Reinforcement learning for HVAC control in intelligent buildings: A technical and conceptual review

Khalil Al Sayed, Abhinandana Boodi, Roozbeh Sadeghian Broujeny, Karim Beddiar

Journal of Building Engineering 95 (2024) 110085

Contents lists available at ScienceDirect

Journal of Building Engineering

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

Reinforcement learning for HVAC control in intelligent buildings: A technical and conceptual review

Khalil Al Sayed [a,b,*], Abhinandana Boodi [a], Roozbeh Sadeghian Broujeny [c], Karim Beddiar [d]

[a] CESI LINEACT (UR 7527), CESI Brest Campus, Brest, 29200, France
[b] École nationale supérieure d'Arts et Métiers, 151 Bd de l'Hôpital, Paris, 75013, France
[c] CESI LINEACT (UR 7527), CESI Arras Campus, Arras, 62000, France
[d] CESI LINEACT (UR 7527), CESI La Rochelle Campus, La Rochelle, 17140, France

ARTICLE INFO

Keywords:
HVAC systems
Markov decision process
Reinforcement learning
Machine learning
Optimization
Building energy
Meta reinforcement learning

ABSTRACT

Heating, Ventilation and Air Conditioning (HVAC) systems in buildings are a major source of global operational CO_2 emissions, primarily due to their high energy demands. Traditional controllers have shown effectiveness in managing building energy use. However, they either struggle to handle complex environments or cannot incorporate learning from experience into their decision-making processes, leading to increased computational requirements. The potential solution to these drawbacks is reinforcement learning (RL), which can overcome them with its versatile and learning-based characteristics. In this context, this study presents a thorough literature review, focusing on studies published since 2019 that applied RL for HVAC system control. It bridges theoretical concepts and literature findings to identify suitable algorithms for each problem and to find gaps. It was found that RL deployment in real buildings is limited (23% of studies), with common training methods revealing fundamental technical problems that prevent their safe use: lack of diversification in exogenous state components (e.g., occupancy schedule, electricity price, and weather) that the agent receives in each episode during training in a way that reflects the diversity or unexpected change in real life. This necessitates repetitive, extensive retraining before real deployment, which is computationally expensive. Future research should focus on applying RL to real buildings by solving the previous problem. The meta-RL emerges as an up-and-coming solution for the generalization capabilities because it trains an agent on a wide range of tasks, making the agent more adaptive and reducing the computational cost. Further research should explore this direction.

Contents

* Corresponding author at: CESI LINEACT (UR 7527), CESI Brest Campus, Brest, 29200, France.
E-mail addresses: kalsayed@cesi.fr (K. Al Sayed), aboodi@cesi.fr (A. Boodi), rsadeghianbroujeny@cesi.fr (R. Sadeghian Broujeny), kbeddiar@cesi.fr (K. Beddiar).

https://doi.org/10.1016/j.jobe.2024.110085
Received 26 February 2024; Received in revised form 25 June 2024; Accepted 28 June 2024
Available online 3 July 2024

List of Abbreviations

A2C/A3C	Asynchronous advantage actor–critic.
ACER	Actor–critic with experience replay.
AHUs	Air handling units.
BEMS	Building Energy Management Systems.
C51	Categorical 51-Atom.
CEM	Cross-entropy method.
DDPG	Deep deterministic policy gradient.
DQN	Deep Q-networks.
DRL	Deep Reinforcement Learning.
ES	Evolution strategies.
ESC	Exogenous state components.
GAMPS	Gradient-aware model-based policy search.
GLIE MC	GLIE Monte-Carlo Control.
GPI	Generalized policy iteration.
GPS	Guided policy search.
HMB-DDPG	Hybrid-model-based DDPG.
HVAC	Heating, Ventilation, and Air Conditioning.
I2 A	Imagination-augmented agents.
MDP	Markov decision process.
ME-TRPO	Model ensemble trust region policy optimization.
Meta-RL	Meta-reinforcement learning.
MBIE-EB	Model-based interval estimation with exploration bonus.
MBPO	Model-based policy optimization.
MVE	Model-based value expansion.
MPC	Model Predictive Control.
NN	Neural networks.
POMDP	Partially observable Markov decision processes.
PID	Proportional-Integral-Derivative.
PILCO	Probabilistic inference for learning control.
PPO	Proximal policy optimization.

PRISMA	Preferred reporting items for systematic reviews and meta-analyses.
PV	Photovoltaic.
RL	Reinforcement learning.
SGA	Stochastic gradient ascent.
SGD	Stochastic gradient descent.
SAC	Soft actor–critic.
SARSA	State–action–reward–state–action.
SoC	State-of-charge.
SSMDP	Standard stationary Markov decision processes.
SNSMDP	Standard non-stationary Markov decision processes.
TD	Temporal-difference.
TD3	Twin delayed DDPG.
TRPO	Trust region policy optimization.
VAML	Value-aware model learning.
VAV	Variable air volume.

1. Introduction

The Heating, Ventilation, and Air Conditioning system (HVAC) has always been an essential and necessary element in buildings, including residential, commercial, and industrial buildings. Its primary objective is to maintain a comfortable and safe indoor environment for the occupants by controlling temperature, air quality, and domestic hot water.

In 2022, HVAC systems accounted for approximately 16.4% (73.6 EJ) of global final energy consumption, with the majority of this energy derived from the combustion of fossil fuels (coal, oil, and natural gas). As a consequence, these systems are responsible for 14% (5.12 Gigatonnes) of global operational CO_2 emissions, a substantial amount of the greenhouse gas that contributes to global warming [1–3].

This understanding of HVAC systems impact on global carbon emissions has taken on an increased urgency following international commitments to combat climate change. Following the Paris Agreement established during the 2015 united nations climate change conference, which incorporated objectives like carbon emission reduction, there has been a surge of interest among numerous countries to develop their own building energy regulations. In light of these strategic decisions, several countries are adopting strict energy efficiency standards that mandate the use of advanced digital controllers. These advanced systems are designed to optimize the operation of HVAC systems [4], thereby reducing energy consumption while concurrently ensuring the overall comfort of the occupants within buildings.

1.1. Existing HVAC system controllers

In the domain of HVAC systems control methods, various approaches have been employed. Among the most widely recognized traditional methods, as highlighted in [5], are the ON/OFF control and the proportional–integral–derivative (PID) system. The ON/OFF control concept is the fundamental approach to system regulation, operating through straightforward activation or deactivation of the system based on predefined temperature thresholds. In addition, some advanced iterations of ON/OFF controllers have been developed to effectively regulate both energy consumption and occupant comfort [6]. On the other hand, the PID system offers a more advanced mechanism, continually adjusting its outputs to reduce the discrepancy between a desired set-point and the actual process variable [7]. Although these two conventional control systems offer simplicity and cost-effectiveness, they fall behind in terms of forward-looking decision-making capabilities, as well as in managing multiple objectives simultaneously. Addressing this drawback, Model Predictive Control (MPC) has become a well-known way to get around this problem. It has been praised as the best way to balance energy savings with user comfort, making it a leader in the field of occupancy-driven optimal control [8].

In the context of MPC, which involves solving an optimization problem at each time step over an extended time horizon, a model representing the system dynamics is crucial. This model forms the basis of the optimization problem constraints and can adopt either a white-box, grey-box or black-box approach, as discussed in [9]. The white-box models are developed based on fundamental physical laws governing a building thermal dynamics [10], offering a theory-based understanding of the system. In contrast, the black-box approach is typically data-driven, using techniques such as machine learning and deep learning to identify patterns between system inputs and outputs [11,12]. Grey-box models are combination of previous models, integrating a core structure developed from physics-based techniques. The determination of their parameters relies on applying parameter estimation algorithms using the data collected from the system [11,13].

While MPC is identified by its flexibility in responding to dynamic systems, it lacks a mechanism to incorporate learning from previous control actions. This results in a relatively consistent and substantial computational burden over time, as MPC continuously undertakes complex calculations to perform optimization.

1.2. Reinforcement learning controller

The advanced controller, RL, adopts a unique approach compared to other controllers, as it is based on learning through direct interaction with the system environment. This allows RL to easily overcome the challenges faced by previous controllers by implementing effective control strategies based on learning and adapting. Theoretically, RL has the potential to excel beyond the most used HVAC control methods. It possesses capabilities for forward-looking decision-making and efficiently handles multiple objectives simultaneously in complex environments [14]. Furthermore, ability of RL to 'remember' [15] and learn from past experiences and control actions renders it computationally more efficient over the long term compared to MPC, particularly after its initial training phase.

RL has been known for it ability to function and learn in real-time without relying on an explicit predefined system dynamic model, an advantage that has been successful in simulation games like Go [16], Starcraft [17] and shogi [18]. However, this advantage is less evident in the context of HVAC system control. Training RL agents directly within real buildings poses significant challenges, including a long duration required to optimize the controller [19] and the risk of potential damage to the building and discomfort to the occupants due to improper actions taken during training [20]. To overcome these challenges, the development of simulated training environments that accurately mimic the dynamics of a target building has become essential. Such environments facilitate the training of controllers in a risk-free setting before they are deployed in actual buildings.

This shift essentially reduces RL advantage over MPC to its capacity for leveraging past experiences. Yet, this advantage encounters challenges in practical applications, particularly during unexpected events such as sudden weather changes or variations in occupancy patterns [21] that were not encountered during training. These scenarios require the RL controller to undergo fine-tuning, a process that is computationally demanding. In contrast, MPC naturally excels in responding to dynamic environments owing to its built-in capability to sequentially adjust and conform to any encountered disturbances [22].

As a result, it is necessary to devise a technique that enhances the RL controller's ability to adapt more effectively to urgent circumstances [21]. Minimizing the computational demands required for such adaptations could enable RL to outperform MPC in terms of performance and computational cost.

1.3. Related works

Several review papers have addressed RL application on buildings, particularly focusing on their types, applications on cost savings, energy savings, demand-response, load shifting, comfort management, etc. In their critical analysis, Ref. [23] presented the application of RL in demand response mechanisms within smart grids. The study highlights RL flexibility in controlling energy systems, including HVAC systems and smart devices, particularly in contexts that integrate renewable energy and dynamic electricity prices. Additionally, the paper points out existing research deficiencies and proposes potential future research directions, focusing on RL ability to adjust to urban developments.

The review conducted by [19] evaluated how RL agents form actions and their impact on optimization objectives within building contexts. This review classified studies according to their relevance to either single or multiple zones and the nature of actions implemented (binary, discrete, continuous). It focused on the trends and difficulties that have been observed while implementing RL for HVAC systems, highlighting the necessity of continued research in this area.

The study by [24] explored the use of RL in controlling indoor thermal conditions to improve energy efficiency and enhance occupant comfort. The paper provided an overview of current research on allowable temperature variability, occupant thermal comfort, and adjustment. It then focused deeper into RL algorithms used for HVAC control, highlighting their functions in responding within dynamic indoor settings. This study covers a range of RL algorithms, the co-simulation of HVAC systems, and the possibilities for reducing energy consumption. It proposes a framework for RL-based controllers to dynamically regulate indoor temperatures, taking into account environmental factors, algorithmic options, and simulation settings.

A comprehensive review by [25] examines the use of Deep Reinforcement Learning (DRL) in Building Energy Management Systems (BEMS), includes HVAC systems across various building types. The study highlights the growing significance of smart buildings in improving energy efficiency and explores the application of DRL in tackling issues like real-time building modeling, achieving multi-objective optimization, and advancing the generalization of BEMS. It categorizes recent developments in DRL-based BEMS, specifically addressing different types of structures such as residential, office, educational buildings, data centers, and other commercial properties. The paper also identifies distinct challenges and future research trajectories for each building category.

