POLITECNICO DI TORINO
Repository ISTITUZIONALE

Rule extraction from deep reinforcement learning controller and comparative analysis with ASHRAE control sequences for the optimal management of Heating, Ventilation, and Air

Original
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 / Razzano, G., Brandi, S., Piscitelli, M.S., Capozzoli, A.. - In: APPLIED ENERGY. - ISSN 0306-2619. - ELETTRONICO. - 381:(2025). [10.1016/j.apenergy.2024.125046]

Availability:
This version is available at: 11583/2995720 since: 2024-12-20T10:20:32Z

Publisher:
Elsevier Ltd

Published
DOI:10.1016/j.apenergy.2024.125046

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

19 November 2026

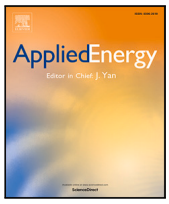
Applied Energy 381 (2025) 125046



Contents lists available at ScienceDirect

Applied Energy

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



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

Giuseppe Razzano, Silvio Brandi, Marco Savino Piscitelli, Alfonso Capozzoli *

Department of Energy (DENERG), TEBE Research Group, BAEDA Lab, Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin, 10129, Italy

ARTICLE INFO

Keywords:
Deep reinforcement learning
Rule extraction
Building energy management
Spawn of energyPlus
HVAC systems
Optimal control

ABSTRACT

The paper introduces a novel methodology for optimizing the operation of a centralized Air Handling Unit (AHU) in a multi-zone building served by VAV boxes with interpretable rules extracted from a Deep Reinforcement Learning (DRL) controller trained to enhance energy efficiency and indoor temperature control. To ensure practical application, a Rule Extraction (RE) framework is developed, translating the DRL complex decision-making process into actionable rules using decision trees. A multi-action approach is proposed by developing three different regression trees for adjusting the supply water temperature, the position of the chiller valve, and the position of the economizer damper of the AHU. The extracted rules are benchmarked against the original DRL controller and two conventional control sequences based on ASHRAE 2006 and ASHRAE Guideline 36 within a high-fidelity co-simulation architecture combining Spawn of EnergyPlus and Python. The co-simulation environment uses EnergyPlus for building envelope and loads while HVAC components and controls are implemented in the equation-based modeling language Modelica. Results show that the RE-based controller closely approximates the performance of the DRL policy with an electric energy consumption only 3% higher, highlighting its ability to effectively mirror a more complex control logic, representing a transparent and easily implementable alternative. The controllers based on ASHRAE 2006 and ASHRAE Guideline 36 lead to higher energy consumption (for both chiller and fan) and violations of indoor temperature compared to both RE-based control and DRL. This study underscores the potential of integrating AI-driven control methods with interpretable rule-based systems, facilitating the adoption of advanced energy management strategies in real-world building automation systems.

1. Introduction

HVAC systems represent a substantial portion of a building energy demand, yet they are crucial for maintaining a comfortable and healthy indoor environment for occupants. This has led to an increasing interest in developing advanced management strategies that can reduce their energy consumption without compromising the quality of the indoor environment.

The energy consumption and efficiency of HVAC systems in buildings is significantly influenced by the behavior of occupants, as well as their preferences for comfort and patterns of occupancy [1]. Fluctuations in comfort preferences and occupancy levels can result in varying thermal loads, which is particularly evident in multi-zone systems served by an Air Handling Unit (AHU). These fluctuations affect the operation of HVAC systems because they require continuous adjustments to maintain desired temperature and air quality levels, often leading to increased energy consumption.

The traditional approach to HVAC control relies on pre-defined rules and parameters and often is not adequate in facing spatial and temporal variations of thermal loads and occupancy patterns. This can lead to significant waste of energy and/or compromised comfort conditions.

In this context, advanced control technologies leveraging real-time data and predictive algorithms enable HVAC systems to dynamically adjust operations to current and future conditions. This improves energy efficiency, operational flexibility, and indoor environment quality by effectively responding to weather, occupancy, and user needs.

In multi-zone buildings, Variable Air Volume (VAV) control systems offer an optimal solution for efficiently managing varying thermal loads while maintaining consistent comfort levels across different

* Corresponding author.
E-mail address: alfonso.capozzoli@polito.it (A. Capozzoli).

https://doi.org/10.1016/j.apenergy.2024.125046
Received 25 September 2024; Received in revised form 18 November 2024; Accepted 29 November 2024
Available online 16 December 2024
0306-2619/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Acronyms

HVAC	Heating, Ventilation, and Air Conditioning
DRL	Deep Reinforcement Learning
IL	Imitation Learning
RE	Rule Extraction
SAC	Soft Actor Critic
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
A2006	ASHRAE 2006 control sequences
G36	ASHRAE Guideline 36 control sequences
AHU	Air Handling Unit
VAV	Variable Air Volume
FMUs	Functional Mock-up Units
FMI	Functional Mock-up Interface
BACS	Building Automation and Controls Systems
AI	Artificial Intelligence
XAI	eXplainable Artificial Intelligence
XRL	eXplainable Reinforcement Learning
SAT	Supply Air Temperature
DP	Difference Pressure
IAQ	Indoor Air Quality
RT	Regression Tree
KPI	key performance indicator
SWT	Supply water temperature
MPC	Model Predictive Control
DT	Decision Tree
ZAT	Zone Air Temperature
COP	Coefficient Of Performance
MAE	Mean Absolute Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
IQR	interquartile range

zones [2]. Standard controllers, such as those outlined in ASHRAE 2006 (A2006) [3] and ASHRAE Guideline 36 (G36) [4], are specifically designed for VAV systems in multi-zone buildings, providing frameworks for optimizing their sequences of operation. These improvements are related to supply air temperature reset, duct static pressure reset, and zone airflow control. Several studies demonstrated the potential of such control sequences in advanced co-simulation environments, which enable their detailed validation. The obtained results demonstrated an average of 31% HVAC energy savings in medium-sized commercial buildings when ASHRAE Guideline 36 is compared with simple rule-based controllers [5].

In this context advanced co-simulation environments enable the definition of highly detailed and accurate representations of real-life HVAC operations, allowing for comprehensive testing of control strategies before they are implemented in real buildings [6,7].

By integrating the strengths of different simulation platforms, co-simulation environments effectively simulate the dynamic interactions between various HVAC system components, making them particularly effective for evaluating complex control strategies [8].

For example, by exporting Modelica models [9] as FMUs [10], detailed dynamic models can be integrated into larger simulation frameworks, allowing for flexible and interoperable simulations. Modelica's object-oriented design simplifies the modeling of complex systems, such as HVAC components, while FMUs ensure seamless integration and extensibility across different simulation tools. The development of frameworks such as BOPTEST (Building Optimization Performance Test) further exemplifies the utility of advanced simulation environments. As highlighted in [11], BOPTEST offers a robust platform for simulation-based testing of advanced control strategies, enabling early performance evaluation, benchmarking against state-of-the-art methods, and practical deployment insights. By reducing implementation costs and verifying performance, it fosters trust among control vendors, building owners, and operators, supporting the adoption of innovative strategies.

Another relevant example, is the co-simulation environment Spawn of EnergyPlus (Spawn) [12] that serves as a valuable tool for bridging the domains of building energy modeling (BEM) and control workflows. The tool reuses EnergyPlus modules for lighting, building envelope, and loads, while re-implementing the HVAC and controls modules in the equation-based modeling language Modelica [9]. This approach enables the execution of fully dynamic, state-based simulations, thereby facilitating the direct simulation of physical control sequences and the estimation of consumption and savings that closely align with reality. The advancements achieved in the field of co-simulation paved the way to a more reliable performance assessment of innovative control solutions for HVAC systems based on the exploitation of Artificial Intelligence (AI). Particularly, DRL offers promising solutions for enhancing HVAC system control, leveraging advancements in co-simulation and testing frameworks. Co-simulation environments address the challenges of training DRL controllers, which require large, high-quality datasets often difficult to obtain from real-world systems.

Moreover, DRL training involves balancing exploration and exploitation, where the algorithm must test various actions to learn effective strategies. This process, however, can result in suboptimal or unsafe actions, making real-world training unsuitable due to safety and comfort concerns. On the other hand, the main benefit of DRL algorithms, is that they can learn optimal control policies through interaction with the environment (real or simulated), continuously improving performance based on feedback [13]. This adaptive capability allows DRL-based controllers to adjust system parameters in response to real-time data, thus achieving significant improvements in both energy efficiency and thermal comfort [14]. DRL demonstrates significant potential for optimizing HVAC systems, achieving notable reductions in energy costs and improving comfort levels when compared to traditional rule-based and model-based strategies [15–18].

However, some challenges need to be addressed to fully exploit the benefit offered by DRL-based controllers. The first challenge is related to their in-field deployment. By now, most of the advancements are mainly validated in co-simulated environments [19,20] with few examples in the literature where DRL has been used in real-world implementations [21–24].

The work presented in [25,26] demonstrates that while DRL can effectively adapt to dynamic energy systems and price signals, further research is necessary to ensure a robust and stable performance in real Building Automation and Controls Systems (BACS) implementations. Together with the need for stability and robustness of DRL controllers when deployed, another challenge is related to the perception that the human users have about the implementation of such advanced solutions. While the benefits of DRL controllers, such as energy savings and comfort improvements, can be demonstrated through simulations, their decision-making process for selecting optimal control actions remains complex and opaque. This complexity can limit their acceptance among HVAC professionals, who require interpretability and validation from a physics-based perspective to trust the proposed control strategies.

To address these challenges, this study explores a rule extraction process as a potential solution for the robust deployment of advanced DRL control strategies, aiming to enhance their transparency, interpretability, and professional acceptance.

Specifically, RE is an approach within the field of Explainable Artificial Intelligence (XAI) and in this application involves surrogate models to extract understandable rules from a DRL controller.

Surrogate models are simpler, more interpretable models that approximate the behavior of the more complex DRL controller. By analyzing the surrogate models, it is possible to derive explicit rules that explain the decision-making process of the DRL controller. This approach contributes in bridging the gap between advanced DRL techniques and the need for transparency and interpretability in practical HVAC applications.

In line with the aims of this paper, the next Section 1.1 reviews and examines the existing literature on the application of RE processes. Section 1.2 then presents and explores the contributions of this study along with the innovative elements it introduces.

1.1. Related works

The need for transparency and interpretability in Artificial Intelligence (AI) has led to significant research in eXplainable Artificial Intelligence (XAI) and RE strategies to provide explanations for the predictions, recommendations, and decisions of intelligent systems [27–29]. RE belongs to the group of post-hoc XAI procedures [30] and it is a process used to derive a set of understandable rules from a trained model. The extracted rules are typically in the form of logical statements, such as IF-THEN rules, which can be easily interpreted by humans. RE helps in validating the reliability of models, especially in safety-critical systems, by providing insight into how decisions are reached [31].

In the domain of building and power system control, RE from complex models is crucial for enhancing interpretability and practical implementation. The framework in [32] used eXplainable Reinforcement Learning (XRL) to optimize control strategies for a parallel cooling system in an office building. By combining deep Q-learning with decision trees, the authors demonstrated that the simplified rule-based control maintained comparable performance to the original complex strategy, with only a 1.2% difference in energy savings.

Similarly, [33] addressed the black-box nature of DRL in power system emergency control by proposing a policy extraction framework. Using an information gain rate-based weighted oblique decision tree (IGR-WODT), the study provided a transparent alternative to DRL models for scenarios such as under-voltage load shedding. This approach improved decision-making transparency and ensured the rule-based controller performed effectively on edge devices with limited computational resources.

In [34], a simulation-based framework optimized dedicated outdoor air systems using a genetic algorithm followed by rule extraction via decision trees. This approach achieved significant energy savings and reduced control complexity, with extracted rules reducing energy costs by 13% and energy consumption by 25%, closely matching the optimal control outcomes. Similarly, [35] used a mixed-integer genetic algorithm to optimize operational parameters across varying climate zones, occupancy, and envelope scenarios. By employing a Decision Tree (DT)-based RE method, the impacts of these variables was evaluated and practical operational rules were extracted.

The study in [36] examined the evolution of intelligent building control strategies by extracting near-optimal rule sets from a database of non-dominated solutions, employing multi-objective Model Predictive Control (MPC) on EnergyPlus models. The study demonstrated that the rule sets, derived from the MPC controller, were able to achieve up to 97% of the energy savings and 92% of the cost savings achieved by the original, more complex control policy, while still maintaining comparable levels of thermal comfort and peak electrical demand.

Similarly in [37] control rules for smart glazing were extracted through a decision tree algorithm from an optimal control strategy developed using an ideal MPC. The performances achieved by the MPC and the extracted rule set resulted very similar (differences in the order of 1%) despite, over a year of simulation, the rules were able to reproduce the control signal of the MPC with an accuracy in the range of 60%–65%. These results demonstrated the effectiveness of RE strategies in mimic sophisticated control logics ensuring complexity reduction without losing key information.

