption were more limited, but overall efficiency remained within acceptable limits. Under more challenging conditions – such as PV outages or extended periods of low electricity prices – sparse temperature violations were observed.

These findings underscore important considerations in multi-objective control and suggest several directions for future research and practical implementation:

- **Integration of Safe Reinforcement Learning**: Incorporating SafeRL techniques could enhance the management of safety constraints, ensuring that temperature violations, excessive energy demand, or system instability are minimized while maintaining energy efficiency. Constraint-aware DRL methods, such as Constrained Policy Optimization (CPO) or Lagrangian-based approaches, could improve the reliability of DRL-based control by explicitly handling hard constraints within the learning process.
- **Exploring Transfer Learning and Domain Adaptation**: investigate more advanced transfer learning strategies to accelerate adaptation to new buildings, different HVAC configurations, or even other types of energy storage systems. Domain adaptation techniques could improve the ability of pre-trained RL controllers to quickly generalize to unseen scenarios with minimal retraining.
- **Real-World Deployment and Validation**: while this study is based on co-simulation with TRNSYS and Python, future research should focus on real-world implementation and validation of DRL-based controllers in actual HVAC systems. Deploying the trained model in a physical building would provide valuable insights into practical challenges, system latency, sensor noise, and control feasibility in real-time operations.

CRediT authorship contribution statement

Giacomo Buscemi: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Francesco Paolo Cuomo:** Writing – review & editing, Validation, Software, Methodology, Data curation, Conceptualization. **Giuseppe Razzano:** Writing – review & editing, Methodology, Conceptualization. **Francesco Liberato Cappiello:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Silvio Brandi:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve the readability and the language quality of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was developed in the framework of the research activities carried out within the PRIN 2020 project: "OPTIMISM—Optimal refurbishment design and management of small energy microgrids", funded by the Italian Ministry of University and Research (MUR).

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

This appendix provides comprehensive technical data supporting the analysis presented in the manuscript. The information is organized into four categories: geometric characteristics of the building structure, thermal properties of the building envelope, internal heat gain parameters, and specifications of the heating, ventilation, and air conditioning (HVAC) system components.

The geometric data presented in Table A.9 establish the dimensional framework of the analyzed building, including volumetric and area measurements for the primary functional zones. Table A.10 details the construction characteristics of the building envelope, specifying thermal transmittance values and material thicknesses for key structural elements. Heat gain parameters for internal loads are documented in Table Table A.11, providing the basis for thermal load calculations. Finally, Table A.12 presents the technical specifications and performance characteristics of the HVAC system components that serve the building's thermal conditioning requirements.

Table A.9
Geometric data (Berta et al., 2017).

Parameter	Value	Unit
Classroom volume	1112.26	m^3
Classroom floor area	218.09	m^2
Corridor volume	1621.39	m^3
Corridor floor area	317.92	m^2
Sanitary facility volume	479.40	m^3
Sanitary facility floor area	94	m^2

Table A.10
Construction characteristics of building envelope.

Building element	Thickness [m]	U-value [W/m^{-2} K^{-1}]
Facades	0.30	0.260
Adjacent walls (between classroom and C-SF)	0.21	0.236
Flooring system[a]	1.00	0.265
Roofing system[a]	0.78	0.129
Window	0.004/0.016/0.004	1.06

[a] The flooring system and roofing system each consists of a series of horizontal layers that provide, respectively, a connection to the ground and to the external environment.

Table A.11
Heat gains parameters for the classrooms (Cammarata, 2003).

Type	Heat gain [W/person]	Heat gain [W/m^2]	Convective [%]	Radiative [%]	Humidity [kg/h person]
Students	115	–	40	60	0.066
Light	–	0.35	80	20	–

Table A.12
Main features of the HVAC system serving the classrooms.