The study by [26] examines the enhancement of energy efficiency in smart buildings through the application of RL. The study looks into the optimization of energy use through various RL algorithms in such environments. This study categorizes and examines multiple RL techniques, outlining their respective strengths, weaknesses, and applicability to diverse control issues in the realm of building energy management. Additionally, it presents case studies to demonstrate the real-world implementation of these algorithms in intelligent buildings. The study concluded by providing a future-oriented viewpoint and discussing potential limitations in the field of reinforcement learning as it relates to improving building energy efficiency.

The research presented by [27] offers a comprehensive review of the application of RL in the context of building control systems. The study investigates how RL algorithms, along with their defined states, actions, rewards, and environments, are implemented and how effective they are in building management. It highlights the current research trends, advancements, and existing gaps, observing that RL application in building controls is predominantly at the research phase with minimal real-world implementations. Significant challenges identified include the lengthy training duration of RL algorithms, the necessity for heightened control security and robustness, and the need to improve the generalization capabilities of RL controllers. The study proposes future research avenues, focusing on the development of RL controllers for practical applications, expediting the training process, fortifying control robustness, and establishing an open-source platform for performance bench-marking.

Table 1
Application of PICOS-logic in RL for HVAC Systems Across Various Building Architectures.

Population (P)	Intervention (I)	Comparison (C)	Outcome (O)	Study design (S)
The group of patients or population relevant to the research question.	The intervention, exposure, or factor being investigated.	The main alternative to compare against the intervention.	The outcomes or results to measure the effectiveness of the intervention.	The type of study or research design suitable for answering the research question.
Different Building Types: – Residential – Office – University Campus/School – Data centers – Other Commercial	Principal RL approaches : – Value-Based – Policy-Based – Actor–Critic	This could be traditional HVAC control systems, other types of AI algorithms, manual control, etc.	Results tie to HVAC systems and occupants : – energy efficiency – cost savings – air quality – CO_2 concentration – user comfort levels	– simulations – real-world implementations

1.4. Aim of this study

Existing review papers typically begin with a literature search of RL for BEMS, categorize the algorithms, conduct statistical surveys, and conclude which algorithm can be used for each new HVAC control problem based on those findings. In contrast, this study takes a different approach. It begins with the detailed theoretical foundations of RL and examines how all the elements of these theoretical foundations are projected by the studies in the literature onto the problem of HVAC control. This method not only provides a comprehensive and detailed view of the current state of research but also helps in proposing more effective future oriented solutions and identifying optimal algorithms that have not been previously applied to buildings. Unlike traditional reviews, which focus solely on algorithms used in the collected papers, this approach allows for a broader and more innovative exploration of potential applications. Additionally, the objective of this study is to build a link between the theoretical concepts and the literature findings in order to validate the practicality and accuracy of using RL algorithms for HVAC system control. As a result of this theoretical-literary link, this study addresses the challenges on the previously mentioned gap 1.2 in RL adaptability to unexpected circumstances, which is unavoidable when RL is practically applied in real buildings, while also presenting some solutions. Briefly, this paper offers a comprehensive overview of the current state of the field and outlines what needs to be addressed in future research to advance RL applications in BEMS.

The outline of this paper is as follows: Section 2 describes review papers selection method, Section 3 introduces the application of RL for HVAC system control, Section 4 discusses the benefits and limitations of employing RL along with future directions, and Section 5 concludes the paper.

2. Review paper selection method

In our literature search, we adhered to the preferred reporting items for systematic reviews and meta-analyses (PRISMA) framework [28], a set of guidelines designed to enhance transparency and completeness in reporting systematic reviews and meta-analyses. This guidelines provides a checklist and a flowchart to document the review process, encompassing the stages of study identification, screening, eligibility, and inclusion. As per [25], PRISMA key principles include: (a) developing a review question; (b) locating applicable studies; (c) selecting and appraising the studies; (d) analyzing the data; (e) presenting the outcomes.

This section examines the first three principles, the remaining parts of the paper focuses on the last two.

Initially, for defining the review question, PICOS-logic framework was used [29], representing Population, Intervention, Comparison, Outcomes, and Study Design. Table 1 outlines the proposition of this review based on the PICOS-logic.

Secondly, the review process included querying the comprehensive Scopus database. The search was structured and keywords were selected as delineated in this equation to guide our inquiry:

query = TITLE-ABS-KEY (“reinforcement learning”)
AND TITLE-ABS-KEY (“building”)
AND TITLE-ABS-KEY (“HVAC systems”)

The literature search was carried out in August 2023, employing keywords such as “reinforcement learning”, “building”, and “HVAC systems”. These terms were chosen for their broad scope to yield the most comprehensive collection of relevant studies. In total, 135 papers were identified for analysis.

Thirdly, we refined our selection of papers by applying a four-step process based on PRISMA methodology as illustrated in Fig. 1. During the identification step, it was observed that the bulk of the papers consisted of scientific journal papers and conference papers. During the screening step, we selectively retained only full scientific reviews and journal papers, excluding conference papers and book chapters. This decision was based on the superior quality and reliability of peer-reviewed literature, which undergoes a stringent evaluation process and provides more elaborate data and analysis. During the eligibility phase, we assessed the suitability of 70 journal papers, having set aside 6 review papers. The assessment was conducted based on the PICOS criteria outlined in Table 1, leading us to select articles whose themes aligned with all components listed in the table. In the final selection, 48 papers were deemed suitable for inclusion in our study Fig. 1. In addition to the selection based on PRISMA criteria, an additional 57 papers were incorporated into the review. These papers were identified and selected by examining the references of the papers already chosen for inclusion.

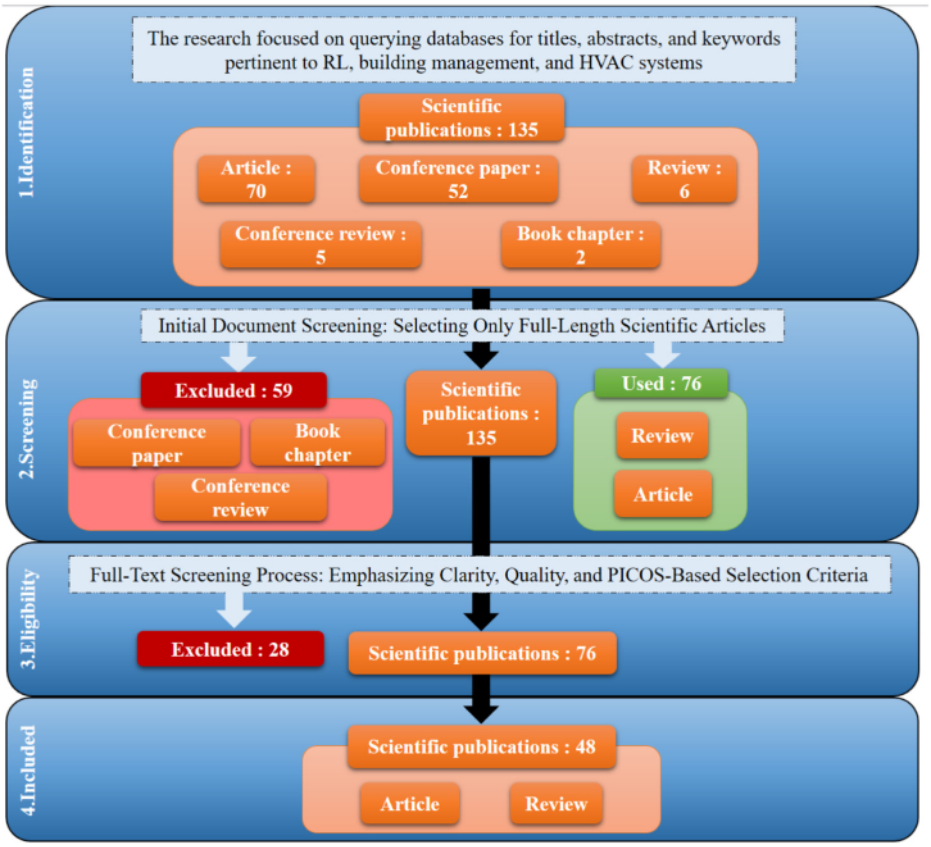

Fig. 1. Structured literature review methodology following the PRISMA principles.

3. Reinforcement learning for HVAC controls

The RL approach is essentially an optimization tool used to identify the optimal decision-making strategy that maximizes the overall accumulated benefit in a decision-making scenario modeled as a Markov decision process (MDP) [30]. If the HVAC systems control problem is not formulated as a MDP, the core principles of RL cannot guarantee the generation of advantageous or optimal results.

3.1. RL modeling environment

MDP is a mathematical framework for modeling decision-making in situations where an agent makes decisions by interacting with an environment [14]. It is defined by a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \mathcal{P}, \Omega, \mathcal{O} \rangle$ where :

- $\mathcal{S}$ is a set of states,
- $\mathcal{A}$ is a set of actions,
- $\mathcal{R}$ is a set of rewards,
- $\mathcal{T}$ is a set of time epochs T [31],
- $\mathcal{P}$ is a probability distribution for going from state s to state s' and getting reward r after taking action a at time epoch e, which has the *Markov property* [14,32]:

$$\begin{aligned} \mathcal{P}(s', r \mid s, a, e) &= P(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a, T = e) \\ &= P(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a, R_t, S_{t-1}, A_{t-1}, \ldots, R_1, S_0, A_0, T = e) \\ &\forall s', r, s, a. \end{aligned} \tag{1}$$

- Ω is a set of observations, where each is a partial description of the state, in a partially observed environment,
- $\mathcal{O}$ is a probability distribution of receiving observation o after taking action a and reaching state s' :

$$\mathcal{O}(o|s', a) = P(O_{t+1} = o|S_{t+1} = s', A_t = a) \tag{2}$$

3.1.1. State $\mathcal{S}$

Signifies the current condition or status of the environment, providing a detailed account of all essential information necessary for decision-making [33]. It encompasses those attributes of the environment that are pertinent to resolving the problem at hand. In HVAC systems control, state representations can vary from simple vectors (e.g., outdoor/indoor temperature, domestic hot water tank temperature, Photovoltaic (PV) production, and hour of the day) as illustrated in [34], to more complex setups (e.g., day of the week, hour of the day, outdoor air temperature, outdoor air relative humidity, wind speed, wind direction, diffuse solar radiation, etc.) as seen in [35].

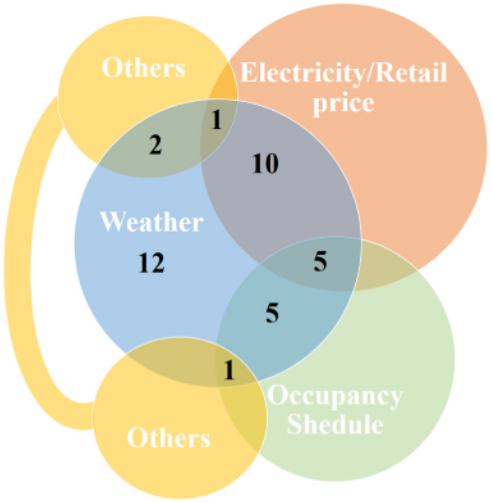

Fig. 2. Exogenous state vector components classification.

In HVAC systems control problems, state vector components fall into two groups: endogenous elements like indoor temperature and power consumption, which are influenced by controller actions, and exogenous factors such as outdoor temperature, electricity prices, and occupancy schedules [36–38], whose transitions are predetermined. Fig. 2 illustrates how ESC are distributed across the various studies identified in the literature search.