In the study proposed by [38] a detailed MPC algorithm using inverse models was implemented in 27 rooms of an institutional building to provide data for a classification learning approach. Decision trees for cooling and heating seasons were generated based on the inputs and outputs of the detailed MPC algorithm. The study found that during the cooling season, energy savings were 42% with MPC and 27% with RE, while during the heating season, energy savings were 18% with MPC and 33% with RE.

As a further example authors in [39] developed MPC controllers for optimizing window operation in mixed-mode buildings using EnergyPlus, demonstrating potential cooling energy savings of over 40% through night cooling strategies. A complementary statistical technique used multi-logistic regression to replicate MPC results, achieving 70%–90% of the original controller's energy savings with much lower computational costs.

The studies above discussed collectively emphasize the potential of RE in enhancing different aspects pertaining to optimal control in buildings. The most relevant advantages, retrieved from the literature can be summarized as follows:

- Transparency: RE enhances the interpretability of complex controllers such as DRL and MPC by translating their decision-making processes into understandable rules, making it easier for HVAC professionals to trust and validate these strategies.
- Ease of Implementation: Extracted rules can be easily implemented in existing building control systems, allowing for the benefits of advanced control strategies without the need for extensive computational resources.
- Real-Time Application: The simplicity of the extracted rules allows for real-time application in direct digital control systems, ensuring efficient and optimal building operation without the complexity of real-time DRL or MPC computations.
- Generalization: RE helps in creating control rules that can have traits of generalization for being applied across different zones or buildings, preventing overfitting to specific conditions and ensuring broader applicability.

In this perspective, the following section outlines the primary contributions and the novel elements this research seeks to bring towards the development of an AI-powered rule-based controller for the management of VAV systems in multizone buildings.

1.2. Novelty and motivation

From the analysis of the current scientific literature, RE emerges as a promising research direction to enhance the scalability and interpretability of advanced control strategies for HVAC systems in buildings.

Among advanced control strategies, DRL presents some interesting features since it does not require the definition of a control oriented model or the direct formalization of an optimization problem enabling a more flexible and scalable application. However, the application of DRL in real world context still face issues related to the amount of time required to converge to near optimal solutions, potential instability of the learned policy and the opaque nature of neural-network-based approaches.

The application of RE methods to convert complex DRL controllers into interpretable rules is a novel approach in the context of HVAC control. This method bridges the gap between high-performing, yet opaque, DRL algorithms and the need for understandable and actionable insights.

In this context, benchmarking DRL and RE controllers against widely adopted rule-based standards, such as ASHRAE 2006 and ASHRAE guideline 36, provides a robust framework for assessing their benefits. These control sequences are foundational in the HVAC

industry and extensively used in research and practice to evaluate energy efficiency and control strategies. Moreover, if benchmarks are evaluated through simulations, the adoption of advanced and detailed simulators and co-simulation frameworks represent a fundamental aspect to consider. To this purpose tools such as Spawn of EnergyPlus represent a significant advancement in the pursuit of achieving simulations that are as closely aligned with reality as possible. In such context, the contributions of the present paper can be summarized as follows:

- Conceptualization of a DRL-based controller, exploiting SAC algorithm, and its application for a centralized AHU in a multi-zone building served by VAV boxes. Specifically an hybrid approach is followed for the definition of the control logic, operating a DRL controller in conjunction with standard control sequences (i.e., ASHRAE 2006). This collaborative approach enabled the system to benefit from advanced decision-making for the control of supply air temperature at AHU level without disrupting the stability provided by the conventional controller at each VAV box level.
- Definition of a RE framework to extract decision rules that mimic the developed DRL controller. The rule extraction is performed by using a decision tree algorithm. As innovative aspect a multi-action approach is followed by developing three different decision trees for adjusting the supply water temperature, the position of the chiller valve, and the position of the economizer damper of the AHU. This approach provides a more granular, efficient, and realistic way to manage an AHU compared to directly set optimal values of the supply air temperature without an explicit control at component level.
- Introduction of a robust benchmark of the proposed solutions (i.e., the DRL and the RE controllers) against traditional yet well-performing baseline controllers following the control sequences suggested in ASHRAE 2006 and ASHRAE Guideline 36. In the literature, RE controllers are typically compared only to the control policies they are designed to mimic. However, in this study a broader comparison with established reference control sequences is carried out, providing a more comprehensive understanding of the added value potentially offered by the proposed approach.
- The implementation of an high-fidelity simulation model of building and HVAC system leveraging an advanced co-simulation architecture combining the simulation tool Spawn of EnergyPlus and Python.

In this context, the study aims to evaluate the effectiveness of DRL-based HVAC control and the feasibility of rule extraction methods in translating advanced control policies into practical and ready-to-implement controllers.

The structure of the paper can be summarized as follows: Section 2 presents the case study, providing context and details of the HVAC system and building under consideration. Section 3 outlines the methodology, including the experimental design, the exploited data, and control logics, detailing the operation of the DRL-based, RE-based, and baseline controllers. Section 4 presents the results, highlighting the performance of the DRL-based controller, the RE-based controller, and baseline controllers. Finally, Section 5 discusses the implications of the findings, and Section 6 concludes with the potential benefits and future research directions.

2. Case study

In this section the analyzed case study is introduced and described in detail. Specifically, the main features of the building and its HVAC system are reported together with specifications on the setting of the co-simulation environment developed to conduct the experiments.

Table 1
Description of building features.

Building feature	Value	Unit
N° of thermal zones	5	[–]
Conditioned floor area	511	[m^2]
Conditioned volume	1559	[m^3]
Transparent/opaque envelope vertical surface ratio	0.27	[–]
Opaque envelope vertical surface	221.80	[m^2]
U-Value Wall	0.78	[W/m^2K]
U-Value Roof	0.20	[W/m^2K]
U-Value Foundation	1.85	[W/m^2K]
U-Value Window	3.24	[W/m^2K]

2.1. Building overview and HVAC system configuration

The building selected as a case study was taken from the U.S. Department of Energy's Commercial Reference Buildings [40]. The building has a simple office configuration organized into five distinct conditioned zones, as shown in the 3D representation reported in Fig. 1.

Details about the geometry of the building and its envelope thermo-physical properties are reported in the Table 1. A value of 24 °C is set as indoor air temperature setpoint during cooling season, which is kept constant in all five thermal zones during the occupancy period. The system operation schedule and building occupancy patterns are defined as follows:

- On non-working days (Saturday and Sunday), the building is considered unoccupied and the HVAC system is turned off.
- On working days (Monday to Friday) the building is considered to be occupied from 08:00 to 19:00. To ensure indoor optimal conditions, the HVAC system is turned on two hours before the expected arrival time of occupants (at 06:00) and remains in operation until 19:00.

Fig. 2 shows a schematic representation of the system under study. The building is equipped with a comprehensive air conditioning system designed to maintain optimal thermal comfort across multiple zones. This system includes a heat generation component consisting of a chiller. The air conditioning system incorporates an Air Handling Unit (AHU) that features an economizer, a heating coil, a cooling coil, a fan, and five VAV boxes. However, since this study focuses exclusively on the cooling season, the heating components and their associated controls are not considered in the following descriptions.

The system operates through six different control signals. The *Economizer Damper Signal* regulates the position of the outdoor air damper and the return air damper, adjusting the mass flow rates of outdoor air ($\dot{m}_{out}$) and recirculated indoor air ($\dot{m}_{ret}$), respectively. The economizer's primary objective is to control the temperature of the mixed air (T_{mix}) by considering both the outdoor air temperature (T_{out}) and the return air temperature (T_{ret}). The total mass flow rate, $\dot{m}_{tot}$, is determined by the fan speed, which is governed by the Fan RPM Signal.

After passing through the economizer, the mixed air flows directly to the cooling coil, bypassing the heating coil, which is not considered in this study. The cooling demand is met by the chiller, and the supply water temperature (T_{swt}) of the chiller is controlled by the *Chiller SWT Signal*. The water mass flow rate ($\dot{m}_{water}$) to the cooling coil is modulated by a valve, whose position is determined by the *Cooling Coil Valve Signal*. Once the supply flow air passes through the cooling coil, it is moved by a fan through the ductwork to the various zones within the building. In each zone, the position of the damper in the VAV box is managed by the *VAV Damper Signal* to regulate the discharge air mass flow rate ($\dot{m}_{dis}$) according to the indoor air temperature setpoint.

2.2. Setup of the co-simulation framework

In the developed simulation environment, the building is modeled using Energy Plus 9.6.0 [41] while the HVAC system and its related

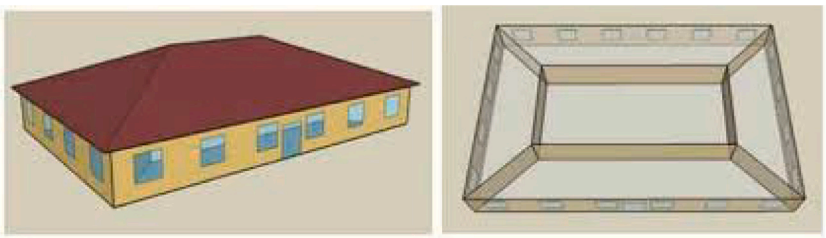


Fig. 1. Building configuration and considered thermal zones.

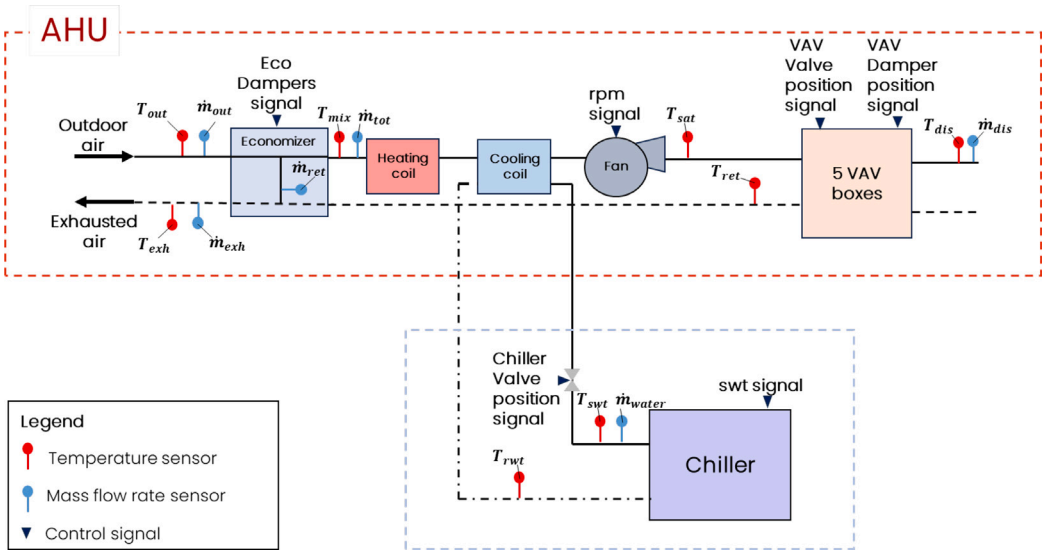


Fig. 2. Schematic representation of the HVAC system.

components are modeled using the Modelica language in the OpenModelica open-source platform [9]. Specifically, the tool Spawn of Energy Plus [12], with the Buildings library 9.0.0 [42], made it possible to connect the Modelica environment and Energy Plus. This integration allows for data exchange between Energy Plus and OpenModelica, by means of the Functional Mock-up Interface (FMI) 2.0 standard [10]. The co-simulation framework was managed entirely through Python, using the FMI standard and the pyfmi package [43]. The FMI standard provides guidelines for packaging and exchanging simulation models in a standardized Functional Mock-up Units (FMUs) format. More specifically, Python was used as the master for the loading, execution and real-time interaction of FMUs, which encapsulate individual components within the building and HVAC system model. Upon this simulation environment the DRL controller was implemented in Python, using the OpenAI Gym framework [44] that allows, through a loop operation, to take actions in the environment, observe the results, and update the DRL agent policy. Once all the controllers (i.e., DRL, RE, A2006, and G36) are defined, they are implemented using simulated real-time data. These controllers make decisions based on predefined/pre-trained control logic and send control signals to the HVAC system simulation model through FMUs. The co-simulation was conducted over a reference period of one month, specifically from July 1st to July 31st, using the weather conditions of the municipality of Turin in northern Italy. For the sake of clarity, Fig. 3 shows the working principle of the employed co-simulation framework.

3. Methodology and methods

Given the co-simulation environment previously introduced, this section explains the main methodological steps behind the development of the DRL controller, the implementation of the RE process and the benchmarking of the tested controllers in terms of energy consumption and indoor temperature violations pertaining to the simulated reference period.

The Fig. 4 presents the methodological framework.

As a first step the control sequences suggested in ASHRAE 2006 (A2006) and ASHRAE Guideline 36 (G36) were individually implemented in the simulation environment to establish baselines for comparison. The results obtained through those simulations provided a solid foundation for benchmarking analysis with the DRL and RE controllers.