Component	Parameter	Value	Unit
HP (NRG-H-0754 (Aermec, 2025))	Rated heating capacity	195.8	kW
	Rated power demand in heating mode	59.0	kW
	Rated COP in heating mode	3.32	–
	Rated total cooling capacity	190.4	kW
	Rated power demand in cooling mode	66.0	kW
	Rated COP in cooling mode	2.88	–
	Rated water flow rate (heating)	33 974	kg/h
	Rated water flow rate (cooling)	32 773	kg/h
DHN-HX	Rated heating capacity	109.0	kW
	Rated air mass flow rate	24.0	kW
	Rated water mass flow rate	24.0	kW
	Rated coil effectiveness	22 000	m^3/h
TK	Volume	12.0	m^3
TK-D	Volume	1.0	m^3
REC	Rated heating capacity	109.0	kW
	Rated total cooling capacity	24.0	kW
	Rated sensible cooling capacity	24.0	kW
	Rated fresh air flow rate	22 000	m^3/h
	Rated exhaust air flow rate	22 000	m^3/h
	Rated sensible effectiveness (pre-heating)	0.59	–
	Rated sensible effectiveness (pre-cooling)	0.54	–
HC	Rated heating capacity	74.0	kW
	Rated air flow rate	22 000	m^3/h
	Rated water mass flow rate	4200	kg/h
	Rated coil effectiveness	0.80	–
CC	Rated cooling capacity	183.38	kW
	Rated air flow rate	22 000	m^3/h
	Rated water mass flow rate	26 541	kg/h
	Rated coil effectiveness	0.74	–
H	Steam inlet temperature	100.0	° C
	Steam rated mass flow rate	90.0	kg/h
RC	Rated heating capacity	28.0	kW
	Rated air mass flow rate	5500	m^3/h
	Rated water mass flow rate	1600	kg/h
	Rated coil effectiveness	0.8	–
	Number of coils	4	–
SF (Claredot, 2025)	Rated air flow rate	55 440	m^3/h
	Rated fan power demand	13.57	kW
	Rated head	690	Pa
	Rated efficiency	0.78	–
RF (Claredot, 2025)	Rated air flow rate	55 440	m^3/h
	Rated fan power demand	13.57	kW
	Rated head	690	Pa
	Rated efficiency	0.78	–

Data availability

Data will be made available on request.

References

Adesanya, M.A., Obasekore, H., Rabiu, A., Na, W.-H., Ogunlowo, Q.O., Akpenpuun, T.D., Kim, M.-H., Kim, H.-T., Kang, B.-Y., Lee, H.-W., 2024. Deep reinforcement learning for PID parameter tuning in greenhouse HVAC system energy optimization: A TRNSYS-python cosimulation approach. Expert Syst. Appl. 252, 124126.

Aermec, 2025. URL https://global.aermec.com/it/industriale/prodotti/scheda-prodotto/?Code=NRG_2002_HP.

Barto, A.G., 1997. Reinforcement learning. In: Neural Systems for Control. Elsevier, pp. 7–30.

Berta, M., Rolfo, D., et al., 2017. Il progetto tra previsione e contingenza. Un tassello eterodosso nel masterplan del politecnico di torino architectural design between prediction and contingency. A heterodox tile in the. Atti Rass. Tec. (1-2-3-Dicembre 2017), 55–64.

Buonomano, A., Calise, F., Ferruzzi, G., 2013. Thermoeconomic analysis of storage systems for solar heating and cooling systems: A comparison between variable-volume and fixed-volume tanks. Energy 59, 600–616. http://dx.doi.org/10.1016/j.energy.2013.06.063, URL https://www.sciencedirect.com/science/article/pii/S0360544213005628.

Cacabelos, A., Eguía, P., Míguez, J.L., Granada, E., Arce, M.E., 2015. Calibrated simulation of a public library HVAC system with a ground-source heat pump and a radiant floor using TRNSYS and GenOpt. Energy Build. 108, 114–126. http://dx.doi.org/10.1016/j.enbuild.2015.09.006, URL https://www.sciencedirect.com/science/article/pii/S0378778815302474.

Calise, F., Cappiello, F.L., Cimmino, L., Dentice d'Accadia, M., Vicidomini, M., 2023b. Renewable smart energy network: A thermoeconomic comparison between conventional lithium-ion batteries and reversible solid oxide fuel cells. Renew. Energy 214, 74–95. http://dx.doi.org/10.1016/j.renene.2023.05.090, URL https://www.sciencedirect.com/science/article/pii/S096014812300722X.

Calise, F., Cappiello, F.L., Cimmino, L., Vicidomini, M., 2023a. Dynamic analysis of the heat theft issue for residential buildings. Energy Build. 282, 112790. http://dx.doi.org/10.1016/j.enbuild.2023.112790, URL https://www.sciencedirect.com/science/article/pii/S0378778823000208.

Cammarata, G., 2003. Illuminotecnica. Dipartimento Ing. Ind. Mecc. Università Catania.

Cammarata, G., 2016. Impianti Termotecnici - Volume IV condizionamento.

Cappiello, F.L., 2024. Energy and economic analysis of energy efficiency actions in surgery rooms: A dynamic analysis. Appl. Energy 373, 123887. http://dx.doi.org/10.1016/j.apenergy.2024.123887, URL https://www.sciencedirect.com/science/article/pii/S0306261924012704.