3.1.2. Action $\mathcal{A}$

The action space in MDP refers to the set of choices available to the controller to impact the environment. Within the context of HVAC systems control, the complexity arises from the diverse range of its components, each with its unique functions and control mechanisms. As outlined in [27], these components include: (a) The condenser is a crucial component in the refrigeration cycle of chillers and heat pumps. It is responsible for releasing heat during the heating mode and absorbing heat during the cooling mode; (b) The Source component, such as boilers, chillers, and heat pumps, which creates the heating or cooling medium. This process is influenced by factors like water temperature set-points [52], operational modes of chillers or boilers, heat pump configurations [45], and primary pump controls; (c) Air handling units (AHUs), which are crucial for air filtration, temperature regulation, and maintaining air quality within buildings. They are managed through adjustments in fan speed & damper positions [48], and filtration systems; and (d) Terminal components, including variable air volume (VAV) boxes, fan coil units, and radiators, that offer localized temperature control. Their operation is regulated by temperature set-points [36], flow control valves, and customized changes for certain rooms or zones.

3.1.3. Reward $\mathcal{R}$

Within the scope of HVAC systems control, the reward function is crucial for evaluating the controller effectiveness in achieving a balance among several key objectives, such as energy efficiency, maintenance of thermal comfort levels, as well as factors like CO_2 concentration [33,37,68,72] and others. Typically, the reward function is structured as a weighted sum of these diverse elements:

$$R_t = \alpha R_{\text{comfort}_t} + \beta R_{\text{CO}_{2\,t}} + \lambda R_{\text{energy}_t} + \delta R_{\text{cost}_t} + \sigma R_{\text{other}_t}$$

Here, the weights $(\alpha, \beta, \lambda, \delta, \sigma)$ where $\lambda\delta = 0$ signify the relative importance of each objective. The terms R_{energy_t} and R_{cost_t} are formulated as negative rewards that are inversely proportional to the amount and cost of energy consumed, respectively. The product $\lambda\delta$ equals zero, indicating that the signal is derived exclusively from either energy consumption (measured in kWh) or the cost of energy consumption (denominated in euros/dollars), but not both simultaneously.

Conversely, R_{comfort_t} and $R_{\text{CO}_{2_t}}$ are positive rewards awarded for keeping temperature and CO_2 levels within predetermined comfort thresholds, with penalties applied for deviations. Similarly, R_{other_t} in the research represents a positive reward, predominantly observed in demand-response scenarios [61]. R_{other_t} can be for example associated with the indoor relative humidity comfort [70] or the management of a battery state-of-charge (SoC), particularly in a PV system. It encourages charging the battery during off-peak hours when electricity prices are lower, and discharging it during peak periods to lessen grid demand [45]. Furthermore,

Table 2
Overview of MDP components utilized in HVAC control problem.

Ref	Markov decision process												
	Exogenous state components				Action				Reward				
	Weather	Electricity price	Occupancy schedule	Others	Source	Condenser	AHU	Terminal	Energy reduction	Cost minimization	Thermal comfort	Air quality	Others
[39]	✓				✓				✓		✓		
[40]								✓	✓		✓		
[41]			✓						✓		✓		✓
[31]	✓		✓				✓	✓	✓		✓		✓
[34]	✓							✓	✓		✓		
[42]	✓		✓					✓	✓		✓		
[20]		✓					✓	✓	✓	✓	✓		✓
[35]	✓		✓				✓	✓	✓		✓		
[43]	✓				✓				✓		✓		
[44]	✓	✓						✓		✓	✓		
[45]	✓	✓	✓		✓	✓			✓		✓		
[12]	✓						✓		✓		✓		
[46]	✓	✓			✓					✓	✓		
[38]	✓	✓	✓					✓	✓		✓		
[47]								✓	✓		✓		
[48]	✓						✓	✓	✓		✓		
[49]	✓	✓	✓				✓	✓	✓		✓		
[50]	✓				✓	✓	✓		✓		✓		
[51]	✓		✓					✓	✓		✓		
[52]	✓		✓			✓		✓	✓		✓		
[53]						✓			✓		✓		
[54]	✓			✓				✓	✓		✓		
[33]	✓					✓	✓		✓		✓	✓	
[55]	✓						✓	✓	✓		✓		
[56]	✓	✓			✓	✓	✓	✓		✓	✓		
[57]	✓	✓						✓		✓	✓		
[58]								✓	✓		✓		
[59]	✓						✓	✓	✓		✓		
[60]	✓			✓			✓	✓		✓	✓		✓
[61]	✓	✓						✓		✓	✓		
[62]	✓	✓						✓		✓	✓		
[63]	✓	✓					✓			✓	✓		
[64]	✓	✓					✓			✓	✓		
[37]	✓	✓	✓				✓			✓	✓	✓	
[65]	✓		✓	✓				✓	✓		✓		
[66]	✓	✓						✓		✓	✓		
[67]	✓	✓		✓	✓	✓				✓	✓		✓
[68]	✓				✓		✓	✓		✓		✓	
[36]	✓	✓	✓					✓		✓	✓		
[69]	✓		✓		✓	✓			✓		✓		
[70]	✓							✓	✓		✓	✓	
[71]	✓							✓	✓		✓		

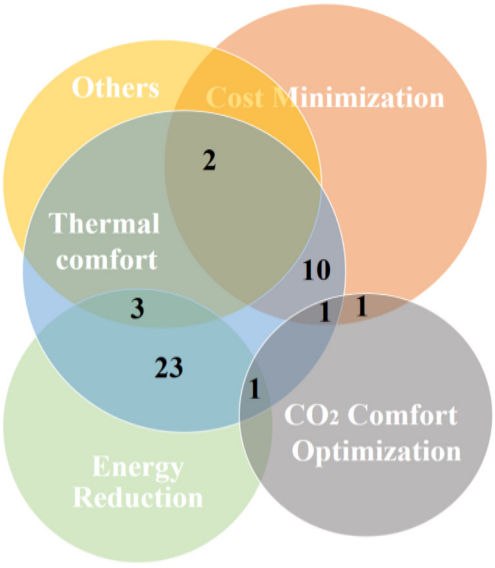

Fig. 3. Classification of Control Objectives by RL.

penalties are imposed for behaviors that could harm the battery's long-term health [20,67]. Fig. 3 illustrates the primary focus objectives according to our literature search. In addition to prioritizing thermal comfort as the main objective, the majority of publications (23 papers) aim to decrease energy consumption, while 11 papers focus on minimizing energy costs. However, certain papers attempt to simultaneously manage three distinct objectives, thereby experimentally validating the RL controller's capacity for use as a multi-objective controller. Table 2 illustrates the specific control objectives for each citation.

3.1.4. Probability $\mathcal{P}, \mathcal{O}$

As stated in (1) and (2), the distributions $\mathcal{P}$ and $\mathcal{O}$ are essential in defining the MDP. These probabilities determine the dynamics of the controllable environment. Therefore, despite the similarities in $\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}$ and Ω across two MDPs, the two probability distributions uniquely express the identity of each MDP. Accordingly, MDP can be classified into three main categories:

1. partially observable Markov decision processes (POMDP),
2. standard stationary Markov decision processes (SSMDP),
3. standard non-stationary Markov decision processes (SNSMDP).

POMDP is a framework in which the agent makes decisions based on observations rather than the true underlying state [73]. This situation arises when sensors are noisy or when some aspects of the environment are hidden from the agent. Such a problem can be modeled through the definition of Ω and $\mathcal{O}$ [61]. In contrast, in the case of full observability, which can be mathematically expressed as ($\Omega = \mathcal{S}$, $\mathcal{O}(s'|s,a) = 1$ if $s' = s$ and $\mathcal{O}(s'|s,a) = 0$ otherwise), standard categories can be defined.

For the two remaining categories: SNSMDP, marked by a time-variant probability distribution $\mathcal{P}$ that varies over time epoch e [32], reflecting the non-stationary nature of real-life room state vectors dynamics. Conversely, the SSMDP is characterized by a static probability distribution $\mathcal{P}$, regardless of the time variable e, representing stationarity and aligning well with the dynamics of room state vectors in simulated environments.

As most studies in the field train and implement their agents within a simulated environment, the SSMDP framework is predominantly adopted in the majority of research papers to model the HVAC systems control problem. In this context, SSMDP can be succinctly represented by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P} \rangle$ where [14]:

$$\begin{aligned} \mathcal{P} : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} &\longrightarrow [0,1] \\ (s', r, s, a) &\longmapsto \mathcal{P}(s', r|s, a) \end{aligned} \tag{3}$$

In the context of HVAC systems control challenges, the uncertainty in state transitions and reward signals following an action, summarized by (3), primarily arises from exogenous components of the state vector. These components, including variables like weather [59], occupancy patterns, and electricity prices, inherently define the probability distribution $\mathcal{P}$, shaping the dynamics of the environment. Furthermore, incorporating real-world data samples for these variables is essential to enhance the realism and applicability of simulation environments, ensuring they closely mimic actual conditions in buildings.

3.1.5. Model environment

As highlighted in the introduction, controllers necessitate training in a virtual simulation environment before deployment in real-world settings Fig. 4. These simulators, which could be based on white-box, grey-box, or black-box models, include tools like EnergyPlus [74], TRNSYS [75], Dymola [76], BOPTEST [77], IESVE [78], Citylearn [79], MATLAB, TESP [80], etc., all details in Table 4. These simulation environments effectively materialize the underlying SSMDP.

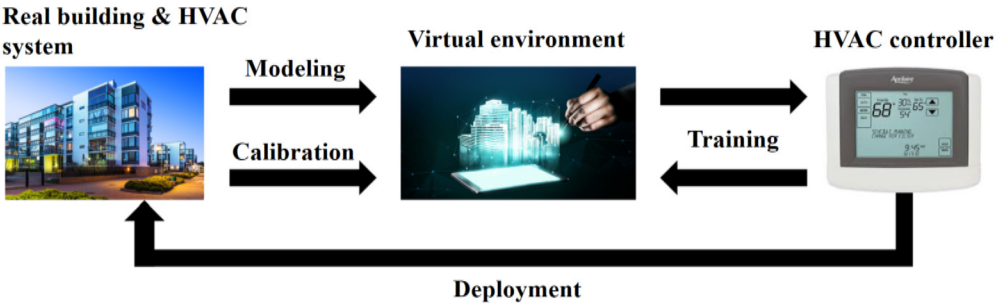

Fig. 4. Stages of RL Controller development: from training to real-world deployment.

Fig. 5 shows the superiority of white/grey-box simulation methods over black-box . This is due to the fact that black-box methods often rely on data-driven approximators, such as neural networks, to simulate the behavior of the building using historical data. However, these methods are susceptible to a problem known as distributional shift.

This problem arises when the internal model of the environment (e.g., a neural-network approximator in the case of a black-box) is used to forecast future states based on the current states and actions of the RL agent. Insufficient training data for specific actions or state–action pairs will result in incorrect predictions about the outcomes of those actions. This can significantly impact the learning process of the RL agent, as it relies on these predictions to make decisions. This clarifies the utilization of the white/grey-box simulation platform, particularly EnergyPlus, which is not solely reliant on historical data but instead on the principles of building physics.