The second step was devoted to the development of the DRL controller. As previously discussed, the DRL-based controller was designed to be implemented at AHU level while the VAV boxes were operated following the ASHRAE 2006 control sequences. This hybrid configuration allowed both controllers to operate concurrently, making the system to benefit from advanced decision-making without disrupting the stability provided by the conventional controller. In particular the DRL-based controller was designed to optimize the control of the Supply Air Temperature (SAT) within the AHU by adjusting the economizer damper position, chiller valve position, and the supply water temperature. A key aspect of the DRL controller design is its ability to manage AHU-related actions efficiently, without expanding the action space as the number of VAV boxes and zones to be served by the AHU system, increases. This feature is critical for maintaining effective control without unnecessary complexity. Therefore, the operation of other components was performed by following the ASHRAE 2006 control sequences, which allowed for controlling fan speed and the damper position for each VAV box in the building. For the sake of clarity, the control actions taken at AHU and VAV box level are summarized in Table 2 with specification of the involved controller.

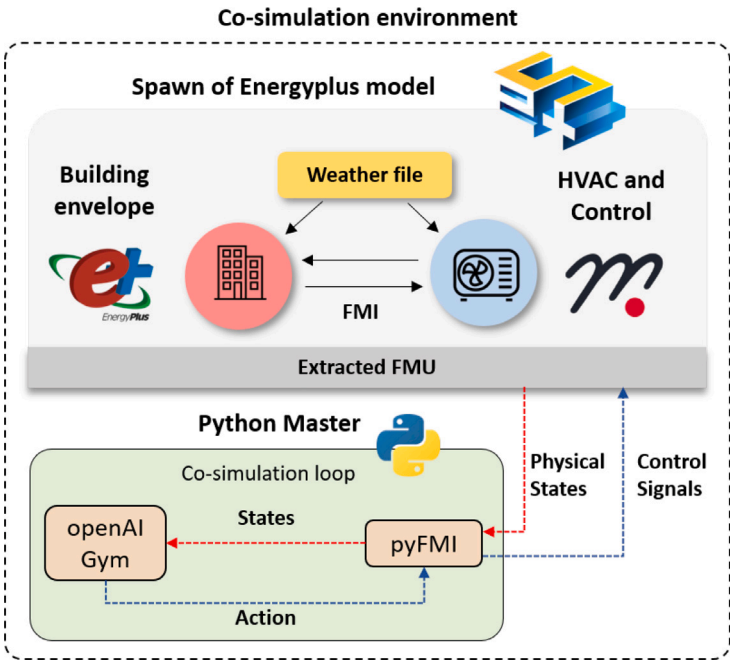


Fig. 3. Employed co-simulation framework.

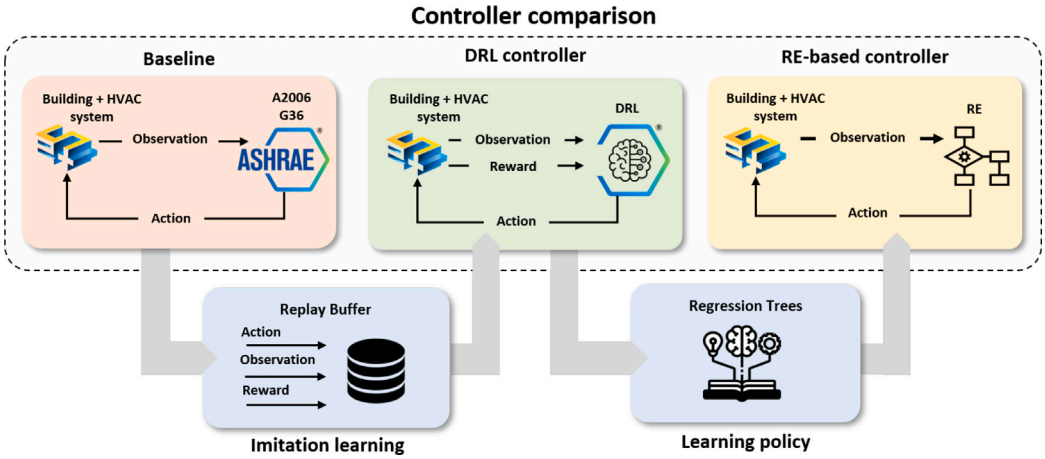


Fig. 4. Methodological framework.

Table 2
Control actions taken at AHU and VAV box level with specification of the involved controller.

Action	Managed by
Economizer damper position	DRL-based controller
Chiller valve position	DRL-based controller
Supply water temperature	DRL-based controller
Fan speed	A2006 control sequences
VAV box damper position	A2006 control sequences

Once the control problem was formulated, the DRL agent was trained by interacting with the simulation environment. However, instead of starting a trial-and-error learning of the control policy from scratch, the DRL agent was preliminary initialized performing an Imitation Learning (IL) process [45]. To this purpose data tuples coming from the simulations of the baseline control strategy ASHRAE 2006 are considered. Those tuples consist of state–action pairs and their resulting outcome capturing what action was taken by the controller when the environment was in a particular state. The outcome refers to the results or consequences of taking a particular action in a given state. The tuples are then collected and stored in a replay buffer or memory, to train the DRL agent, and updating its understanding of which actions are beneficial in specific states based on the rewards received, gradually refining the policy to maximize cumulative rewards over time.

As a consequence, the DRL controller learns from the baseline control strategy ASHRAE 2006 gaining an initial understanding of effective control of the AHU that meet ASHRAE standards. This starting phase of IL enabled the DRL controller to subsequently develop and optimize its strategies through further reinforcement learning, thereby improving its performance beyond the baseline strategy [45,46].

The next step of the methodological process, aimed to extract a set of IF-THEN rules from the simulation of DRL implementation in order to mimic its control policy in the most accurate way as possible. The IF-THEN rules were extracted through the development of three Regression Tree (RT) models i.e., one decision tree for each control action that the DRL controller can take.

Eventually, a comprehensive comparison among all the tested controllers (A2006, G36, DRL and RE) is conducted in order to assess and benchmark their performance through a set of key performance

indicator (KPI)s. By means of this analysis it was possible to understand the benefits associated to the DRL-based controller in reducing energy consumption and indoor air temperature violations respect to the baselines and at the same time assess the performance loss of the RE-based control respect to the target DRL control policy.

3.1. Description of the employed control strategies

This section provides a detailed overview of the four control strategies implemented in the case study. Specifically, it explains the control sequences used for the analyzed HVAC system under the ASHRAE 2006 control sequences (A2006) and ASHRAE Guideline 36 control sequences (G36), while discussing the algorithms employed by the DRL and RE-based controllers.

ASHRAE 2006. The A2006 standard introduced comprehensive control strategies aimed at optimizing the operation of HVAC systems, particularly through the implementation of VAV control sequence, as outlined in the "Sequences of Operation for Common HVAC Systems" [3]. These strategies encompass control sequences for supply and return fans, economizer dampers, VAV boxes (including valves and dampers), and zone control:

- The supply fan speed is regulated according to the static pressure of the duct. The duct static pressure is adjusted so that at least one VAV damper is 90% open. This strategy optimizes the distribution of airflow, reduces energy consumption, and ensures proper ventilation throughout the building by maintaining a desired pressure setpoint.
- The economizer dampers are modulated to follow the dry bulb temperature setpoint of the mixed air. The objective is to ensure that a minimum outside air flow rate is maintained.
- In each zone, the VAV damper is adjusted to achieve the desired room temperature in both cooling and heating mode.
- A finite state machine is responsible for regulating the operational mode of the HVAC system. This machine transitions the system between the following operation modes: occupied, unoccupied, off, unoccupied night setback, unoccupied warm-up, and unoccupied pre-cool.

To provide a comprehensive overview, the A2006 includes several additional functions to enhance the performance of HVAC systems. Frost protection serves to prevent the freezing of coils and other components in cold conditions by maintaining a minimum temperature in critical areas of the HVAC system, thereby ensuring efficient operation even in low-temperature environments. Furthermore, the standard specifies minimum outdoor air requirements to guarantee sufficient fresh air intake, which is important for maintaining Indoor Air Quality (IAQ) and complying with ventilation standards. Additionally, supply air cooling through economizing systems leverages outdoor air for cooling when conditions are favorable, thereby reducing the reliance on mechanical cooling and lowering energy consumption.

ASHRAE Guideline 36. Guideline 36 provides enhanced control sequences for VAV systems aimed at optimizing energy efficiency, comfort, and system performance. The main difference with A2006 is represented by the Trim& Respond control strategy. The T&R system is a dynamic control mechanism designed to facilitate continuous adjustment of HVAC system parameters with the objective to optimize performance and energy efficiency. In the "Trim" phase, the control system gradually lowers (or trims) the setpoint. The idea is to reduce the setpoint to the lowest possible value that still meets the needs of the most demanding zone. This minimizes energy consumption, as the fan do not have to work as hard to maintain a higher static pressure, or the cooling system do not need to work as hard to produce unnecessarily cooled air. The system continuously monitors all the zones served by the VAV system, identifying the "critical zone". This is typically the zone where the VAV damper is most open, indicating that it is the most difficult to satisfy in terms of airflow or temperature control. The setpoint is trimmed as long as the critical zone remains satisfied. If, at any point, the critical zone cannot be satisfied the system enters in the "Respond" phase. In the Respond phase, the setpoints are gradually adjusted to deliver more air by increasing the static pressure setpoint or to provide warmer or cooler air by raising or lowering the supply air temperature setpoint, until the critical zone is once again satisfied. The two main setpoint reset strategies are in the following explained:

- SAT Reset: The SAT is dynamically adjusted based on the outdoor air temperature (OAT) and setpoint requests from zone terminals to balance fan and cooling energy consumption as shown in Fig. 5(a). Specifically, the SAT setpoint is adjusted from the minimum cooling SAT (Min_ClgSAT) when the OAT is at its maximum (OAT_Max) and increases proportionally to the maximum SAT (T-max) as the OAT decreases to its minimum (OAT_Min). T-max is further refined using (T&R) logic based on zone-level reset requests that occur when the system detects significant zone temperature deviations from setpoint or high activity in the cooling loop. Additional reset requests are sent if the zone temperature exceeds the setpoint for an extended period of time, and requests continue until the cooling loop activity decreases, ensuring efficient temperature control and energy use.
- Static DP reset: The static DP setpoint is dynamically adjusted based on damper position opening requirements of VAV boxes (Fig. 5(b)). The airflow control logic works by monitoring the actual airflow relative to the setpoint airflow and damper position, ensuring that the system dynamically adjusts its response based on airflow discrepancies and damper positions to maintain optimal airflow. If the airflow deviates significantly from the setpoint and the damper is nearly fully open, the system sends multiple requests to resolve the discrepancy. The number of requests decreases as the severity of the deviation decreases. This approach ensures that the system tends to maintain the minimum static pressure while effectively responding to increasing demand from the zone terminals.

Conversely, according to ASHRAE 2006, SAT and the static DP setpoints are constant and defined according to operating schedules.

The control logic for a VAV reheat terminal unit adjusts damper and valve positions to maintain optimal airflow and temperature based on zone status. In cooling mode, the system modulates airflow between minimum and maximum setpoints and disables the heating coil unless the discharge air temperature drops below 10 °C. In deadband mode, the airflow is set to a minimum and the heating coil is disabled unless the discharge temperature is too low. The sequences of controlling damper and valve position for VAV reheat terminal unit are described in Fig. 6.

The economizer damper control system is designed to dynamically adjust the damper positions to optimize the use of outside air for cooling purposes and to ensure adequate ventilation. First, the system calculates the minimum and maximum outdoor air damper positions based on the specific requirements and functions of the economizer. This is achieved through the implementation of a strategy that may entail airflow sensing, differential pressure sensing, or a combination of both approaches to the damper. The system activates or deactivates the economizer based on a number of factors, including the outdoor temperature, enthalpy (if applicable), status of the supply fan, frost protection level, and zone status. This ensures that the economizer operates under favorable conditions. Ultimately, the positions of the outside air and return air dampers are modulated based on the SAT setpoint loop.

The valves of the heating and cooling coils and the Supply water temperature (SWT) are regulated according to the same strategies for both A2006 and G36.