Cappiello, F.L., Erhart, T.G., 2021. Modular cogeneration for hospitals: A novel control strategy and optimal design. Energy Convers. Manage. 237, 114131. http://dx.doi.org/10.1016/j.enconman.2021.114131, URL https://www.sciencedirect.com/science/article/pii/S0196890421003071.

Chen, C., An, J., Wang, C., Duan, X., Lu, S., Che, H., Qi, M., Yan, D., 2023. Deep reinforcement learning-based joint optimization control of indoor temperature and relative humidity in office buildings. Buildings 13 (2), 438.

Claredot, 2025. URL https://www.claredot.net/it/sez_Aeraulica/potenza-assorbita-di-un-ventilatore.php.

Coraci, D., Brandi, S., Hong, T., Capozzoli, A., 2024. An innovative heterogeneous transfer learning framework to enhance the scalability of deep reinforcement learning controllers in buildings with integrated energy systems. In: Building Simulation. 17, (5), Springer, pp. 739–770.

Della Repubblica, P., 1993. Decreto del presidente della repubblica 26 agosto 1993, n. 412. Regolamento recante norme per la progettazione, l'installazione, l'esercizio e la manutenzione degli impianti termici degli edifici ai fini del contenimento dei consumi di energia. Available online: https://www.normattiva.it/uri-res/N2Ls.

di Unificazione, E.N.I., 1995. UNI 10339-impianti aeraulici al fini di benessere. Generalità, classificazione e requisiti. Regole per la richiesta d'offerta, l'offerta, l'ordine e la fornitura.

Diller, T., Soppelsa, A., Nagpal, H., Fedrizzi, R., Henze, G., 2024. A dynamic programming based method for optimal control of a cascaded heat pump system with thermal energy storage. Optim. Eng. 25 (1), 229–251.

Du, Y., Zandi, H., Kotevska, O., Kurte, K., Munk, J., Amasyali, K., Mckee, E., Li, F., 2021. Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning. Appl. Energy 281, 116117. http://dx.doi.org/10.1016/j.apenergy.2020.116117, URL https://www.sciencedirect.com/science/article/pii/S030626192031535X.

G. u. d. Europa, 2019. Raccomandazione (UE) 2019/786 della commissione dell'8 maggio 2018 sulla ristrutturazione degli edifici.

Feng, Z., Nekouei, E., 2024. A privacy-preserving framework for cloud-based HVAC control. IEEE Trans. Control Syst. Technol..

François-Lavet, V., Henderson, P., Islam, R., Bellemare, M.G., Pineau, J., 2018. An introduction to deep reinforcement learning. Found. Trends® Mach. Learn. 11 (3–4), 219–354. http://dx.doi.org/10.1561/2200000071.

Frilingou, N., Xexakis, G., Koasidis, K., Nikas, A., Campagnolo, L., Delpiazzo, E., Chiodi, A., Gargiulo, M., McWilliams, B., Koutsellis, T., Doukas, H., 2023. Navigating through an energy crisis: Challenges and progress towards electricity decarbonisation, reliability, and affordability in Italy. Energy Res. Soc. Sci. 96, 102934. http://dx.doi.org/10.1016/j.erss.2022.102934, URL https://www.sciencedirect.com/science/article/pii/S2214629622004376.

Fu, Y., Xu, S., Zhu, Q., O'Neill, Z., Adetola, V., 2023. How good are learning-based control v.s. model-based control for load shifting? Investigations on a single zone building energy system. Energy 273, 127073. http://dx.doi.org/10.1016/j.energy.2023.127073, URL https://www.sciencedirect.com/science/article/pii/S036054422300467X.

Gestore dei Mercati Energetici, 2024. GME – mercati dell'energia. URL https://www.mercatoelettrico.org/. (Accessed 6 June 2025).

Haarnoja, T., Zhou, A., Abbeel, P., Levine, S., 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: International Conference on Machine Learning. PMLR, pp. 1861–1870.

of Heating, R.A.S., Engineers, A.C., Atlanta, G., 2009. ASHRAE Handbook: Fundamentals. ASHRAE.

Henze, G.P., Schoenmann, J., 2003. Evaluation of reinforcement learning control for thermal energy storage systems. HVAC& R Res. 9 (3), 259–275.

International Energy Agency, 2023. Tracking Clean Energy Progress 2023. IEA, Paris, URL https://www.iea.org/reports/tracking-clean-energy-progress-2023.

ISO, B., 2017. 17772-1: 2017. Energy performance of buildings. Indoor environmental quality. Indoor environmental input parameters for the design and assessment of energy performance of buildings.