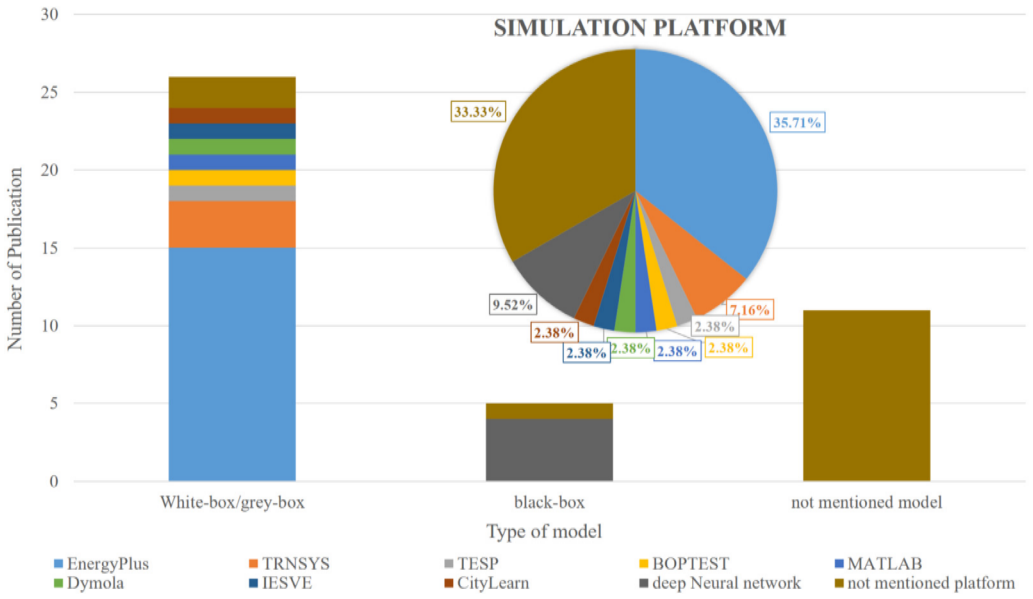

Fig. 5. Setup of the simulation platform utilized for training the RL controller.

3.2. Classification of reinforcement learning algorithms

As indicated before, RL is a mathematical framework with the objective of creating the optimal decision-making policy within the context of a decision-making situation modeled as an MDP. As SSMDP is the predominant category used in literature to model the HVAC systems control problem, this section focuses on the theoretical concepts and its application of RL assigned exclusively to this category of MDP.

In each time steps $t \in \mathbb{N}$ (e.g., 15 min), the agent interacts with the environment (building) receiving a state $S_t \in S$ consisting of endogenous and exogenous HVAC factors 3.1.1, and selects an action $A_t \in \mathcal{A}$ 3.1.2 accordingly. In the next time step $t+1$, and as a result of its action A_t, a scalar reward $R_{t+1} \in \mathcal{R}$ 3.1.3 is given to the agent that has entered a new state S_{t+1} until the agent reaches a terminal state (in the case of an episodic task) that only transitions to itself and generates zero rewards (Fig. 6).

The terminal state is defined as any state vector that outlines the final duration of the ongoing simulation (e.g., 24:00:00 31/12/XXXX) in scenarios of annual episodes commencing in January, or it can be the preceding state plus any state where the indoor temperature component violates comfort thresholds. In the latter scenario, the HVAC control problem is approached similarly to how games like cartpole or Flappy Bird are handled, which accelerates the training at the beginning.

At every time step, the agent interacts with the environment according to a function-based on policy π:

$$\pi : S \times \mathcal{A} \longrightarrow [0,1]$$
$$(s,a) \longmapsto \pi(s,a) = P(A_t = a \mid S_t = s) \tag{4}$$

With the goal of maximizing the total quantity of long-term reward received:

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \tag{5}$$

where R_t is a random variable with possible values $r \in \mathcal{R}$ and corresponding probabilities $(\sum_{s' \in S} \mathcal{P}(s', r|s,a))_{r \in \mathcal{R}}$, and $0 \leq \gamma \leq 1$ is the *discount rate*, which assists in changing the agent behavior to accomplish short-term (γ close to 0) or long-term (γ close to 1) goals [14]. In the context of HVAC control problem, a γ close to 0 is typically used when the goal is to achieve optimal occupant comfort with minimal energy consumption at each timestep. In contrast, a γ close to 1 allows for occasional comfort violations or even excessive energy consumption during some timesteps, in order to achieve long-term energy savings as a final result at the end of the day, month, or year (depending on the episode length).

RL algorithms can be broadly categorized into two main types: model-free and model-based [25]. Model-based algorithms leverage the transition probability distribution $\mathcal{P}(s', r|s,a)$ when deducing the optimal policy, which might either be provided a priori or implied from interactions with the environment. On the other hand, model-free algorithms avoid the need for explicit knowledge of the transition dynamics $\mathcal{P}(s', r|s,a)$ and instead approximate the optimal policy through trial-and-error experiences gathered from direct engagements with the environment. In the context of the HVAC control problem, model-free algorithms are impractical to apply without creating a simulator of the target building, for the reasons outlined in 1.2. This contrasts with model-based, which do not require extensive interaction with a simulator. Model-based algorithms use a predictive model that estimates how actions

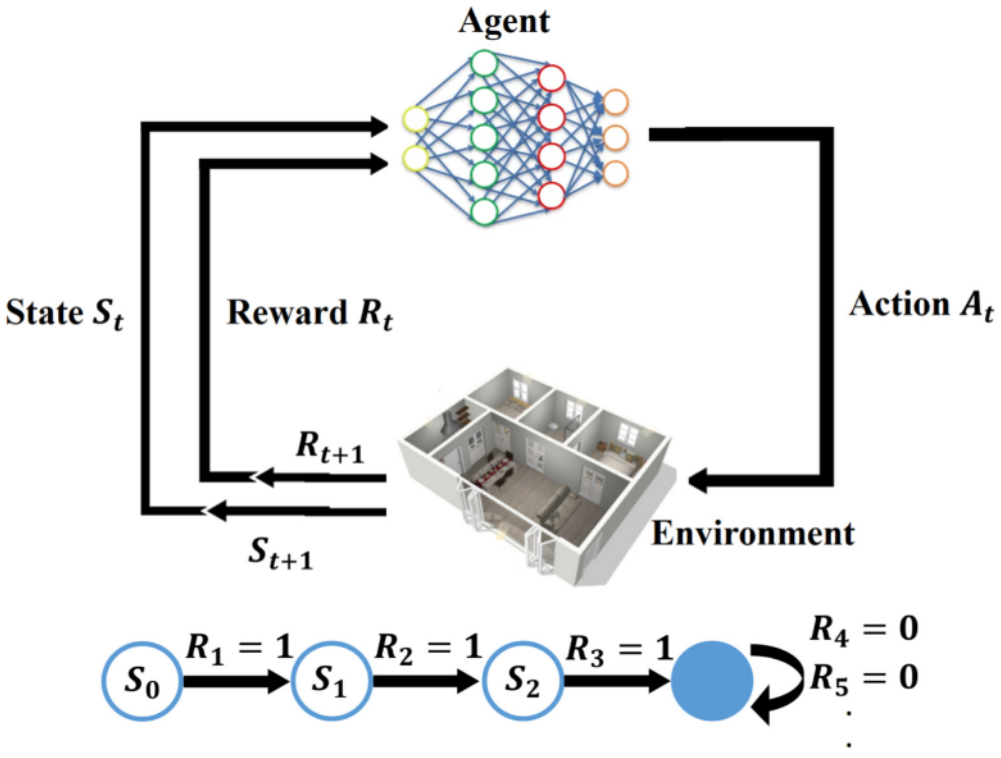

Fig. 6. Process of interaction between agent and environment [38,81].

will influence indoor conditions of the building, allowing for the anticipation of future states and rewards prior to implementing actual measures within the building [40,41]. However, constructing such predictive models is particularly challenging for complex control tasks.

For SSMDP, and Regardless of whether a RL algorithm falls under model-free or model-based categorization, the algorithm's strategies for formulating the optimal decision-making policy can be segmented into three principal approaches: value-based, policy-based, and actor–critic Fig. 7.

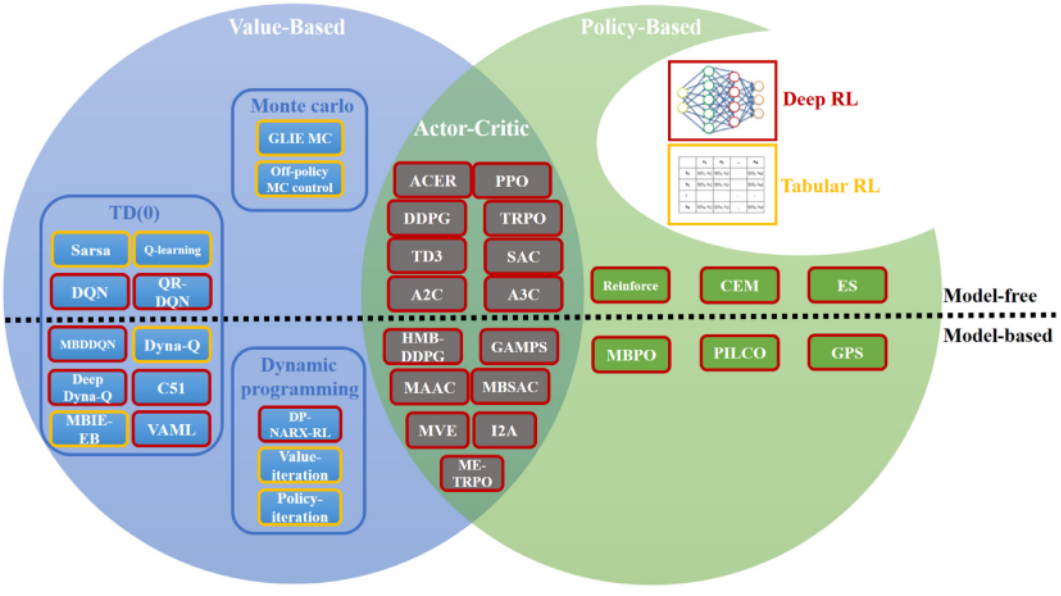

Fig. 7. Categorization of Reinforcement Learning Algorithms.