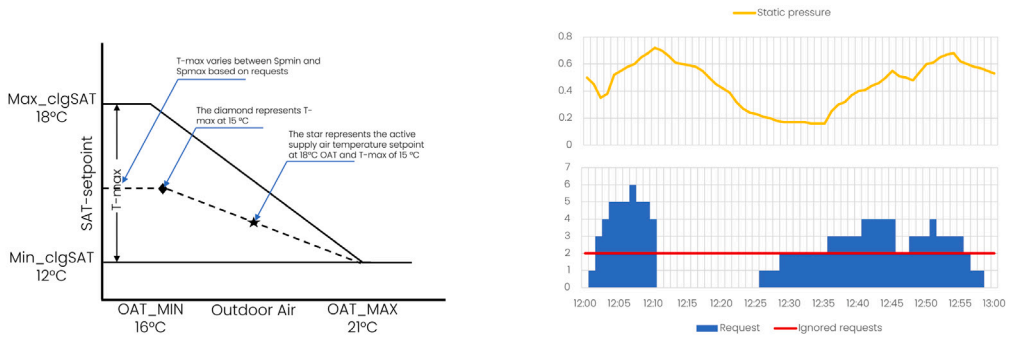


Fig. 5. Trim-and-respond control strategies from [4] for the reset of Supply Air Temperature (SAT) based on outside air conditions driven by terminal box cooling requests (a) and the reset of Static pressure (DP) setpoint based on VAV box damper positions and driven by pressure requests (b).

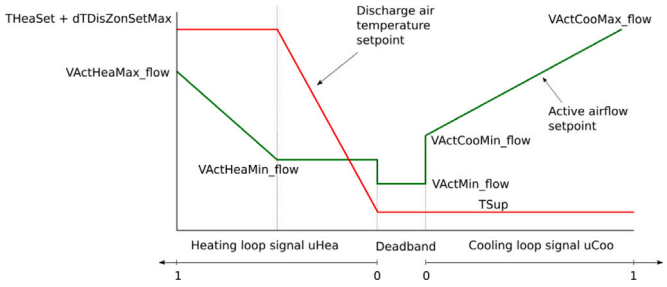


Fig. 6. Damper and valve position for VAV Boxes from [4].

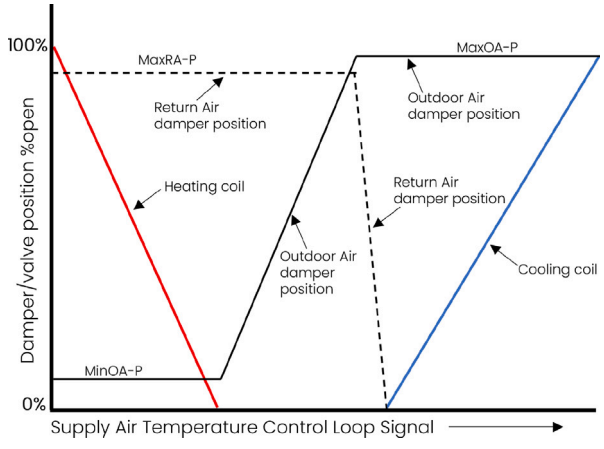


Fig. 7. Damper position in economizer and valve position of heating and cooling coil in G36 control [4].

The valves of the heating and cooling coils operate based on the SAT control loop signal, which is managed by a PI controller that tracks the SAT setpoint. When the fan is off, the control signal is set to 0. For cooling, as the SAT increases from its minimum value, the cooling valve control signal similarly increases linearly from 0 to 1, gradually opening the valve (see Fig. 7).

A weather-compensated control strategy was implemented for regulating the SWT of chillers according to summer conditions. This control method dynamically adjusts the chiller supply water temperature setpoint based on outdoor temperature fluctuations. During periods of high outdoor temperatures, the control system reduces the chilled water temperature setpoint to meet the increased cooling demand. Conversely, during relatively cooler summer periods, slightly higher setpoints are utilized to reduce the operational load on the chiller.

3.1.1. Implementation of DRL control strategy

DRL is a branch of machine learning where an agent learns the optimal control policy for a specific problem through a trail-and-error approach. The learning in the DRL framework is driven by a feedback mechanism formalized in the form of reward or penalty signal. The objective of the agent is to learn a policy that maximizes the cumulative reward over time.

DRL is typically formulated as a Markov Decision Process (MDP) [47], defined by a tuple (S, A, P, R, γ), where:

- S represents the set of states,
- A represents the set of actions,
- P is the state transition probability function,
- R is the reward function, and
- γ is the discount factor, which determines the importance of future rewards.

During the learning process, the agent seeks to identify the optimal mapping between states and actions in order to maximize reward return. This goal is achieved while balancing the exploration of unseen control trajectories and the exploitation of learned knowledge. According to DRL the control policy, i.e. the mapping between state and actions, is formalized through deep neural networks.

In this study, the Soft Actor Critic (SAC) algorithm [13] was implemented. SAC algorithm can handle continuous action spaces and employs an off-policy evaluation mechanism encoded within specific Actor–Critic architecture. Two distinct DNNs are employed: the Actor network, which maps the current state to an estimated optimal action, and the Critic networks, which evaluate the goodness of taking specific actions given a certain state of the environment by estimating the corresponding Q-values. Specifically, SAC uses two critic networks

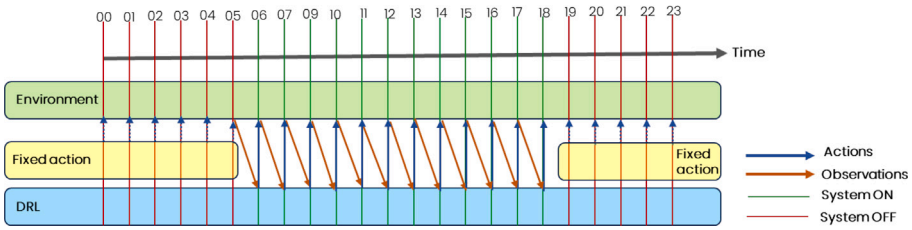


Fig. 8. DRL interaction with simulated environment.

to mitigate overestimation bias and an additional value network for stable training. This dual-network configuration enhances SAC ability to effectively learn and optimize policies in complex and continuous action domains [48–50].

A significant aspect of the SAC algorithm is the incorporation of entropy regularization [48]. This algorithm is based on the maximum entropy reinforcement learning framework, which aims to maximize both the expected reward and entropy. The objective can be expressed as follows:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \left(r_t + \alpha H(\pi_t)\right)\right] \tag{1}$$

The term H represents the Shannon entropy, which quantifies the agent's propensity for taking random actions. The coefficient α serves to balance the relative importance of entropy against the reward. In conventional reinforcement learning algorithms, α is typically set to zero. Maximizing this objective function is inherently linked to the exploration–exploitation trade-off. This ensures that the agent actively explores new policies while avoiding suboptimal behavior traps.

In the present application, the DRL agent based on the SAC framework is operated exclusively during the operational hours of the system. As illustrated in Fig. 8, the SAC agent continuously interacts with the environment during these periods of activity, receiving observations, selecting actions and adjusting the control policy based on the corresponding rewards.

Conversely, during periods when the system is inactive, the SAC agent does not receive new observations from the controlled environment and the implemented actions are defined according an expert-based schedule.

Table 3 lists the specific observations processed by the SAC algorithm with evidence of the variable names and their corresponding descriptions. The DRL is designed to work in conjunction with the A2006 controller as specified in Table 2. The primary role of the DRL is to optimize the supply air temperature (SAT) dynamically adjusting the following variables with a control timestep of 30 min:

- Economizer damper position (**E_damper_position**). The controller operates within a range between a minimum opening position that provides adequate air quality and a maximum opening.
- Supply water temperature setpoint (**SWT**): The SWT setpoint can be dynamically adjusted in the range between 4 °C and 15 °C.
- Chiller valve position (**C_valve_position**): The opening can be adjusted between 0 and 1 with 0.1 increments.

This optimization aims to minimize Zone Air Temperature (ZAT) violations ($ZAT_{violation}$), SAT violations ($SAT_{violation}$) and energy cost ($E_{electricity_consumption}$). Control of the supply fan and VAV boxes follows A2006 control. The reward R for the DRL is reported in Eq. (2) and each term is explained in detail below:

1. **Energy Consumption** ($E_{electricity_consumption}$): This term represents the total electrical energy consumed by the HVAC system. Specifically, it includes the energy used by the supply fan (E_{el_fan}) and the chiller ($E_{el_chiller}$), which are the primary contributors to system energy use.
2. **Zone Air Temperature Violations** ($ZAT_{violation}$): This term quantifies deviations of the zone air temperature ($T_{\text{in,zone},t}$) from the comfort band, defined as ±1 °C around the setpoint temperature (T_{setpoint}).
   - For temperatures below the comfort band ($T_{\text{in,zone},t} - T_{\text{setpoint}} < -1$), the penalty increases linearly with the distance from this bound, with the absolute value ensuring that the $ZAT_{violation}$ term remains positive.
   - For temperatures exceeding the upper comfort limit ($T_{\text{in,zone},t} - T_{\text{setpoint}} > 1$), the penalty increases with the square of the distance from the upper bound of the comfort band. The quadratic function imposes stronger penalties for overheating, reflecting its significant impact on thermal comfort during the cooling season.
3. **Supply Air Temperature Violations** ($SAT_{violation}$): This term penalizes deviations of the supply air temperature (T_{sat}) from the operational limits, which are set between 12 °C and 18 °C. The penalty is calculated as the square of the deviation whenever the supply air temperature falls outside this range.

$$\mathrm{R} = -(E_{electricity_consumption} + ZAT_{violation} + SAT_{violation}) \tag{2}$$

Where:

$$E_{electricity_consumption} = E_{el_fan} + E_{el_chiller}$$

$$ZAT_{\text{violation}} = \begin{cases} \sum_{\text{zone}} \sum_t |T_{\text{in,zone},t} - T_{\text{setpoint}}| & \text{if } T_{\text{in,zone},t} - T_{\text{setpoint}} < -1 \\ \sum_{\text{zone}} \sum_t (T_{\text{in,zone},t} - T_{\text{setpoint}})^2 & \text{if } T_{\text{in,zone},t} - T_{\text{setpoint}} > 1 \\ 0 & \text{otherwise} \end{cases}$$

$$SAT_{\text{violation}} = \begin{cases} \min((T_{\text{sat}} - 12), (18 - T_{\text{sat}}))^2 & \text{if } T_{\text{sat}} < 12\ ^\circ\text{C or} \\ & T_{\text{sat}} > 18\ ^\circ\text{C} \\ 0 & \text{otherwise} \end{cases}$$

The reinforcement learning control agent was developed using the Python Stable Baselines package with the SAC algorithm. The SAC agent used a multi-layer perceptron (MLP) policy with two hidden layers, each consisting of 64 neurons. Notably, learning starts were set to zero because the replay buffer was preloaded with observations, actions, and rewards extracted from the baseline simulation (Imitation learning process). The batch size for training was set to 128, and the learning rate was kept at 1e−4 to facilitate stable and effective learning. Initially, a gradient step of 100 was used for the first time step only to speed up learning, and then it was reduced to 1 to ensure smoother convergence. A total of 20 episodes were simulated during the training process, with the final episode dedicated to deploying the learned policy to ensure robustness and adaptability of the trained agent. In this case study, an episode refers to a sequence of interactions between the agent (the controller) and the co-simulation environment that corresponds to a month of implementation during the cooling season. During each episode, the agent interacts with the environment making decisions, receiving feedback, and adjusting its actions to improve its performance.

Table 3
Observations for DRL Control.

Observation	Description	Unit
SAT	Supply air temperature	[°C]
V_{flow_air}	Supply air volumetric flow rate	[m^3/s]
$V_{flow_outdoor_air}$	Outdoor air volumetric flow rate	[m^3/s]
ZAT_{south}	South zone air temperature	[°C]
ZAT_{east}	East zone air temperature	[°C]
ZAT_{north}	North zone air temperature	[°C]
ZAT_{west}	West zone air temperature	[°C]
ZAT_{core}	Core zone air temperature	[°C]
$T_{outdoor_air}$	Outside dry bulb temperature	[°C]
$VAVBox_{south_damper}$	VAV Box south zone damper position	[–]
$VAVBox_{west_damper}$	VAV Box west zone damper position	[–]
$VAVBox_{north_damper}$	VAV Box north zone damper position	[–]
$VAVBox_{east_damper}$	VAV Box east zone damper position	[–]
$VAVBox_{core_damper}$	VAV Box core zone damper position	[–]
rpm_{signal}	Supply fan control signal	[–]
T_{mix}	Mixed air temperature	[°C]
m_{flow_water}	Cooling water volumetric flow rate	[m^3/s]
SWT	Supply water temperature	[°C]
RWT	Return water temperature	[°C]
hour	Hour of the day	[h]
occupancy	Building occupancy	[–]
T_{out_1h}	Outside temperature 1 h ahead	[°C]
T_{out_2h}	Outside temperature 2 h ahead	[°C]
T_{out_3h}	Outside temperature 3 h ahead	[°C]
T_{out_4h}	Outside temperature 4 h ahead	[°C]

3.1.2. Rule extraction-based control strategy

The policy learned by the DRL controller was used for the extraction of a set of IF-THEN control rules. Those rules were identified by means of regressive DTs which aim to mimic the actions taken by the DRL controller [51]. A regressive decision tree is a type of decision tree algorithm where the goal is to predict a continuous outcome variable (e.g., a continuous control action of the DRL controller). In this algorithm, the data is recursively split into subsets based on input feature values in a manner that reduces the variance within each resulting subset. The tree structure consists of nodes, where each internal node represents a decision or test on a particular feature, and each leaf node represents a predicted value for the target variable. The process begins with the root node, which contains the entire dataset. The algorithm selects a feature and a threshold value that best splits the data into two subsets, aiming to minimize the sum of squared differences between the predicted values and the actual values within each subset [52]. This splitting continues recursively, with the algorithm selecting features and thresholds that further reduce variance, until a stopping criterion is met. This criterion could be a maximum tree depth, a minimum number of samples in a node, or a minimum reduction in variance [53].