Jang, Y.-E., Kim, Y.-J., Catalão, J.P.S., 2021. Optimal HVAC system operation using online learning of interconnected neural networks. IEEE Trans. Smart Grid 12 (4), 3030–3042. http://dx.doi.org/10.1109/TSG.2021.3051564.

Klein, S., et al., 2017. TRNSYS 18: A Transient System Simulation Program. Solar Energy Laboratory, University of Wisconsin, Madison, USA, URL http://sel.me.wisc.edu/trnsys. (Accessed on 23 January 2025).

Krinidis, S., Tsolakis, A.C., Katsolas, I., Ioannidis, D., Tzovaras, D., 2018. Multi-criteria HVAC control optimization. In: 2018 IEEE International Energy Conference. ENERGYCON, pp. 1–6. http://dx.doi.org/10.1109/ENERGYCON.2018.8398747.

Kropas, T., Streckienė, G., Kirsanovs, V., Dzikevics, M., 2022. Investigation of heat pump efficiency in baltic states using trnsys simulation tool. Environ. Clim. Technol. 26 (1), 548–560.

Lissa, P., Schukat, M., Keane, M., Barrett, E., 2021. Transfer learning applied to DRL-based heat pump control to leverage microgrid energy efficiency. Smart Energy 3, 100044.

Lu, X., Fu, Y., O'Neill, Z., 2023. Benchmarking high performance HVAC rule-based controls with advanced intelligent controllers: A case study in a multi-zone system in modelica. Energy Build. 284, 112854. http://dx.doi.org/10.1016/j.enbuild.2023.112854, URL https://www.sciencedirect.com/science/article/pii/S0378778823000841.

Meteonorm, 2024. Meteonorm - global meteorological data. URL https://meteonorm.com/. (Accessed 6 June 2025).

Parliament, I., 2013. DPR n. 74/2013.

Pavirani, F., Gokhale, G., Claessens, B., Develder, C., 2024. Demand response for residential building heating: Effective Monte Carlo tree search control based on physics-informed neural networks. Energy Build. 311, 114161. http://dx.doi.org/10.1016/j.enbuild.2024.114161, URL https://www.sciencedirect.com/science/article/pii/S0378778824002779.

Politecnico of Turin, 2024. URL https://www.polito.it/.

Puterman, M.L., 1990. Markov decision processes. Handbooks Oper. Res. Management Sci. 2, 331–434.

Rashad, M., Żabnieńska-Góra, A., Norman, L., Jouhara, H., 2022. Analysis of energy demand in a residential building using TRNSYS. Energy 254, 124357. http://dx.doi.org/10.1016/j.energy.2022.124357, URL https://www.sciencedirect.com/science/article/pii/S0360544222012609.

Razzano, G., Brandi, S., Piscitelli, M.S., Capozzoli, A., 2025. Rule extraction from deep reinforcement learning controller and comparative analysis with ASHRAE control sequences for the optimal management of heating, ventilation, and air conditioning (HVAC) systems in multizone buildings. Appl. Energy 381, 125046. http://dx.doi.org/10.1016/j.apenergy.2024.125046, URL https://www.sciencedirect.com/science/article/pii/S0306261924024309.

Sutton, R., Barto, A., 1998. Reinforcement learning: An introduction. IEEE Trans. Neural Netw. 9 (5), http://dx.doi.org/10.1109/TNN.1998.712192, 1054–1054.

Team, T., 2023. Calling python from TRNSYS with CFFI. URL https://trnsys.de/static/77828438acd0697c30be234f0f248eff/Calling-Python-from-TRNSYS-with-CFFI.pdf. (Accessed 15 January 2025).

Wang, X., Kang, X., An, J., Chen, H., Yan, D., 2023. Reinforcement learning approach for optimal control of ice-based thermal energy storage (TES) systems in commercial buildings. Energy Build. 301, 113696.

Wei, T., Wang, Y., Zhu, Q., 2017. Deep reinforcement learning for building HVAC control. In: Proceedings of the 54th Annual Design Automation Conference 2017. pp. 1–6.

Xu, S., Fu, Y., Wang, Y., Yang, Z., Huang, C., O'Neill, Z., Wang, Z., Zhu, Q., 2025. Efficient and assured reinforcement learning-based building HVAC control with heterogeneous expert-guided training. Sci. Rep. 15 (1), 7677.

Xu, S., Wang, Y., Wang, Y., O'Neill, Z., Zhu, Q., 2020. One for many: Transfer learning for building hvac control. In: Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation. pp. 230–239.