3.2.1. Value-based

This methodology is pertinent to scenarios with a finite action space, e.g., in [42], the agent was limited to selecting actions from five possible choices [20, 40, 50, 60, 70], which represent the setpoints for the supply water temperature. The key concept in this approach is the estimation of the state/action-value function according to a specific policy. These functions predict the future rewards that can be expected for the agent to be in a given state or for doing a given action in a given state based on the actions that the policy will take during the following time steps. The state and action-value functions can be formally defined for the policy π as [14]:

$$V_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}|S_t = s\right] \tag{6}$$

$$Q_\pi(s,a) = \mathbb{E}_\pi[G_t|S_t = s, A_t = a]. \tag{7}$$

The Bellman equation for V_π and Q_π, states that each value has its own fundamental recursive relationships [82]:

$$V_\pi(s) = \sum_a \pi(s,a)\sum_{s',r}\mathcal{P}(s',r|s,a)\left[r + \gamma V_\pi(s')\right] \tag{8}$$

$$Q_\pi(s,a) = \sum_{s',r}\mathcal{P}(s',r|s,a)\left[r + \gamma V_\pi(s')\right] \tag{9}$$

Given that the primary objective of RL is to identify the optimal policy yielding the highest expected return over an extended period, the dominance of one policy over another can be expressed as:

$$\pi \geq \pi' \iff V_\pi(s) \geq V_{\pi'}(s) \quad \forall s, \tag{10}$$

According to the definition, the optimal policy π^* is one that has the following value function called *optimal state-value function*:

$$V_*(s) = \max_\pi V_\pi(s) \quad \forall s, \tag{11}$$

Thus, the *optimal action-value function* can be defined as:

$$Q_*(s,a) = \max_\pi Q_\pi(s,a) \quad \forall s, a, \tag{12}$$

which has a crucial relationship with V_* follows:

$$V_*(s) = \max_{a\in\mathcal{A}} Q_*(s,a) \quad \forall s, \tag{13}$$

The Bellman optimality equation for V_* and Q_* can be defined by applying Bellman Eq. (9) for (12), and Eq. (13):

$$V_*(s) = \max_{a\in\mathcal{A}}\sum_{s',r}\mathcal{P}(s',r|s,a)\left[r + \gamma V_*(s')\right] \tag{14}$$

$$Q_*(s,a) = \sum_{s',r}\mathcal{P}(s',r|s,a)\left[r + \gamma \max_{a'\in\mathcal{A}} Q_*(s',a')\right] \tag{15}$$

In accordance with (13), a policy is deemed optimal if it assigns non-zero probabilities solely to actions that maximize the optimal action-value function Q_* for each state s. According to this definition, it is not requisite for an optimal policy to be singular. Rather, a general formulation for any optimal policy, denoted as π^*, can be stated as follows:

$$\begin{aligned}&\pi^* : \mathcal{S}\times\mathcal{A} \longrightarrow [0,1]\\ &\qquad (s,a) \longmapsto \pi^*(s,a) = 0 \quad \text{if} \quad a \notin \Sigma(s)\end{aligned} \tag{16}$$

where $\Sigma(s) = \underset{a\in\mathcal{A}}{argmax}\sum_{s',r}\mathcal{P}(s',r|s,a)\left[r + \gamma V_*(s')\right] = \underset{a\in\mathcal{A}}{argmax}\, Q_*(s,a)$ [83].

Based on (16), the key elements for devising an optimal policy are V_* and Q_*. Despite the fact that in ideal situations — where state and action sets, $\mathcal{S}, \mathcal{A}$, are finite and the environmental dynamics, $\mathcal{P}(s',r|s,a)$, are well-defined — it is possible in theory to determine V_* and Q_* by solving (14) and (15) through methods suited for non-linear equation systems [14]. However, the real-world application of these methods is computationally intensive and often impractical. To address this, various approximation methods come into play, highlighting the significance of the value-based RL approach.

All value-based RL algorithms employs the framework of generalized policy iteration (GPI) to ascertain V_* (or Q_*) and π_*. This technique intertwines two key processes, (1) the policy improvement, where the policy is iteratively refined based on the current value/action-value function:

$$\begin{aligned}\pi'(s) = \underset{a}{argmax} Q_\pi(s,a)\ \forall s &\Longrightarrow V_\pi(s) \leq Q_\pi(s,\pi'(s))\ \forall s\\ &\Longrightarrow V_\pi(s) \leq V_{\pi'}(s)\ \forall s \Longrightarrow \pi \leq \pi'.\end{aligned}$$

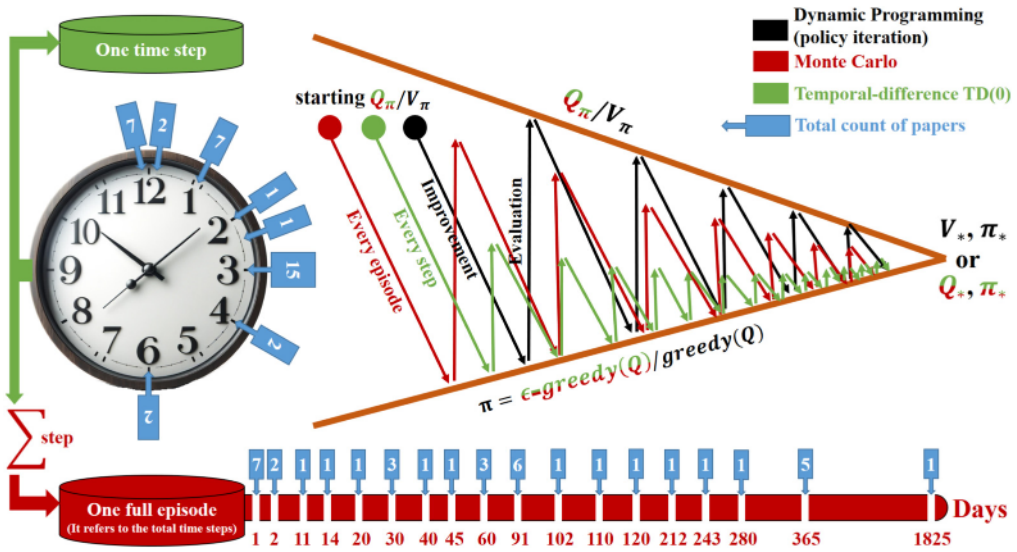

Fig. 8. The concept of generalized policy iteration and its relationship with real-world applications and outcomes. The combination of green and red colors indicates that the symbol is effective for both TD(0) and Monte Carlo scenarios.

Subsequently, (2) the policy evaluation, where the approximated value/action-value function incrementally aligns with its true value, whether wholly (in dynamic programming case) or in part (Monte-Carlo and TD(0)), for the designated policy. Upon convergence of these evaluation and improvement phases, both the value/action-value function and the policy are deemed optimal Fig. 8 .

Furthermore, an algorithm is designated as on-policy when it involves the evaluation and improvement of the policy currently in use for decision-making. In contrast, it is deemed off-policy when the evaluation and improvement pertain to a policy that is not necessarily the one under which behavior is generated, thus facilitating the integration of data from external sources or alternate strategies into the decision-making framework.

In a value-based RL framework, the approximation methods for V_* (or Q_*) can be classified into three primary categories, each with its unique frequency of updates in GPI. These three categories are:

1. Dynamic programming,
2. Monte Carlo methods,
3. Temporal-difference (TD) learning.

Dynamic Programming (RL context): Based on the author's understanding, dynamic programming stands as the sole approach within the category of value-based model-based algorithms that is employed for the control of HVAC systems. Generally, the application of this approach in research papers is limited, stemming from the inherent characteristics of the algorithm that necessitates having predefined values for the environmental dynamics $\mathcal{P}(s', r|s, a)$. Attaining these specific values in the context of HVAC control problems is particularly challenging, often necessitating reliance on approximations derived from interactions with the environment (learn the model).

Algorithm 1 Value Iteration

Parameter: a small threshold $\theta > 0$
Initialize: $V(s)$ for all $s \in S^+$ arbitrarily except $V(\text{terminal}) = 0$
repeat
  $\Delta \leftarrow 0$
  for each $s \in S$ **do**
    $v \leftarrow V(s)$
    $V(s) \leftarrow \max_a \sum_{s',r} p(s', r|s, a)[r + \gamma V(s')]$
    $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
  end for
until $\Delta < \theta$
Output: deterministic policy π such that
$\pi(s) = \arg\max_a \sum_{s',r} p(s', r|s, a)[r + \gamma V(s')]$

Within the dynamic programming domain for BEMS, two key algorithms are significant:

1. value iteration,
2. policy iteration.

The value iteration algorithm draws directly from the Bellman optimality Eq. (14). By adapting this equation into an iterative update rule, V_* is straightforwardly derived as seen from lines 3 to 10 in Algorithm 1 . Consequently, the optimal policy emerges through the application of (16). This algorithm and its derivatives serve as the core control mechanism for BEMS challenges in [40,41,84]. Ref. [33] utilizes a modified version of the value iteration algorithm to manage variables such as chilled water flow rate, fresh air damper ratio, on/off lighting, and open/close windows. This is done with the goal of reducing energy usage while maintaining optimal thermal and CO_2 concentration comfort within a single zone. The results of this modified version showed a marked improvement in energy efficiency and control effectiveness, outperforming both PID control and conventional RL and value iteration techniques.

The policy iteration algorithm, which to the best of the author's knowledge has not yet been applied in BEMS, implements a variant of GPI. This is characterized by a systematic process that alternates between specific policy evaluation and subsequent policy improvement. Their specific policy evaluation is derived explicitly from the Bellman Eq. (8), which is integrated into an iterative update mechanism akin to that outlined in Algorithm 1, yielding the entire value function corresponding to the current policy.

Monte Carlo (RL context): Originating from (7) and fall under the umbrella of model-free algorithms, this approach prioritizes the approximation of the action-value function Q over the state-value function V during policy evaluation. Unlike dynamic programming for V, the policy evaluation procedure does not entail a thorough approximation of Q (i.e., the red arrow does not reach the objective fully during evaluation Fig. 8). For policy improvement, an ϵ-greedy strategy is employed to ensure comprehensive exploration of states throughout environmental interaction. One of the important algorithms, GLIE Monte-Carlo Control (GLIE MC) (Algorithm 2) [14,85] samples episodes and updates the visitation count $N(s_t, a_t)$ for each state–action pair and adjusts the action-value function Q based on the incremental returns G_t; it then progressively refines the policy using an ϵ-greedy approach that becomes less exploratory as the number of encounters increases. As more episodes are processed, the estimated Q converges to the optimal action-value function Q_* under the optimal policy π_* Fig. 8.

As indicated before, Within the context of HVAC systems in an MDP framework, the state vector is typically composed of various parameters that reflect the immediate conditions and operational status of the building energy systems. This encompasses indoor environmental metrics like temperature, humidity, and CO_2 levels, alongside occupancy counts, external weather conditions, and real-time energy usage data. The extensive nature of these combined parameters results in a vast state space, which presents challenges for storing the action-value for every state in a conventional lookup table. Additionally, ensuring that an ϵ-greedy policy explores each state sufficiently can require a prohibitively large number of episodes. To address these issues, function approximators can be employed, specifically neural networks (NN), to estimate the action-value function for each state–action pair, denoted as $\hat{Q}(s, a, w)$, where the NN is parameterized by the weights w.

In the adaptation of Monte Carlo methods to approximate solutions, significant changes to the foundational approach are minimal. For instance, in the GLIE MC algorithm, the standard procedure involving steps 2–5 is substituted by implementing Eq. (18). This involves training the neural network $\hat{Q}$, where the weights w are adjusted via stochastic gradient descent (SGD) to minimize a designated loss function L[85]:

$$L(w) = \mathbb{E}_{\pi}\left[(G_t - \hat{Q}(s, a, w))^2\right] \tag{17}$$

$$\Delta w = -\frac{1}{2}\alpha \nabla_w L(w) = \alpha(G_t - \hat{Q}(s, a, w))\nabla_w \hat{Q}(s, a, w) \tag{18}$$

Algorithm 2 GLIE MC

1: Sample the k^{th} episode using $\pi : \{s_1, a_1, r_2, \ldots, s_T\} \sim \pi$
2: **for** each state s_t and action a_t in the episode **do**
3: $N(s_t, a_t) \leftarrow N(s_t, a_t) + 1$
4: $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \frac{1}{N(s_t, a_t)}(G_t - Q(s_t, a_t))$
5: **end for**
6: Improve policy based on new action-value function:
$\epsilon \leftarrow \frac{1}{k}$
$\pi \leftarrow \epsilon$-greedy(Q)

To the author's understanding, there has been no application of the (RL) Monte Carlo method within the realm of BEMS. This method offers an advantage especially in situations where delayed rewards are involved, a common occurrence in BEMS in case of real-world deployment. Here, the outcomes of certain actions, such as thermostat adjustments or the implementation of energy conservation strategies, may not be immediately apparent owing to the gradual nature of thermal dynamics. Monte Carlo methods, by evaluating the aggregate reward at the conclusion of an episode, could potentially provide a more accurate modeling of these delayed outcomes.