In the analyzed case study the observations collected from the deployment simulation of the DRL controller were used to develop the regression trees. Specifically, only the operational hours of the HVAC system were considered, avoiding to include OFF hours in the training set. The number of the developed decision trees is equal to three i.e., one for each action of the DRL controller (SWT setpoint, chiller valve position, economizer damper position). The input variables for these decision trees consisted in a subset of DRL observations reported in Table 3, including supply air conditions from AHU, supply water conditions from the chiller, air temperatures of zones, current and predicted values of outdoor air temperature (a perfect prediction was considered), and the positions of valves and dampers of VAV boxes. Once the decision trees have been developed, the rule extraction was performed. In detail, rule extraction from a decision tree involved following each path from the root node to a leaf node, collecting the conditions encountered along the way, and then combining these conditions into a rule that describes the decision-making process for that particular path. The result were three sets of rules (one for each action) that comprehensively describe how the trees attempted to emulate the DRL controller based on the input features. The deployment of the extracted sets of rules was then performed in the co-simulation environment in order to assess the performance of the RE-based controller.

3.2. Benchmarking analysis

Eventually, a comparison process between the 4 different implemented controllers was performed. This comparison was based on the calculation of the following KPIs considering the reference simulation period of 1 month (1st to 31st of July):

- Energy Consumption (kWh): it measures the total amount of electrical energy consumed by the HVAC system components.
- ZAT Violations (°C): it quantifies the deviations of the zone air temperature values from the acceptable range of 23 °C to 25 °C.

The controller with the lowest values for the defined set of KPIs represents the best-performing one.

4. Results

This section outlines the results of the study. Initially, the performance of the DRL algorithm is presented, focusing on cumulative reward, energy consumption, and indoor air temperature violations. This is followed by an analysis of the RE-based controller performance, which is then directly compared to the DRL controller. Finally, DRL and RE-based controller are compared against the two baseline control policies (A2006 and G36) to quantify the added value provided by proposed approach.

4.1. DRL results

The SAC algorithm was trained by interacting with the co-simulation environment for 20 episodes. After training, its final performance was assessed by deploying the learned control policy in a static manner on a single episode (where the episode is the month of July). The simulations were performed on a workstation featuring an Intel Core i9 processor (3.70 GHz) and 128 GB of RAM. Training the DRL control policy over 20 episodes required approximately 3 h, while the deployment episode took about 10 min. The performance of the DRL controller is evaluated through the cumulative of the three reward terms reported in Eq. (2): (i) the energy consumption term; (ii) the comfort term expressed in terms of ZAT violations; and (iii) the SAT violations. The Fig. 9 reports the trend of the three reward components over all the 20 training episodes and the deployment one. Fig. 9 provides a clear visualization of the relative contributions of different reward components to the overall performance of the controller. By tracking the progression of each component the figure highlights how the system optimization balances these factors to achieve the cumulative reward. Overall, the results demonstrated the DRL controller ability to effectively optimize multiple objectives, with the comfort component being optimized first and most successfully, followed by the energy consumption and the SAT term. The fluctuations in the SAT term highlight the complexity of balancing such objectives, but the stabilization across all components indicates the successful training and deployment of the control policy.

4.2. Rule extraction results

The three decision trees were developed using data collected from the deployment episode of the DRL, as explained in Section 3. The accuracy results in terms of Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) ere reported in Table 4.

For the Supply water Temperature the obtained MAE value indicates that the decision tree predictions deviate by about 0.75 °C on average from the actual values of the DRL controller. This error, lower than 1 °C, can be considered acceptable. For the Economizer Damper



Fig. 9. Cumulative reward for the DRL controller broken down by each reward component. The comfort term is represented by the blue solid line, the energy consumption term is shown in red, and the SAT term is depicted in green.

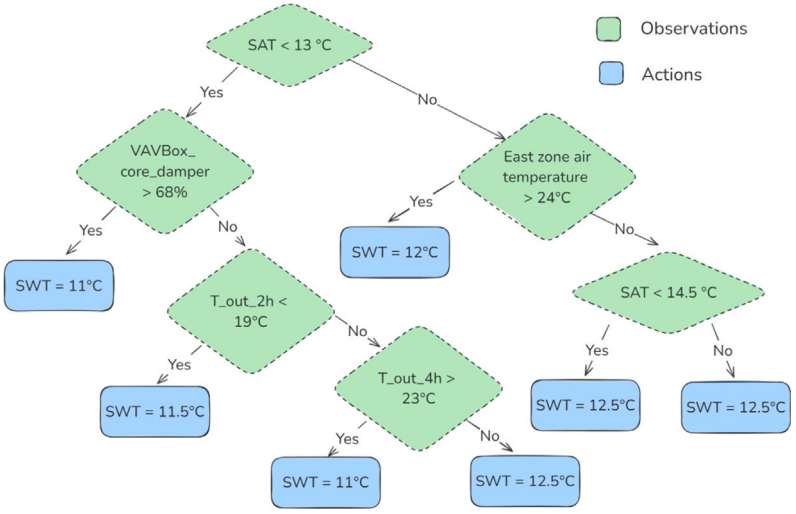


Fig. 10. Decision tree for the estimation of the SWT.

Table 4
Performance metrics evaluated for the developed decision trees (i,e., Supply water temperature, Economizer damper position, and Chiller valve position).

Metrics	Supply water temperature	Economizer damper position	Chiller valve position
MAE	0.748	0.024	0.056
MSE	1.132	0.004	0.022
RMSE	1.064	0.068	0.149

position, which has a range of 0–1, the MAE of 0.0244 suggests that the model predictions are off by about 2.44% of the full range. This relatively small error indicates that the model is quite accurate in predicting the damper position, and this level of precision would likely be acceptable in most HVAC control scenarios. For the Chiller valve position, also with a range of 0–1, the MAE of 0.0557 means that the model predictions are off by about 5.57% of the full range indicating a fairly accurate model. For completeness, in Fig. 10, is reported the decision tree developed for mimic the DRL actions on SWT.

The final model has a depth equal to 4, with 7 leaf nodes representing predicted actions. It means that it can be easily translated in a set of 7 IF-THEN rules offering a clear and interpretable view of the logic used by the deep reinforcement learning controller. By mapping out the decision paths, the DT provides insight into how the controller makes choices in various scenarios into a form that can be easily understood and analyzed. For regression tasks, the prediction at a leaf node corresponds to the mean of the target variable for all training data points within that node. The extracted rules are then in the following reported:

- **Rule 1: IF** $SAT < 13$ °C **AND** $\text{VAVBox}_{\text{core_damper}} > 0.68$ **THEN** Set SWT to 11.0 °C
- **Rule 2: IF** $SAT < 13$ °C **AND** $\text{VAVBox}_{\text{core_damper}} \leq 0.68$ **AND** $T_{\text{out_2h}} < 19.0$ °C **THEN** Set SWT to 11.5 °C
- **Rule 3: IF** $SAT < 13$ °C **AND** $\text{VAVBox}_{\text{core_damper}} \leq 0.68$ **AND** $T_{\text{out_2h}} \geq 19.0$ °C **AND** $T_{\text{out_4h}} > 23.0$ °C **THEN** Set SWT to 12.5 °C
- **Rule 4: IF** $SAT < 13$ °C **AND** $\text{VAVBox}_{\text{core_damper}} \leq 0.68$ **AND** $T_{\text{out_2h}} \geq 19.0$ °C **AND** $T_{\text{out_4h}} \leq 23.0$ °C **THEN** Set SWT to 11.0 °C
- **Rule 5: IF** $SAT \geq 13$ °C **AND** $\text{ZAT}_{\text{east}} > 24.0$ °C **THEN** Set SWT to 12.0 °C
- **Rule 6: IF** $SAT \geq 13$ °C **AND** $\text{ZAT}_{\text{east}} \leq 24.0$ °C **AND** $SAT < 14.5$ °C **THEN** Set SWT to 13.0 °C
- **Rule 7: IF** $SAT \geq 13$ °C **AND** $\text{ZAT}_{\text{east}} \leq 24.0$ °C **AND** $SAT \geq 14.5$ °C **THEN** Set SWT to 14.0 °C

where SAT is the Supply air temperature of the AHU, VAVBox_core_damper is the position of the Damper VAV in the core zone of the building, $T_{\text{out_2h}}$ and $T_{\text{out_4h}}$ are the values of outside air

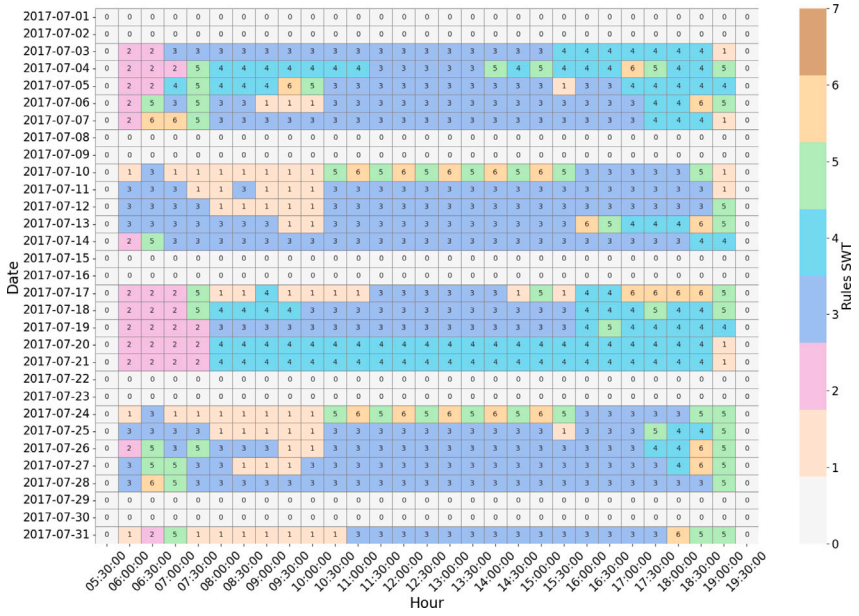


Fig. 11. Rules pertaining to the management of SWT implemented by the RE-based controller during deployment.

temperature 2 and 4 h ahead to the time of decision (considering a perfect prediction), ZAT_east is the current air temperature in the eastern zone of the building. According to the above reported decision rules it is possible to infer some key aspects about what the RE process made it possible to learn from the DRL control policy.

One of the key considerations is the sensitivity of the RE-based controller to external conditions. The rules emphasize the importance of future outside air temperatures, such as the temperatures over the next two and four hours (T_{out_2h} and T_{out_4h}) allowing for exploiting the predictive capabilities of the reference DRL controller. The position of the Damper VAV in the core zone also plays a significant role, especially when the SAT is below 13 °C. This indicates that the controller considers the internal airflow and distribution needs within the building. Zone-specific adjustments are another crucial aspect of the decision rules. The system takes into account the temperature in specific zones, particularly in the east zone of the building (ZAT east). This suggests that the control strategy is able to identify localized temperature variations, which are important for maintaining consistent comfort across different areas of the building. Considering the above listed set of rules, a detailed analysis of rule usage was conducted to determine when specific rules are likely to be implemented by the RE-based controller during its deployment in the co-simulation environment. Fig. 11 illustrates the implementation of the rules over different hours in the simulated month of July. The *y*-axis represents the date, while the *x*-axis shows the time of day, starting from 06:00 and ending at 19:00 with a timestep of 30 min (that is the amount of time between two consecutive actions taken from the controller). Each cell within the grid is color-coded according to a specific rule number, ranging from 0 to 7, as indicated by the color scale on the right side of the figure where Rule 0, in gray, indicates periods when no active control is applied.

As shown in Fig. 11, Rule 3 stands out as the most frequently applied control action, particularly during the mid-morning and early afternoon hours when outdoor temperatures exceed 23 °C. Its consistent application across multiple days highlights its significant role in maintaining thermal comfort during the peak cooling demand of the day. Furthermore, Rules 1 and 2, which are employed earlier in the day, typically between 06:00 and 09:00 a.m., are essential for initially adjusting the building indoor air temperature within the comfort band. These rules, characterized by lower values of SWT setpoint, effectively prepare the system for the higher cooling demand that comes later in the day. The results also reports the targeted use of Rule 5 during afternoons that revealed to be the hottest in the simulated period, ensuring continued comfort under peak temperature conditions, while Rule 6 is applied selectively in transitional periods from mid to late afternoon.

From the analysis of the implemented rules within the co-simulation environment, it emerged that Rule 7, despite being extracted from the deployment episode of the DRL controller, was never actually employed by the RE-based controller. This suggests that, in the action sequence executed by the RE-based controller in the co-simulation environment, the specific conditions required to trigger Rule 7 were never encountered. This aspect underscores a potential limitation in the rule extraction and application process indicating that the RE-based controller is not fully replicating the decision-making pathways of the DRL controller, potentially overlooking certain strategies that could be beneficial under specific conditions. The DTs related to the Economizer damper position and the Chiller valve position are detailed in Appendix. Following the same approach as discussed earlier, IF-THEN rules were also derived from these DTs. Specifically, 6 and 9 rules were extracted from the two trees, respectively. This resulted in three set of decision rules that are applied simultaneously: 7 rules for setting the SWT, 6 rules for determining the Economizer Damper position, and 9 rules for controlling the chiller valve opening.

4.3. Comparison results

This section presents the results of the comparative analysis conducted across all controllers. To achieve this, the KPIs related to energy consumption and thermal comfort are summarized in Table 5.

As shown in Table 5, the DRL-based controller significantly outperforms the other controllers in terms of electrical energy consumption, with a total of 938 kWh, which is about 20% lower than both the A2006 and G36 controllers. The RE-based controller also shows improved performance, leading to a final consumption of energy equal to 967 kWh, which, although higher than that of the DRL-based controller, remains lower than the consumption levels of the A2006 and G36 controllers.

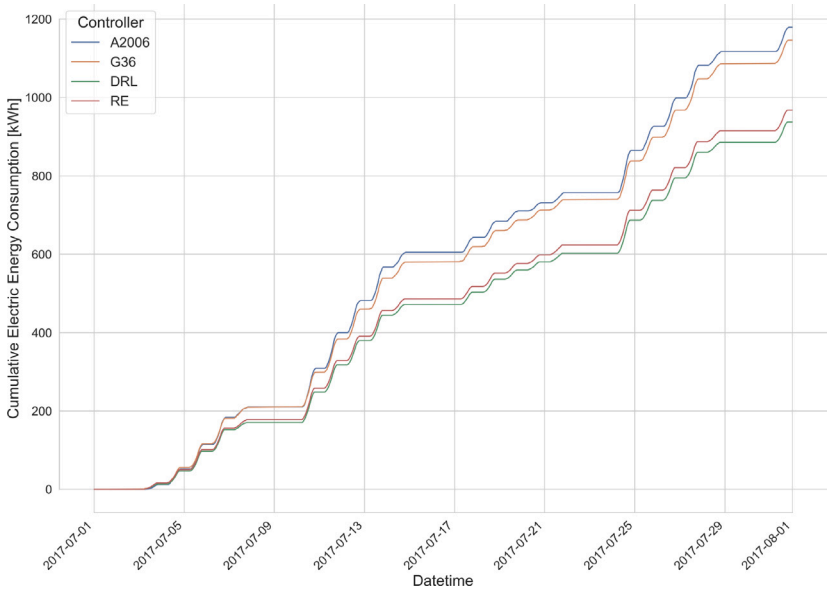


Fig. 12. Cumulative of the total energy consumption achieved by the different controllers.

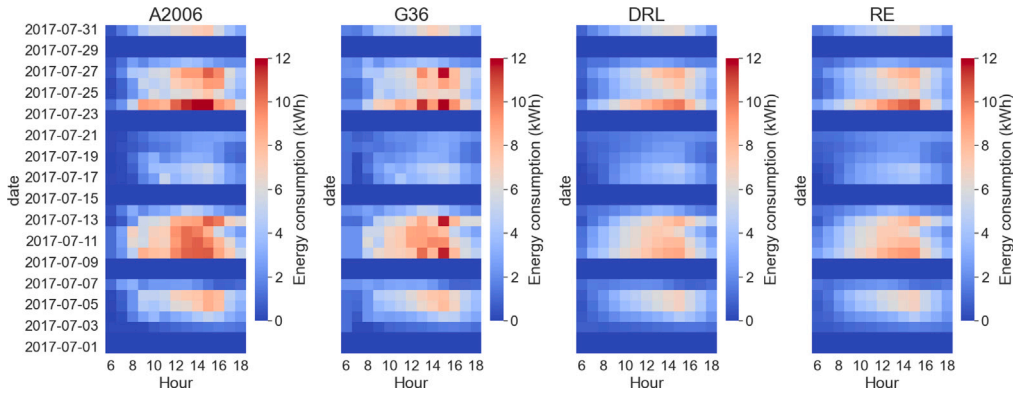


Fig. 13. Energy consumption for the different controllers (A2006, G36, DRL, and RE) throughout the operation hours of the deployment episode. Each subplot depicts the hourly energy consumption for a specific controller, with intensity indicated by color.

Table 5
KPIs for different controllers.

Controller	Electric energy consumption [kWh]	ZAT violations [°C]
A2006	1179	9.9
G36	1146	14.8
DRL	938	0.6
RE	967	4.1

Fig. 12 shows the cumulative of the total energy consumption achieved by the different controllers. It can be observed that the baseline controllers exhibit comparable energy consumption, with the A2006 controller exhibiting a higher consumption than the G36. On the other hand, the RE controller determined a consumption profile that is nearly identical to that of the DRL controller, suggesting that their behavior throughout the period is analogous. This aspect is further validated in Fig. 13 where energy consumption patterns across controllers are reported with hourly detail.

In terms of ZAT violation, which is an indirect measure of thermal comfort, the DRL-based controller again demonstrated superior performance with only 0.6 °C of violation. This indicates the strong capability of the controller in maintaining the desired level of indoor air temperature. The RE-based controller also achieved good performances, with a ZAT violation of 4.1 °C, which is notably lower than both the A2006 and G36 controllers. In this sense, Fig. 14 shows, for a period of one week in July, the indoor air temperature trends for the five thermal zones of the building and the SAT for all considered controllers. The green band in each subplot represents the assumed thermal comfort band, which is the range of indoor temperatures that should be maintained during occupied hours (highlighted by a vertical gray band). The fluctuations observed in ZAT more evident when the G36 was implemented, are mainly due to the trim and respond logic. Similarly, especially during the start-up phase of the HVAC system, also the A2006 determined the occurrence of indoor air temperature violations in the five zones. On the other hand, both the DRL and RE-based controllers effectively managed the control of SAT near the start of the occupied period, thereby reducing the risk of losing control over the indoor air temperature.

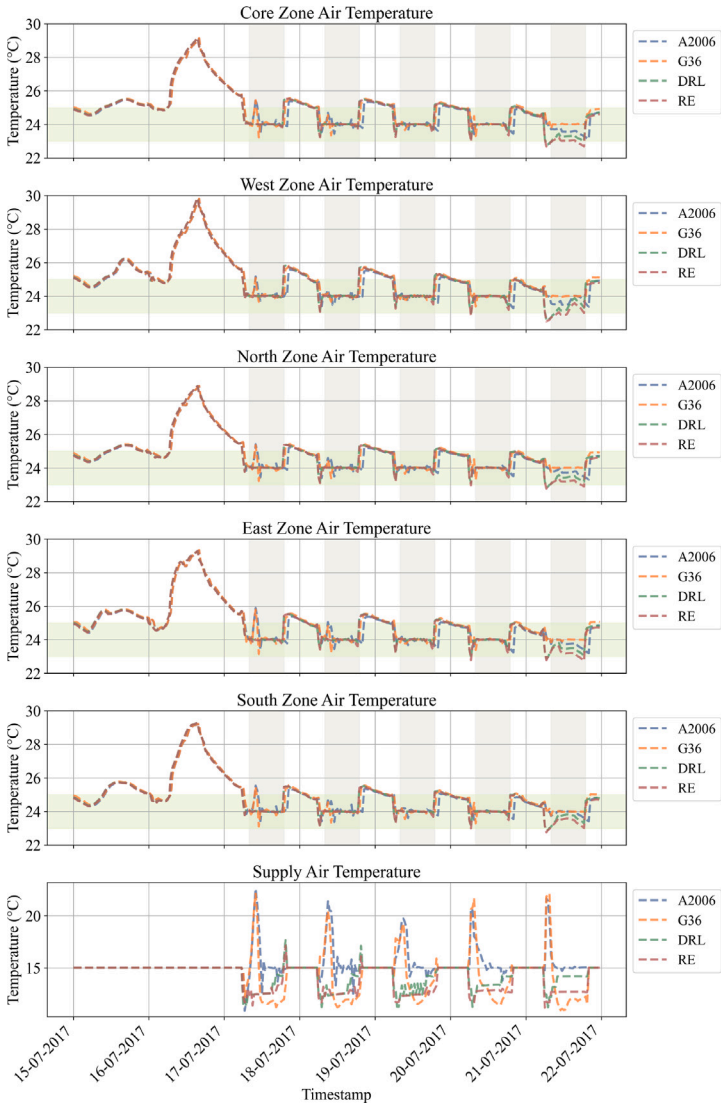


Fig. 14. Indoor temperature trends across five zones and SAT for all controllers.

Table 6
Break down of the electrical energy consumption between fan and chiller across different controllers.

Controller	Energy consumption fan [kWh]	Energy consumption chiller [kWh]
A2006	213	966
G36	179	967
DRL	71	867
RE	73	894

The lower electrical energy consumption observed with the DRL controller and the RE-based controller, compared to the baseline controllers, can be attributed to an improved management of both the AHU fan and chiller as reported in Table 6.

The operation of the DRL-based controller results in the lowest energy consumption for the AHU fan, equal to 71 kWh. This is a significant reduction compared to the A2006 and G36 controllers, which consume 213 kWh and 179 kWh, respectively. At the same time, the chiller electrical energy consumption for the DRL controller is 867 kWh, which is approximately 100 kWh lower than the consumption of the two baseline controllers.

In terms of fan energy consumption, the RE-based controller consumed 73 kWh, which is only 2.8% higher than the DRL controller. This slight increase suggests that the rule extraction process has successfully captured from the DRL the strategy for optimizing the fan usage, maintaining a very close level of efficiency.

Regarding chiller energy consumption, the RE-based controller consumes 894 kWh, which is approximately 3.2% higher than the DRL controller. This modest increase in energy use still reflects a strong ability of the rule extraction process to replicate the DRL controller chiller management strategy.

Fig. 15(a) shows the box plots of the Coefficient Of Performance (COP) of the chiller under the four considered control scenarios, considering as a calculation timestep 30 min. The DRL controller stands out with the highest median COP among the four controllers, indicating superior efficiency in operating the chiller. The relatively narrow interquartile range (IQR) range suggests that the DRL controller consistently maintains this high efficiency across various conditions,

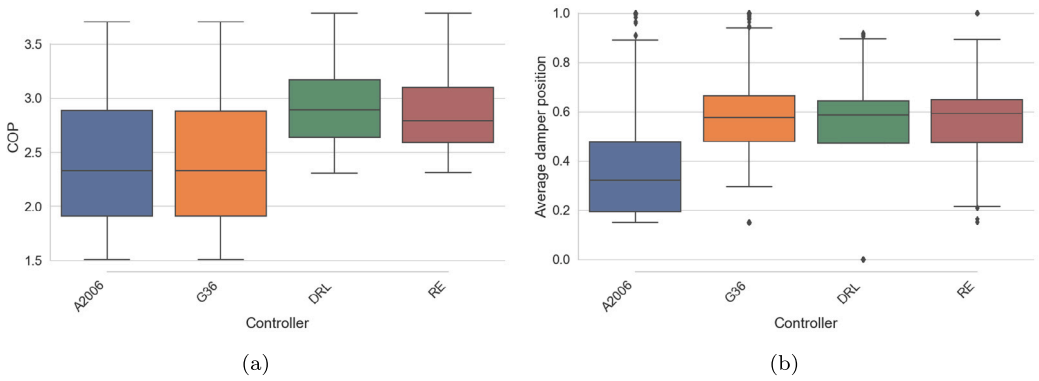


Fig. 15. Box plots of the chiller COP (a) and mean damper position of the VAV boxes in the five thermal zone of the building (b) under the different controllers.

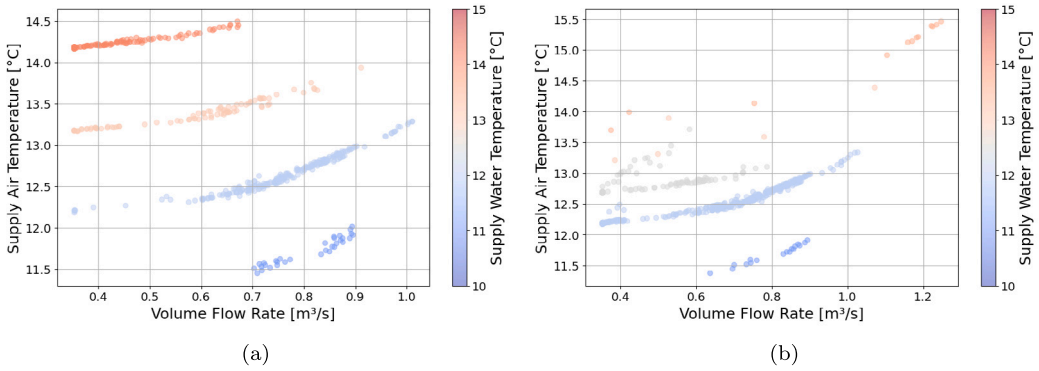


Fig. 16. Scatter plots comparing relationships between SAT, supply air flow rate and SWT in the implementation scenario of DRL controller (a) and RE-based controller (b).

with minimal variability. The RE-based controller also shows strong performance, with a median COP slightly lower than the DRL controller but still significantly higher than the A2006 and G36 controllers.