Temporal-difference TD(0): Derived from the bellman Eq. (9), this approach is fully categorized within model-free algorithms. Similar to Monte Carlo methods, it focuses on approximating the action-value function Q and the policy evaluation does not necessitate a detailed approximation of this function (i.e., the green arrow does not reach the objective fully during evaluation Fig. 8), and it incorporates an ϵ-greedy tactic for policy improvement. In contrast to Monte Carlo methods, which adjust the policy at the end of each episode, this approach operates on a one-step basis for evaluation and improvement, with the immediate target $r + \gamma Q(s', a')$ as the objective, as opposed to the total expected return G_t.

Two well-known algorithms are state–action–reward–state–action (SARSA) and Q-learning [14]; SARSA (Algorithm 3) is an on-policy algorithm where both the evaluation and improvement are applied to the ϵ-greedy policy in use. On the other hand, Q-learning

is an off-policy algorithm that operates using an ϵ-greedy policy for action selection while concurrently evaluating and improving a separate greedy policy.

Algorithm 3 SARSA

Initialize $Q(s,a), \forall s \in S, a \in A(s)$, arbitrarily, and $Q(\text{terminal-state}, \cdot) = 0$
for each episode **do**
  Initialize s
  Choose a from s using policy derived from Q (e.g., ϵ-greedy)
  repeat for each step of episode
    Take action a, observe r, s'
    Choose a' from s' using policy derived from Q (e.g., ϵ-greedy)
    $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$
    $s \leftarrow s'; a \leftarrow a';$
  until s is terminal
end for

Similar to Monte Carlo, TD(0) also uses function approximators, such as a NN, for the estimation of the action-value function. When adapting SARSA methods for approximation, the foundational method remains largely unaltered. In Algorithm 3, rather than following the original step 8, Eq. (18) is applied, with $r + \gamma \hat{Q}(s', a', w)$ serving as the target instead of G_t. This process involves the tuning of the neural network weights through SGD with the aim of reducing the loss function specified in Eq. (17), where G_t is replaced by $r + \gamma \hat{Q}(s', a', w)$.

A variety of algorithms are classified within the TD(0) category, including categorical 51-Atom (C51) [86], model-based interval estimation with exploration bonus (MBIE-EB) [87], and value-aware model learning (VAML) [88]. In the domain of HVAC system control, deep Q-networks (DQN) [89] stands as the most renowned TD(0) algorithm, representing an advanced iteration of the Q-learning algorithm. This variant employs a neural network to estimate the action-value function and incorporates techniques such as experience replay and fixed Q-targets to enhance learning stability and efficiency. Ref. [39] employs a DQN algorithm to regulate chilled water temperature set-points, aiming to minimize energy consumption while preserving thermal comfort in a single zone. Consequently, this approach exhibited improved performance in energy efficiency and control efficacy, surpassing rule-based control and traditional RL, and closely matching the performance of MPC.

In [52], a DQN algorithm is employed for managing the supply air temperature and chilled water temperature set-points in a multi-zone environment. This approach aims at minimizing energy usage while preserving thermal comfort. The study contrasts this DQN-based control system with conventional fixed temperature set-point controls. Findings reveal that the DQN method is superior in coordinating energy use with indoor air temperature, outperforming traditional controls. Particularly, the DQN demonstrates enhanced stability in control actions post-training, efficiently optimizing the balance between energy expenditure in HVAC systems and maintaining comfortable indoor air conditions.

3.2.2. Policy-based

In this methodology, the policy is parameterized directly to optimize the cumulative reward obtained from the environment, bypassing the need for value or action-value function estimation. Control over the policy parameters enables the manipulation of the action selection distribution (19). This allows for the learning of stochastic policies, proving highly efficient within the POMDP framework. It is particularly effective in environments with high-dimensional or continuous action spaces, demonstrated by the case of BEMS, where actions such as HVAC temperature settings, lighting dimming levels, renewable energy integration, operational timing, load shifting require continuous, and multi-dimensional decision-making capabilities. Numerous studies have aimed to make the action spaces of their agents continuous to provide a broader range of action choices at each timestep, thereby enabling more precise and effective management of energy consumption. For example, in [20], the action spaces for certain actuators, such as the supply air temperature and airflow setpoints of the AHU, are continuous. Similarly, [44] describes a continuous action space for the indoor heating setpoint, and [65] details continuous action spaces for both heating and cooling temperature setpoints across five zones.

$$\begin{aligned} &\pi_\theta : S \times \mathcal{A} \longrightarrow [0,1] \\ &\quad (s,a) \longmapsto \pi_\theta(s,a) = P(A_t = a \mid S_t = s, \theta) \end{aligned} \tag{19}$$

where π_θ is generally a NN with weights θ .

In this approach, the quest for the optimal policy involves training the neural policy, π_θ, with data derived from engaging with the environment. Here, the parameters, θ, are optimized through stochastic gradient ascent (SGA) to amplify a specified objective function, J [85]:

$$J(\theta) = \sum_s d^{\pi_\theta}(s) \sum_a \pi_\theta(s,a) r_s^a \tag{20}$$

where d^{π_θ} is stationary distribution of Markov chain for π_θ.

$$\Delta\theta = \alpha \nabla_\theta J(\theta) = \alpha \nabla_\theta \log \pi_\theta(s_t, a_t) v_t \tag{21}$$

where v_t is G_t (episodic updates) or r_t (per step updates).

Among the foundational algorithms in policy-based methods, REINFORCE is significant. It uses the cumulative rewards from entire episodes to iteratively refine the policy.

Algorithm 4 REINFORCE

```
Initialize policy parameters θ arbitrarily
for each episode {s_1, a_1, r_2, ..., s_{T-1}, a_{T-1}, r_T} ~ π_θ do
    for t = 1 to T − 1 do
        θ ← θ + α∇_θ log π_θ(s_t, a_t) G_t
    end for
end for
return θ
```

Additional algorithms categorized as model-free encompass the cross-entropy method (CEM) [90] and evolution strategies (ES) [91]. On the other hand, examples of model-based algorithms include model-based policy optimization (MBPO) [92], probabilistic inference for learning control (PILCO) [93], and guided policy search (GPS) [94].

Despite the popularity of policy-based approaches in the field of RL, it appears that, to the best of the author's knowledge, no algorithms from this category have been implemented in BEMS as of yet.

3.2.3. Actor–critic

This methodology combines elements of both value-based and policy-based strategies. In a policy-based manner, it directly parameterizes the policy to maximize the accumulated reward from the environment, employing the approximate policy gradient theorem (22) [85]. Furthermore, it adopts a value-based perspective by evaluating the current NN policy π_θ with the aid of a critic network, labeled $\hat{Q}(\cdot,\cdot,w)$, which serves to estimate the action-value function Q_{π_θ}.

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a)\hat{Q}(s,a,w)] \tag{22}$$

This strategy seeks the optimal policy by training the neural policy, π_θ, and the critic network, $\hat{Q}(\cdot,\cdot,w)$, using data acquired through interaction with the environment. Within each step of the episode, parameter optimization occurs in two phases: first, the parameters θ are fine-tuned via SGA (23) to increase the objective function J; subsequently, the parameters w undergo refinement through SGD (24) to decrease the loss function L.

$$\Delta\theta = \alpha\nabla_\theta \log \pi_\theta(s,a)\hat{Q}(s,a,w) \tag{23}$$

$$\Delta w = \alpha(r + \gamma\hat{Q}(s',a',w) - \hat{Q}(s,a,w))\nabla_w \hat{Q}(s,a,w) \tag{24}$$

In scenarios where the critic is solely defined by the action-value function $\hat{Q}$, as seen in (22), this can lead to significant variance. An improvement to this approach involves the adoption of an advantage function A(25), within the framework of advantage actor–critic algorithms, to potentially mitigate this issue. However, this modification introduces complexity, necessitating the calibration of two distinct parameter sets, w and v. To streamline this process, one may opt for the temporal difference actor–critic method, which employs the TD-Error (δ_v) (26) in place of the advantage function for the critic evaluation. This adjustment simplifies the optimization challenge by requiring the refinement of a single parameter set rather than two.

$$A(s,a) = \hat{Q}(s,a,w) - \hat{V}(s,v) \tag{25}$$

$$\delta_v = r + \hat{V}(s',v) - \hat{V}(s,v) \tag{26}$$

Numerous algorithms have emerged within the field of actor–critic, such as asynchronous advantage actor–critic (A2C/A3C) [95], deep deterministic policy gradient (DDPG) [96], twin delayed DDPG (TD3) [97], soft actor-critic (SAC) [98], proximal policy optimization (PPO) [99], alphaZero [18], imagination-augmented agents (I2 A) [100], trust region policy optimization (TRPO) [101], model ensemble trust region policy optimization (ME-TRPO) [102], hybrid-model-based DDPG (HMB-DDPG) [63], model-based value expansion (MVE) [103], gradient-aware model-based policy search (GAMPS) [104], and actor-critic with experience replay (ACER) [105].

In [62], the study utilizes a DDPG algorithm to optimize the temperature scheduling of an HVAC system in a single-zone setting. The primary objective is to lower the energy costs associated with HVAC operation while ensuring thermal comfort. This study compares the DDPG-controlled system with a controller that operates on a fixed temperature schedule. Results show that the DDPG system effectively reduces energy expenses and enhances occupant comfort.

In [51], an A3C algorithm is used for managing heating and cooling air temperature set-points, with the goal of decreasing electrical power demand while preserving thermal comfort in both single-zone and multi-zone environments. In certain test scenarios, this algorithm effectively learned a dynamic control policy, achieving a 12.8% reduction in HVAC systems energy use compared to traditional controls, while maintaining comfort levels. Nonetheless, the study revealed a tendency for over-fitting, indicating that future research should concentrate on enhancing the generalizability of DRL approaches.

In [54], a PPO algorithm is implemented to control indoor temperature set-points, with the objective of lowering HVAC system energy demand while ensuring thermal comfort across multiple zones. The findings indicate that PPO offers superior control performance in air conditioning systems compared to value optimization-based methods such as Q-learning and DQN algorithms.

This implies that PPO is more efficient in temperature regulation, leading to energy conservation while simultaneously upholding indoor comfort levels.

In [48], the study utilizes the SAC algorithm to regulate various components in a single-zone environment, including damper position, supply airflow, temperature set-point, and AHU fan speed. The primary objective of this approach is to decrease the energy usage of the HVAC system while ensuring thermal comfort is maintained. The research compares the effectiveness of the SAC-based control system with a conventional PID controller. In terms of measurable outcomes, the SAC controller attained a 17.4% reduction in energy usage and a 16.9% enhancement in thermal comfort relative to the performance of the existing PID controller.

3.3. Overview of algorithm implementation in current literature

In reference to Fig. 9, the current trend in literature is the adoption of value-based approaches, indicative of a preference for control within a discrete action space. The dominant algorithm in this category is DQN, signifying an inclination towards management in extensive or continuous state spaces. Actor–critic methods also represent a significant portion, showing a preference for controlling continuous state/action spaces, with DDPG and SAC being the leading algorithms in this area. Finally, based on this literature study, policy-based approaches have seen limited use, aligning with the scarcity of algorithms in this category.