Similarly, Fig. 15(b) shows the box plots pertaining to the mean damper position of the VAV boxes in the five thermal zone of the building. The data reveals that the DRL and RE-based controller were able to maintain the VAV box dampers on average significantly more opened than the A2006 controller, justifying their lower electrical energy consumption for the AHU fan. In fact, when dampers are more open, the fan operates in a lower resistance regime, meaning less pressure is required to maintain the same volume of air circulation. Consequently, the fan electrical energy consumption decreases significantly due to the reduction in pressure drop across the system. In comparison, the box plot pertaining to the G36 controller also exhibit a median value for damper opening around 55% (close to DRL and RE-based controller) but with a wider range of values below the 1st quartile that are associated to more closed positions of the VAV box dampers.

To highlight the main differences between the DRL controller and the RE-based controller, Fig. 16 presents the relationship between the supply air flow rate and the SAT, averaged over 30-minute intervals for both controllers. The data points are color-coded according to the SWT, with the color gradient indicating variations in SWT, as shown in the color bar on the right side of the figure. Fig. 16(a) illustrates the observations related to the deployment of the DRL controller, while Fig. 16(b) pertains to the RE-based controller. From Fig. 16(a) it can be inferred that DRL controller learned four specific operational patterns that clearly describe the possible relations between SWT, SAT and the air volume flow rate. Regarding the RE-based controller, in Fig. 16(b) it can be observed that for many data points its behavior is largely analogous to the DRL controller, suggesting a comparable response of both controllers under the same boundary conditions. However, for other points, the RE-based controller does not seem to follow the same policy as the DRL controller, especially when the SWT is set to its highest values. This inconsistency was previously discussed and observed in Fig. 11, where it was noted that Rule 7 of the SWT decision tree (which sets the SWT to 14 °C), although derived from the deployment episode of the DRL controller, was never triggered during the deployment of the RE-based controller.

5. Discussion

This study presented a rule-extraction methodology to derive a rule-based controller from a DRL control policy previously trained for an office building in Turin, Italy. The results section examined the framework strengths and limitations, outlying potential directions for future research. The discussion section is then organized into subsections, providing a structured analysis of these findings.

5.1. Optimization strategies for the development of the RE-based controller

The main advantage of the RE-based controller over a sophisticated DRL agent lies in its easier implementation. IF-THEN rules can be easily

integrated within modern BACS architectures, even on edge devices or on-premises applications. However, a significant drawback is that, despite being trained from a DRL control policy, RE-based controller may represent a sub-optimal approximation of the optimal policy. Furthermore, it remains static and tailored to the specific case study, lacking the adaptability of the DRL approach.

To enhance the ability of the proposed RE-based controller to emulate DRL control policy, specific design strategies were employed. The DT models include predictions of outdoor temperature up to four hours ahead, enabling them to effectively anticipate climate variations. Additionally, the maximum depth of the DT models and the maximum number of leaf nodes were carefully optimized. This optimization ensured that the extracted rules were both sufficiently detailed with traits of generalizability, allowing the controller to handle a wide range of operational scenarios. Moreover, the transparent nature of the DT models represents an effective opportunity to simplify the validation of decision rules by HVAC professionals, thereby bridging the gap between advanced control and practical, real-world applications.

5.2. Robust benchmarking for assessing advanced control benefits

To establish a robust benchmark for the proposed method, two baseline strategies based on ASHRAE guidelines (A2006 and G36) were introduced. In terms of energy consumption, from the simulations it was found that the implementation of the G36 led to an energy reduction respect to the A2006 controller. This difference is primarily due to the T&R strategy implemented in the G36. The trained DRL controller achieved a 18% reduction in energy consumption compared to the best-performing baseline controller, G36. This reduction is primarily due to decreases in both fan and chiller energy consumption. Although the DRL controller was not designed to directly modulate fan speed, it effectively optimized SAT to balance the building thermal loads. In addition, by determining a higher VAV damper opening, the controller minimized pressure drops, thereby reducing fan energy consumption. Furthermore, differently from baseline controllers that exploit a weather compensation strategy to set the cooling supply water temperature the DRL controller can determine the optimal values also considering observations pertaining to the actual indoor environmental conditions. This approach resulted in higher values of chiller COP, further contributing to the overall reduction of energy consumption.

All controllers successfully maintained ZAT values within the predefined comfort limits, with only minor temperature violations observed. However, the baseline controllers (A2006 and G36) exhibited slightly higher temperature deviations compared to the DRL controller. The DRL controller demonstrated a superior ability to maintain steady temperatures within the comfort band, making more precise adjustments. Its predictive capabilities also enabled it to anticipate and prevent ZAT violations, particularly those that occurred during the initial hours of operation in the baseline controller implementations.

The developed RE-based controller demonstrated satisfactory performance in terms of both energy consumption and thermal comfort. Specifically, the RE-based controller achieved energy savings comparable to those of the DRL control policy. This suggests that the rule-based approach was able to effectively incorporate the most relevant energy-efficient strategies learned by the DRL controller, such as the optimal management of SAT and damper positions to reduce fan and chiller energy usage.

5.3. Challenges and opportunities for RE -based controllers

Rule extraction from DRL control policy, while potentially simplifying the implementation in real world of an advanced controller, also comes with several drawbacks. One key issue is the loss of precision, as the extracted rules could oversimplify the original DRL agent decisions, leading to suboptimal performance in complex scenarios. On the other hand, in environments with high-dimensional state or action spaces, the complexity of the rules can rapidly increase, making them harder to be interpreted and applied, de facto undermining their usefulness. Temporal dependencies, which are often crucial in DRL controllers, pose another problem. DRL agents rely on the temporal relationships between states and actions to make decisions, but extracting static rules that capture these dependencies could be particularly challenging. This can result in a loss of the dynamic behavior that the original DRL controller exhibits. Additionally, extracted rules are often highly context-dependent, which limits their effectiveness when applied to scenarios different from those learned during training. In this context, techniques such as transfer learning and imitation learning can help mitigate these limitations, but they require significant expertise and still pose challenges in practical implementation.

The authors believe that RE-based controllers offer a valid alternative between traditional and more advanced control systems. In an era where digital twins of buildings and energy systems are essential for developing advanced controllers, such as DRL and MPC, learning an optimal control policy through experimentation and optimization in a risk-free environment is becoming increasingly accessible. However, integrating such advanced control policies into existing BACS can be challenging. Many BACS are not equipped to easily accommodate these sophisticated algorithms, often requiring extensive customization or even hardware upgrades. This integration process typically requires specialized knowledge and expertise, which can pose a significant barrier for many organizations. In contrast, RE-based controllers, which can be translated in a set of IF-THEN rules, can be seamlessly implemented within existing BACS, avoiding the need for the complex architectures required by fully advanced solutions. In essence, RE-based controllers offer a practical solution for leveraging existing BACS infrastructure to exploit advanced, data-driven control strategies without the need for extensive system modifications. While there may be a slight trade-off in performance, the ability to understand, validate, and implement these advanced controls in a more accessible and straightforward manner makes RE-based controllers a promising option worthy of further investigation.

6. Conclusions

In this study, a novel rule-extraction methodology was developed and evaluated to develop a rule-based controller derived from a DRL policy for HVAC system control in an office building. The RE-based controller was compared against traditional baseline controllers, specifically ASHRAE 2006 and ASHRAE Guideline 36 control sequences, demonstrating that the DRL controller outperforms the baselines in both energy efficiency and indoor air temperature violations and the RE-based controller closely approximates the performance of the DRL policy. The RE methodology offers several practical benefits, particularly in making advanced control strategies more accessible and easier to implement within existing BACS. By translating complex DRL policies into interpretable decision tree models, RE-based controller provide a transparent and actionable framework that can be seamlessly implemented into conventional HVAC systems. This approach ensures much of the energy-saving potential and thermal comfort benefits of the DRL controller while simplifying the deployment process.

In addition, the developed co-simulation environment played a crucial role in this research, providing a robust and realistic platform for evaluating and refining the proposed advanced control strategy. By integrating EnergyPlus for building energy modeling with Modelica for detailed HVAC system simulation, the co-simulation framework allowed for an accurate and dynamic representation of the system behavior under the considered control scenarios. In this context, future research will focus on enhancing the generalizability of RE controllers, exploring their application across different building and system types, and further refining the rule extraction process to fully capture the dynamic aspects of DRL policies. This could lead to even more effective and widely applicable solutions for optimizing HVAC operations, contributing to energy savings and improved indoor environmental quality in buildings.

CRediT authorship contribution statement

Giuseppe Razzano: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Silvio Brandi:** Writing – original draft, Visualization, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Marco Savino Piscitelli:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Conceptualization. **Alfonso Capozzoli:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Conceptualization.

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

The work of Giuseppe Razzano and Alfonso Capozzoli was carried out within the project FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (Piano Nazionale di Ripresa E Resilenza (PNRR) – Missione 4 Componente 2, Investimento 1.3 – D.D. 1555 11/10/2022, PE00000013). The work of Silvio Brandi was carried out within the project NODES - Digital and Sustainable North Western Italy and received funding from European Union Next-GenerationEU (Piano Nazionale di Ripresa E Resilenza (PNRR) – Missione 4 Componente 2, Investimento 1.5 – D.D. 1054 23/06/2022, ECS00000036). The work of Marco Savino Piscitelli was carried out within the Ministerial Decree no. 1062/2021 and received funding from the FSE REACT-EU - PON Ricerca e Innovazione 2014–2020. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Appendix. Rule extraction decision trees

This section reports the decision trees developed for extracting control rules from the DRL controller. Fig. A.17 shows the decision tree that estimates the action pertaining to the position of the economizer damper while Fig. A.18 shows the decision tree pertaining to the position of the chiller valve.

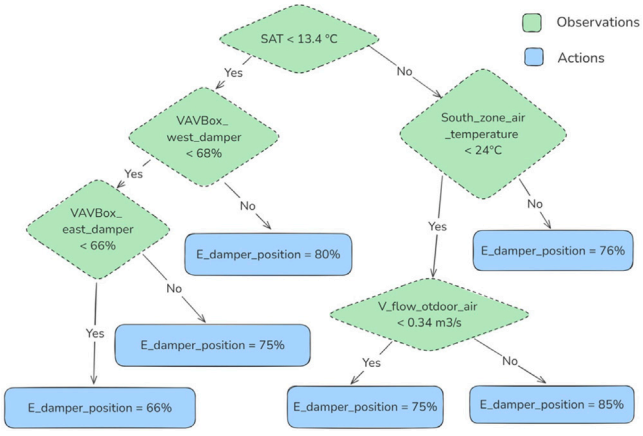


Fig. A.17. Decision tree for the estimation of the economizer damper position.

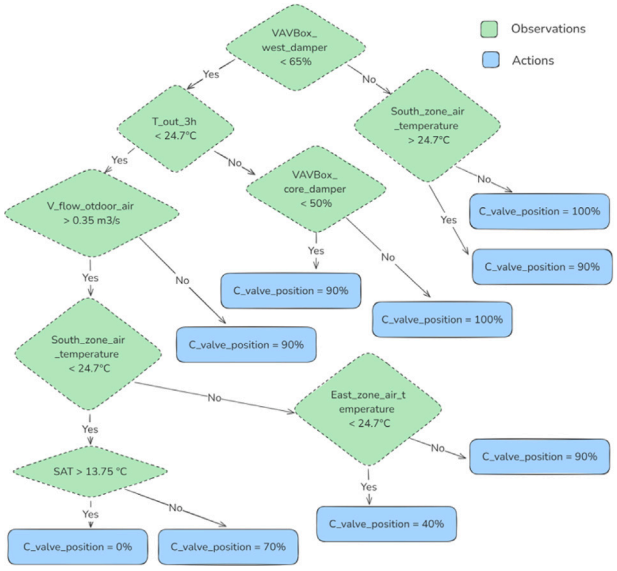


Fig. A.18. Decision tree for the estimation of the chiller valve position.

Data availability

Data will be made available on request.

References

[1] Chen S, Zhang G, Xia X, Chen Y, Setunge S, Shi L. The impacts of occupant behavior on building energy consumption: A review. Sustain Energy Technol Assess 2021;45. http://dx.doi.org/10.1016/j.seta.2021.101212.