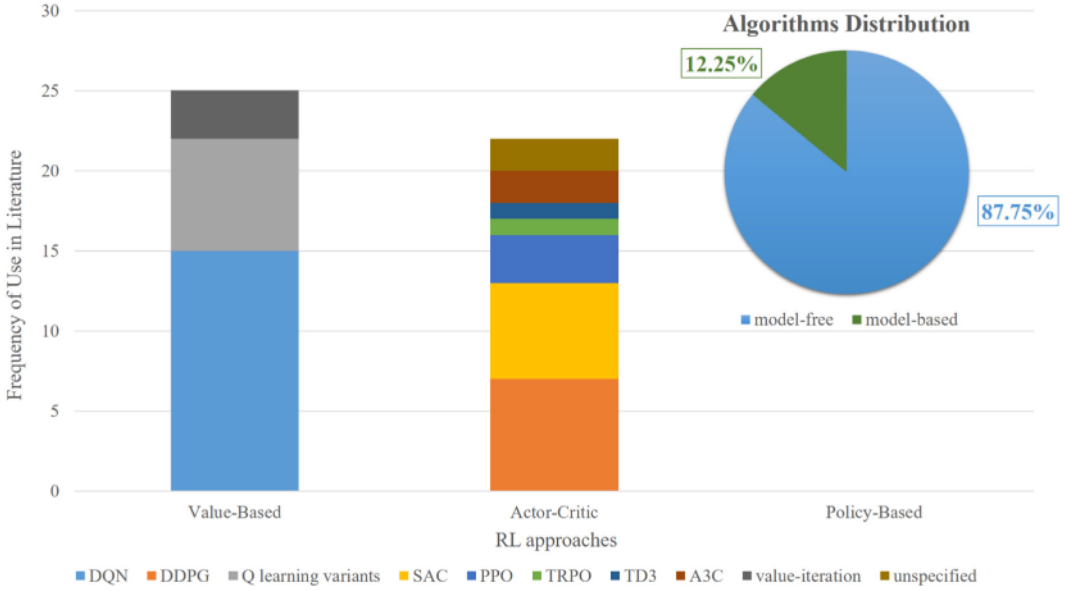

Fig. 9. The categorization of RL algorithms employed for HVAC control. Any algorithm listed in Table 3 that is recognized as a variant of one of the algorithms depicted in the picture can be classified under its standard name (DQN, DDPG, . . . , value-iteration).

Model-free RL algorithms emerge as the predominant type of controllers. Their popularity can be attributed to their simplicity, adaptability, and resilience, especially in situations involving complex or unknown environmental dynamics. In contrast, model-based algorithms, despite their inherent sample efficiency — an advantageous trait in scenarios where environmental interaction is expensive or restricted — only constitute 12.25% of the controllers used. This lower prevalence is likely due to the fact that most training processes occur within simulated environments, rendering the sample-efficiency of model-based approaches less critical.

Fig. 8 reveals that in existing literature, the most common configuration for RL controllers in HVAC applications involves using 15 min as the duration for a single time-step and one day for an episode length. This setup is logical: temperature changes require time to be observed, making the 15-minute time-step suitable, while the daily cycle of actions aligns well with the episodic structure. However, numerous studies also use one hour as a time-step, which could be excessively long. Remarkably, some papers extend the episode length to an entire year, or even five years. This extended duration may lead to significant disruptions, especially with sudden changes in predefined exogenous state vector components. The likelihood of such disruptions escalates with longer episode duration.

Table 3 indicates that the predominant approach in addressing the HVAC system control problem is to employ a single RL agent, as evidenced by 33 studies. In contrast, only 9 studies utilize multi-agent RL. This discrepancy can be attributed to the high computational burden associated with training multiple RL agents, which contradicts the goals of BEMS. According to Table 4, it is evident that only 10 papers implement the agent in a physical building, compared to 32 papers that utilize a virtual environment. The difference is reasonable as deploying a reinforcement learning agent in a real structure poses safety challenges, because in real building, the problem is framed as a SNSMDP whereas the agent is trained within a SSMDP framework.

4. Discussion

This section aims to investigate the advantages and limitations of RL algorithms in HVAC applications, providing valuable insights for researchers and engineers. It also aims to shed light on prospective future developments in RL implementations in this domain.

4.1. Benefits of employing RL

When the training environment in simulations accurately replicates real-world scenarios, RL controllers exhibit a significant advantage over traditional and MPC methods in HVAC systems. Post the intensive computational training phase, an RL agent acquires the capability to execute actions based on its established policy, doing so with minimal computational burden at each time step. Moreover, these agents are adept at managing multiple objectives concurrently and possess an 'offline' ability to anticipate future scenarios.

These attributes position RL agents as an ideal solution in complex settings, enabling them to devise innovative action sequences that optimally reduce energy consumption.

Incorporating RL controllers into HVAC systems has proven to significantly enhance energy efficiency and cost reduction, outperforming traditional control methods. Particularly, RL controllers exhibit a reduction in energy usage by 27%–30% compared to rule-based controllers, while still maintaining optimal thermal comfort levels [12]. They also offer considerable cost savings, up to 39.6% over standard controllers, while preserving thermal comfort [20]. Alongside heating energy savings ranging from 5%–12%, with the added benefit of improved interior climate control [42]. Additionally, these controllers achieve 8% energy savings relative to rule-based algorithms, ensuring comfort for both indoor environments and domestic hot water [34]. Furthermore, RL controllers are effective in reducing electricity costs and peak energy demands by 23% in comparison to manually designed rule-based systems, without compromising on thermal comfort [45].

4.2. Limitations & future directions

4.2.1. Restricted set of algorithms

A primary limitation in current studies on RL algorithms for HVAC control systems is their reliance on popular RL algorithms successful in gaming, such as DQN and DDPG. These studies often overlook other RL approaches (e.g., Monte-Carlo, dynamic programming, model-based, policy-based) (Fig. 7) and algorithms (e.g., SARSA, REINFORCE, policy iteration, ...), which might yield superior results in the distinctively dynamic environment of thermal control, as opposed to game environments. Therefore, further development and exploration of these less-utilized approaches and algorithms for HVAC control, followed by comparative analysis with well-known ones like DQN and DDPG, present a valuable direction for future research.

4.2.2. Virtual environments-challenges and prospects

The most apparent limitation identified in our research is the requirement to develop a comprehensive virtual environment that accurately mimics the behaviors of various building environments. This element is essential for agent training, yet simultaneously it demands a considerable amount of time, especially when implementing white-box and grey-box modeling methods. As indicated before, these methods are predominantly preferred by researchers (Fig. 5) due to their enhanced accuracy and the extensive exploratory freedom they provide to the agent. This is in contrast to black-box modeling techniques, which are reliant on real-world input data that is typically quite restricted in scope. Nevertheless, conducting a comparative analysis of outcomes from two identical RL agents, each trained in distinct simulation environments one in a white-box or grey-box and the other in a black-box setting could constitute a valuable research direction.

4.2.3. Adapting to real-world challenges

Another significant limitation identified in numerous research papers (78% of papers according to Table 3) is the reliance on a static, predefined sequence of ESC (such as weather conditions, occupancy schedules, electricity prices, etc.) during the training phase. This approach can lead to challenges during real-world deployment, where these external components may shift suddenly and dramatically, deviating from the scenarios simulated in training [81]. Such abrupt changes can result in inappropriate actions by the agent, potentially leading to thermal regulation issues or, in more severe cases, damaging electrical components in actual buildings. Furthermore, the capability of online adaptation inherent in RL may not effectively address these challenges due to the real-time nature of these episodes, which can delay the agent ability to adjust to new conditions promptly.

To address this issue, a feasible solution involves fitting every potential transition sequence that could occur during a real deployment episode into each simulated episode within the virtual environment. This approach involves adjusting the frequency of specific transition sequences in simulated episodes to be proportional to their likelihood of occurrence in the real world. By adopting this strategy, the dynamics of the simulated environment align more closely to the real world case with an virtual distribution $\mathcal{P}$, effectively transforming the virtual model into a realistic SSMDP [35,44,48,50,59,67]. To achieve this, we recommend using all available yearly historical ESC data transitions for the specific target building and location, which includes variables like outdoor temperature, electricity price, and occupancy schedules. Compress all this yearly data into a single package, and augment it with synthetic data representing potential unforeseen future changes. Consequently, a training episode in this context differs from traditional RL episodes. Each new training episode involves a stochastic pass through the data package, ensuring each sample

Table 3
Reinforcement learning algorithms.

Ref	Algorithm	Number of training episode	Time steps	Single-Agent	Multi-Agent	Single-ESC[a]	Multi-ESC[b]
[39]	DQN	20	1 h over 102 days	✓		✓	
[40]	value-iteration	–	1 min over 2 days		✓	✓	
[41]	value-iteration	–	–	✓		✓	
[31]	(actor-critic) with some variants of relearning	–	1 h over 1 day	✓			✓
[34]	DQN	–	5 min over 8 months	✓		✓	
[42]	DQN	50	15 min over 2 months	✓		✓	
[20]	DDPG	–	15 min over 1 year	✓		✓	
[35]	DQN	Single-zone: 400 Multi-zone: 250	15 min over 3 months	✓			✓
[43]	Q-learning (table lookup)	1000	(10:00 AM, 2:00 PM & 6:00 PM over 40 days)	✓		✓	
[44]	DDPG	300	1 h over 1 day	✓			✓
[45]	SAC	30	1 h over 3 months	✓		✓	
[12]	DDPG	200	15 min over 3 months		✓	✓	
[46]	DDQN,SAC, MBDDQN,MBSAC	50	1 min over 2 weeks	✓		✓	
[38]	ED-DQN	100	30 min over 4 months	✓		✓	
[47]	Q-learning (table-lookup)	1000	30 min over 1 day	✓		✓	
[48]	SAC	120	5 min over 1 day	✓			✓
[49]	MARL (Q-learning variant)	–	–		✓		
[50]	SAC	non-episodic	12 min over 20 days	✓			✓
[51]	A3C	20 millions timesteps 772 episodes	5 min over 3 month		✓	✓	
[52]	DQN	30	10 min over 1 month	✓		✓	
[53]	DQN	7	15 min over 1 year	✓			✓
[54]	PPO	200	15 min over 45 days	✓		✓	
[33]	DP-NARX-RL	1000	15 min (Residential) & 4 min (airport)	✓		✓	
[55]	DQN	–	–	✓		✓	
[56]	DQN & DDPG	DDPG : 300 DQN : 75	DDPG : 1 h over 1 month DQN : 5 min over 3 month	✓		✓	
[57]	MARL DDQN (DQN variant)	–	2 h 30 min over 280 days		✓	✓	
[58]	TQBER (tabular Q-learning variant)	–	5 min over 1 year	✓		✓	
[59]	SAC, TD3, TRPO and PPO	20	15 min over 1 year	✓			✓
[60]	DQN	100	15 min over 1 month		✓	✓	
[61]	DQN	–	15 min over 5 year	✓		✓	
[62]	DDPG	–	5 min over 212 days	✓		✓	
[63]	HMB-DDPG (model-based)	2880	–	✓		✓	
[64]	DQN	–	20 min over 110 days	✓	✓	✓	
[37]	MAAC (SAC variant)	5000	15 min over 1 day		✓	✓	
[65]	PPO	1000	15 min over 2 months	✓		✓	
[66]	DQN	50	15 min over 2 months	✓		✓	
[67]	DDPG	3000	1 h over 1 day (for 2 month)	✓			✓
[68]	DQN	1000	15 min over 11 days in July	✓		✓	
[36]	Q-learning variant	–	1 h over 1 day		✓	✓	
[69]	A3C	3.25 millions timesteps 376 episodes	15 min over 3 months	✓		✓	
[70]	Q-learning	–	20 min over 1 year	✓		✓	
[71]	MCAC (Monte Carlo actor-critic)	8 episodes	5 min over 2 days	✓		✓	

[a] This means that the same sequences of exogenous state components are used in all episodes.
[b] This means that various sequences of exogenous state components are used across different episodes.