[2] Anand P, Sekhar C, Cheong D, Santamouris M, Kondepudi S. Occupancy-based zone-level VAV system control implications on thermal comfort, ventilation, indoor air quality and building energy efficiency. Energy Build 2019;204. http://dx.doi.org/10.1016/j.enbuild.2019.109473.

[3] ASHRAE. Sequences of operation for common HVAC systems. Atlanta, GA: ASHRAE; 2006.

[4] American Society of Heating, Refrigeration and Air-Conditioning Engineers. ASHRAE guideline 36-2021: High-performance sequences of operation for HVAC systems. Atlanta, GA: ASHRAE; 2021.

[5] Zhang K, Blum D, Cheng H, Paliaga G, Wetter M, Granderson J. Estimating ASHRAE Guideline 36 energy savings for multi-zone variable air volume systems using Spawn of EnergyPlus. J Build Perform Simul 2022;15:215–36. http://dx.doi.org/10.1080/19401493.2021.2021286.

[6] Wetter M. Co-simulation of building energy and control systems with the building controls virtual test bed. J Build Perform Simul 2011;4(3):185–203. http://dx.doi.org/10.1080/19401493.2010.518631.

[7] Mu Y, Zhang J, Ma Z, Liu M. A novel air flowrate control method based on terminal damper opening prediction in multi-zone VAV system. Energy 2023;263. http://dx.doi.org/10.1016/j.energy.2022.126031.

[8] Alfalouji Q, Schranz T, Falay B, Wilfling S, Exenberger J, Mattausch T, et al. Co-simulation for buildings and smart energy systems — A taxonomic review. Simul Model Pract Theory 2023;126. http://dx.doi.org/10.1016/j.simpat.2023.102770.

[9] Fritzson P, Pop A, Aronsson P, Lundvall H, Nyström K, Saldamli L, et al. The OpenModelica modeling, simulation, and development environment. 2005, URL: https://www.researchgate.net/publication/252264811.

[10] Blockwitz T, Otter M, Akesson J, Arnold M, Clauss C, Elmqvist H, et al. Functional mockup interface 2.0: The standard for tool independent exchange of simulation models. In: Proceedings of the 9th international MODELICA conference, vol. 76. Linköping University Electronic Press; 2012, p. 173–84. http://dx.doi.org/10.3384/ecp12076173.

[11] Blum D, Jorissen F, Huang S, Chen Y, Arroyo J, Benne K, et al. Prototyping the BOPTEST framework for simulation-based testing of advanced control strategies in buildings. 4, International Building Performance Simulation Association; 2019, p. 2737–44. http://dx.doi.org/10.26868/25222708.2019.211276,

[12] Wetter M, Nouidui TS, Brooks C, Lee EA, Lorenzetti D, Roth A. Prototyping the next generation EnergyPlus simulation engine. In: Proceedings of building simulation 2015: 14th conference of IBPSA. Building simulation, 14, Hyderabad, India: IBPSA; 2015, p. 403–10. http://dx.doi.org/10.26868/25222708.2015.2419.

[13] Haarnoja T, Zhou A, Hartikainen K, Tucker G, Ha S, Tan J, et al. Soft actor-critic algorithms and applications. 2018, URL: http://arxiv.org/abs/1812.05905.
[14] Michailidis P, Michailidis I, Vamvakas D, Kosmatopoulos E. Model-free HVAC control in buildings: A review. Energies 2023;16. http://dx.doi.org/10.3390/en16207124.
[15] Lu X, Fu Y, Xu S, Zhu Q, O'Neill Z. Comparison study of high-performance rule-based HVAC control with deep reinforcement learning-based control in a multi-zone VAV system. 2022, URL: https://docs.lib.purdue.edu/ihpbc/407.
[16] Lu X, Fu Y, O'Neill Z. Benchmarking high performance HVAC rule-based controls with advanced intelligent controllers: A case study in a multi-zone system in modelica. Energy Build 2023;284. http://dx.doi.org/10.1016/j.enbuild.2023.112854.
[17] Fu Y, Xu S, Zhu Q, O'Neill Z, Adetola V. How good are learning-based control v.s. model-based control for load shifting? Investigations on a single zone building energy system. Energy 2023;273. http://dx.doi.org/10.1016/j.energy.2023.127073.
[18] Quang TV, Phuong NL. Using deep learning to optimize HVAC systems in residential buildings. J Green Build 2024;19(1):29–50. http://dx.doi.org/10.3992/jgb.19.1.29.
[19] Silvestri A, Coraci D, Wu D, Borkowski E, Schlueter A. Comparison of two deep reinforcement learning algorithms towards an optimal policy for smart building thermal control. 2600, Institute of Physics; 2023, http://dx.doi.org/10.1088/1742-6596/2600/7/072011,
[20] Du Y, Zandi H, Kotevska O, Kurte K, Munk J, Amasyali K, et al. Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning. Appl Energy 2021;281. http://dx.doi.org/10.1016/j.apenergy.2020.116117.
[21] Zhang Z, Lam KP. Practical implementation and evaluation of deep reinforcement learning control for a radiant heating system. In: BuildSys 2018 - proceedings of the 5th conference on systems for built environments. Association for Computing Machinery, Inc; 2018, p. 148–57. http://dx.doi.org/10.1145/3276774.3276775.
[22] Blad C, Bøgh S, Kallesøe C, Raftery P. A laboratory test of an offline-trained multi-agent reinforcement learning algorithm for heating systems. Appl Energy 2023;337. http://dx.doi.org/10.1016/j.apenergy.2023.120807.
[23] Heidari A, Khovalyg D. DeepValve: Development and experimental testing of a reinforcement learning control framework for occupant-centric heating in offices. Eng Appl Artif Intell 2023;123. http://dx.doi.org/10.1016/j.engappai.2023.106310.
[24] Silvestri A, Coraci D, Brandi S, Capozzoli A, Borkowski E, Köhler J, et al. Real building implementation of a deep reinforcement learning controller to enhance energy efficiency and indoor temperature control. Appl Energy 2024;368:123447. http://dx.doi.org/10.1016/j.apenergy.2024.123447.
[25] Schreiber T, Eschweiler S, Baranski M, Müller D. Application of two promising reinforcement learning algorithms for load shifting in a cooling supply system. Energy Build 2020;229. http://dx.doi.org/10.1016/j.enbuild.2020.110490.
[26] Brandi S, Fiorentini M, Capozzoli A. Comparison of online and offline deep reinforcement learning with model predictive control for thermal energy management. Autom Constr 2022;135. http://dx.doi.org/10.1016/j.autcon.2022.104128.
[27] Ridley M. Explainable artificial intelligence (XAI). Inf Technol Libr 2022;41. http://dx.doi.org/10.6017/ITAL.V41I2.14683.
[28] Jiménez-Raboso J, Manjavacas A, Campoy-Nieves A, Molina-Solana M, Gómez-Romero J. Explaining deep reinforcement learning-based methods for control of building HVAC systems. Commun Comput Inf Sci 2023;1902 CCIS:237–55. http://dx.doi.org/10.1007/978-3-031-44067-0_13.
[29] Zhang K, Zhang J, Xu PD, Gao T, Gao DW. Explainable AI in deep reinforcement learning models for power system emergency control. IEEE Trans Comput Soc Syst 2022;9:419–27. http://dx.doi.org/10.1109/TCSS.2021.3096824.
[30] Barbado A, Corcho Ó, Benjamins R. Rule extraction in unsupervised anomaly detection for model explainability: Application to OneClass SVM. Expert Syst Appl 2022;189:116100. http://dx.doi.org/10.1016/j.eswa.2021.116100.
[31] Hailesilassie T. Rule extraction algorithm for deep neural networks: A review. IJCSIS Int J Comput Sci Inf Secur 2016;14. URL: https://sites.google.com/site/ijcsis/.
[32] Cho S, Park CS. Rule reduction for control of a building cooling system using explainable AI. J Build Perform Simul 2022;15:832–47. http://dx.doi.org/10.1080/19401493.2022.2103586.
[33] Dai Y, Chen Q, Zhang J, Wang X, Chen Y, Gao T, et al. Enhanced oblique decision tree enabled policy extraction for deep reinforcement learning in power system emergency control. Electr Power Syst Res 2022;209. http://dx.doi.org/10.1016/j.epsr.2022.107932.
[34] Choi Y, Lu X, O'Neill Z, Feng F, Yang T. Optimization-informed rule extraction for HVAC system: A case study of dedicated outdoor air system control in a mixed-humid climate zone. Energy Build 2023;295. http://dx.doi.org/10.1016/j.enbuild.2023.113295.
[35] Gunay B, Ouf M, O'Brien W, Newsham G. Building performance optimization for operational rule extraction. 4, International Building Performance Simulation Association; 2019, p. 2819–26. http://dx.doi.org/10.26868/25222708.2019.210271,
[36] Yu MG, Pavlak GS. Extracting interpretable building control rules from multi-objective model predictive control data sets. Energy 2022;240. http://dx.doi.org/10.1016/j.energy.2021.122691.
[37] Piscitelli MS, Brandi S, Gennaro G, Capozzoli A, Favoino F, Serra V. Advanced control strategies for the modulation of solar radiation in buildings: MPC-enhanced rule-based control. 2, International Building Performance Simulation Association; 2019, p. 869–76. http://dx.doi.org/10.26868/25222708.2019.210609,
[38] Bursill MJ, O'Brien L, Beausoleil-Morrison I. Multi-zone field study of rule extraction control to simplify implementation of predictive control to reduce building energy use. Energy Build 2020;222. http://dx.doi.org/10.1016/j.enbuild.2020.110056.
[39] May-Ostendorp PT, Henze GP, Rajagopalan B, Corbin CD. Extraction of supervisory building control rules from model predictive control of windows in a mixed mode building. J Build Perform Simul 2013;6:199–219. http://dx.doi.org/10.1080/19401493.2012.665481.
[40] Deru M, Field K, Studer D, Benne K, Griffith B, Torcellini P, et al. U.S. department of energy commercial reference building models of the national building stock. 2025, URL: http://www.osti.gov/bridge.
[41] Crawley DB, Lawrie LK, Winkelmann FC, Pedersen CO. EnergyPlus: A new-generation building energy simulation program. 2001, URL: https://www.researchgate.net/publication/268390672.
[42] Wetter M, Zuo W, Nouidui TS, Pang X. Modelica Buildings library. J Build Perform Simul 2014;7(4):253–70. http://dx.doi.org/10.1080/19401493.2013.765506.
[43] Andersson C, Åkesson J, Führer C. PyFMI: A Python package for simulation of coupled dynamic models with the functional mock-up interface. 2016, URL: https://api.semanticscholar.org/CorpusID:218002023.
[44] Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, et al. Openai gym. 2016, arXiv preprint arXiv:1606.01540.
[45] Coraci D, Brandi S, Hong T, Capozzoli A. Online transfer learning strategy for enhancing the scalability and deployment of deep reinforcement learning control in smart buildings. Appl Energy 2023;333. http://dx.doi.org/10.1016/j.apenergy.2022.120598.
[46] Liu M, Guo M, Fu Y, O'Neill Z, Gao Y. Expert-guided imitation learning for energy management: Evaluating GAIL's performance in building control applications. Appl Energy 2024;372:123753. http://dx.doi.org/10.1016/j.apenergy.2024.123753.
[47] van Otterlo M, Wiering M. Reinforcement learning and Markov decision processes. In: Wiering M, van Otterlo M, editors. Reinforcement learning: state-of-the-art. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012, p. 3–42. http://dx.doi.org/10.1007/978-3-642-27645-3_1.
[48] Haarnoja T, Zhou A, Hartikainen K, Tucker G, Ha S, Tan J, et al. Soft actor-critic algorithms and applications. 2019, arXiv:1812.05905.
[49] Pinto G, Piscitelli MS, Vázquez-Canteli JR, Nagy Z, Capozzoli A. Coordinated energy management for a cluster of buildings through deep reinforcement learning. Energy 2021;229:120725. http://dx.doi.org/10.1016/j.energy.2021.120725.
[50] Coraci D, Brandi S, Piscitelli MS, Capozzoli A. Online implementation of a soft actor-critic agent to enhance indoor temperature control and energy efficiency in buildings. Energies 2021;14. http://dx.doi.org/10.3390/en14040997.
[51] Song Y, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry 2015;27(2):130–5. http://dx.doi.org/10.11919/j.issn.1002-0829.215044.
[52] Capozzoli A, Piscitelli MS, Brandi S, Grassi D, Chicco G. Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings. Energy 2018;157:336–52. http://dx.doi.org/10.1016/j.energy.2018.05.127.
[53] Gao Y, Miyata S, Akashi Y. How to improve the application potential of deep learning model in HVAC fault diagnosis: Based on pruning and interpretable deep learning method. Appl Energy 2023;348:121591. http://dx.doi.org/10.1016/j.apenergy.2023.121591.