Table 4
Summary of models.

Ref	Office	Commercial	Residential	University	Data centers	*NA*	Single-zone	Multi-zone	Virtual	Real	EnergyPlus	Trnsys	Tesp	Boptest	Matlab	Dymola	Iesve	Citylearn	DNN	*NA*
[39]						✓	✓		✓		✓								✓	
[40]		✓						✓		✓										✓
[41]		✓						✓		✓										✓
[31]						✓		✓		✓										✓
[34]			✓				✓		✓								✓			
[42]	✓						✓		✓		✓									
[20]		✓					✓			✓	✓									
[35]		✓					✓	✓	✓		✓									
[43]						✓			✓			✓								
[44]			✓					✓	✓											✓
[45]		✓						✓	✓		✓							✓	✓	
[12]	✓							✓	✓										✓	
[46]			✓				✓		✓					✓						
[38]			✓					✓	✓											✓
[47]	✓							✓	✓											✓
[48]	✓						✓		✓										✓	
[49]						✓		✓		✓									✓	
[50]		✓						✓		✓										✓
[51]	✓						✓	✓	✓		✓									
[52]	✓							✓	✓		✓									
[53]		✓					✓		✓		✓									
[54]				✓				✓	✓			✓								
[33]		✓	✓				✓		✓						✓					
[55]		✓						✓		✓	✓									
[56]			✓					✓	✓	✓										✓
[57]			✓					✓	✓							✓				
[58]			✓				✓		✓		✓									
[59]					✓			✓	✓		✓									
[60]	✓				✓			✓	✓		✓									
[61]	✓						✓		✓											✓
[62]						✓	✓		✓				✓							
[63]						✓		✓	✓											✓
[64]	✓						✓	✓	✓			✓								
[37]		✓						✓	✓											✓
[65]		✓						✓	✓		✓									
[66]			✓					✓		✓										✓
[67]			✓				✓		✓											✓
[68]	✓							✓	✓		✓									
[36]				✓				✓	✓										✓	
[69]	✓						✓			✓	✓									
[70]			✓				✓		✓											✓
[71]	✓						✓		✓		✓									

(classical episode of yearly transitions) is presented exactly once in a random sequence. This approach minimizes correlation and improves the accuracy of the virtual distribution $\mathcal{P}$.

Even with the perturbation issue addressed by the virtual distribution $\mathcal{P}$, and despite accounting for all potential surprises, the possibility of sudden, abnormal changes in the real-world environment remains (i.e., SNSMDP), particularly in exogenous components like occupancy schedules and electricity prices. These changes can render the existing virtual $\mathcal{P}$ outdated, necessitating the retraining or fine-tuning of the agent in an updated simulator rather than in the actual building [31]. However, this process is computationally demanding, which contradicts the agent primary goal of reducing energy consumption. This challenge steers the research towards discovering and refining an optimal RL method that conducts the retraining process with minimal computational and energy resources. The meta-reinforcement learning approach (Meta-RL) emerges as a particularly promising solution in this context.

Meta-RL [106,107] trains an agent on a wide range of tasks and scenarios, equipping it with a form of meta-knowledge. This meta-knowledge allows the agent to quickly adapt its strategies based on the new conditions it encounters, using its understanding of how to learn efficiently in various contexts. Unlike standard RL agents that might require expensive retraining or retraining from scratch when encountering new environments, Meta-RL agents can make quick adjustments to their behavior [108]. This is because they have been trained to anticipate and adapt to changes, often requiring only minimal additional offline learning to handle new situations effectively.

Another recommendation to address the two main problems above that initially originate from the extended duration of single episodes (e.g., 1 year, 6 months, 3 months), which increase the likelihood of significant future changes in ESC, is to reduce the length of each episode to the shortest feasible period while preserving a cyclic pattern. The simulation model should be updated regularly with all relevant ESC data for that particular period, including weather conditions from meteorological centers, electricity pricing from local utility providers, and occupancy schedules from calenders or prediction models. The agent should then be trained on this simulation model just before the real start of the period when real deployment happens, with this process being repeated frequently, allowing for the fine-tuning of the agent based on the prior period's learning.

This methodology should resolve our two main challenges; however, it introduces a new issue: significant cumulative electricity consumption from the short-period retraining (or fine-tuning) of the agent, driven by GPUs or CPUs. Although the consumption is minor on a short-period basis, it builds up over time. Further research should be directed towards this methodology to examine if the electricity consumption required for fine-tuning over the shortest feasible period decreases incrementally, eventually becoming negligible. This potential reduction is particularly due to the 'remember' capability in reinforcement learning, which is enhanced by consistent frequent fine-tuning.

5. Conclusion

In conclusion, this study presented a comprehensive literature review of the application of RL in HVAC systems after 2019, highlighting its potential to better optimize and control BEMS. However, several significant challenges remain to be resolved. In this investigation, papers were reviewed from two perspectives; (1) Markov Decision Process (MDP): this represents the model of the environment, where novel concepts of endogenous and exogenous state components are clearly defined, where endogenous components are influenced by actions, while exogenous components are predetermined. This definition led to a novel classification of MDP categories, depending on the form of the probability distribution $\mathcal{P}$, which represents the dynamics of the interaction environment. It was found that any virtual building environment is categorized as SSMDP due to the static nature of their probability distribution, and thus can be resolved by any RL algorithm from the three traditional approaches (value-based, policy-based, and actor–critic). In contrast, real-world cases are categorized as SNSMDP. (2) Reinforcement Learning Algorithms: this pertains to the 'brain' of the active agent. In all studies, the agent is trained in a virtual environment SSMDP before real deployment, explaining the use of traditional algorithms (DQN, QPG, SAC, etc.). However, theoretically, agents trained with these algorithms may not perform well during real deployment due to the change in the environment model to SNSMDP. The only solution to fix the disturbance in the agent's performance is to undergo periodic extensive retraining, which is computationally expensive. In future works, the focus should be on addressing the high computational cost associated with retraining the RL agent before it is deployed in an actual building for HVAC system control, which is modeled as an SNSMDP. In this context, the use of a meta-RL technique is proposed. This approach has been theoretically proven to facilitate rapid adaptation to the disturbances generated by the nature of SNSMDP during real-world deployment. Meta-RL can improve adaptability and reduce the computational cost associated with retraining. Future research should focus on applying Meta-RL to HVAC system control.

CRediT authorship contribution statement

Khalil Al Sayed: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Abhinandana Boodi:** Writing – review & editing, Validation, Supervision, Resources, Conceptualization. **Roozbeh Sadeghian Broujeny:** Writing – review & editing, Validation, Supervision, Resources, Conceptualization. **Karim Beddiar:** Writing – review & editing, Validation, Supervision, Resources, 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.

Data availability

Data will be made available on request.

References

[1] IEA, Buildings, IEA, Paris, 2023, https://www.iea.org/energy-system/buildings, License: CC BY 4.0.
[2] IEA, Heating, IEA, Paris, 2023, https://www.iea.org/energy-system/buildings/heating, License: CC BY 4.0.
[3] IEA, Space Cooling, IEA, Paris, 2022, https://www.iea.org/energy-system/buildings/space-cooling, License: CC BY 4.0.
[4] United Nations Environment Programme, 2022 Global status report for buildings and construction: Towards a zero-emission, efficient and resilient buildings and construction sector, 2022, URL https://wedocs.unep.org/20.500.11822/41133.
[5] A. Afram, F. Janabi-Sharifi, Theory and applications of HVAC control systems – A review of model predictive control (MPC), Build. Environ. (ISSN: 0360-1323) 72 (2014) 343–355, http://dx.doi.org/10.1016/j.buildenv.2013.11.016, URL https://www.sciencedirect.com/science/article/pii/S0360132313003363.
[6] K. Chinnakani, A. Krishnamurthy, J. Moyne, A. Arbor, F. Gu, Comparison of energy consumption in HVAC systems using simple ON-OFF, intelligent ON-OFF and optimal controllers, in: 2011 IEEE Power and Energy Society General Meeting, 2011, pp. 1–6, http://dx.doi.org/10.1109/PES.2011.6039823.
[7] Y. Wang, Y. Shao, C. Kargel, Demand controlled ventilation strategies for high indoor air quality and low heating energy demand, in: 2012 IEEE International Instrumentation and Measurement Technology Conference Proceedings, 2012, pp. 870–875, http://dx.doi.org/10.1109/I2MTC.2012.6229554.
[8] M. Esrafilian-Najafabadi, F. Haghighat, Occupancy-based HVAC control systems in buildings: A state-of-the-art review, Build. Environ. (ISSN: 0360-1323) 197 (2021) 107810, http://dx.doi.org/10.1016/j.buildenv.2021.107810, URL https://www.sciencedirect.com/science/article/pii/S0360132321002171.
[9] J. Arroyo, C. Manna, F. Spiessens, L. Helsen, Reinforced model predictive control (RL-MPC) for building energy management, Appl. Energy (ISSN: 0306-2619) 309 (2022) 118346, http://dx.doi.org/10.1016/j.apenergy.2021.118346, URL https://www.sciencedirect.com/science/article/pii/S0306261921015932.
[10] Y. Balali, A. Chong, A. Busch, S. O'Keefe, Energy modelling and control of building heating and cooling systems with data-driven and hybrid models—A review, Renew. Sustain. Energy Rev. (ISSN: 1364-0321) 183 (2023) 113496, http://dx.doi.org/10.1016/j.rser.2023.113496, URL https://www.sciencedirect.com/science/article/pii/S1364032123003532.
[11] Z. Afroz, G. Shafiullah, T. Urmee, G. Higgins, Modeling techniques used in building HVAC control systems: A review, Renew. Sustain. Energy Rev. (ISSN: 1364-0321) 83 (2018) 64–84, http://dx.doi.org/10.1016/j.rser.2017.10.044, URL https://www.sciencedirect.com/science/article/pii/S1364032117314193.
[12] Z. Zou, X. Yu, S. Ergan, Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network, Build. Environ. 168 (2020) 106535.
[13] R.Z. Homod, Review on the HVAC system modeling types and the shortcomings of their application, in: M. Benghanem (Ed.), J. Energy (ISSN: 2356-735X) 2013 (2013) 768632, http://dx.doi.org/10.1155/2013/768632.
[14] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, 2018.

[15] G. Novati, P. Koumoutsakos, Remember and forget for experience replay, 2019, arXiv:1807.05827.

[16] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, D.G. van den, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Mastering the game of Go with deep neural networks and tree search, Nature (ISSN: 1476-4687) 529 (7587) (2016) 484–489, http://dx.doi.org/10.1038/nature16961.

[17] A. Shantia, E. Begue, M. Wiering, Connectionist reinforcement learning for intelligent unit micro management in StarCraft, in: The 2011 International Joint Conference on Neural Networks, 2011, pp. 1794–1801, http://dx.doi.org/10.1109/IJCNN.2011.6033442.

[18] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al., Mastering chess and shogi by self-play with a general reinforcement learning algorithm, 2017, arXiv preprint arXiv:1712.01815.

[19] S. Sierla, H. Ihasalo, V. Vyatkin, A review of reinforcement learning applications to control of heating, ventilation and air conditioning systems, Energies (ISSN: 1996-1073) 15 (10) (2022) http://dx.doi.org/10.3390/en15103526, URL https://www.mdpi.com/1996-1073/15/10/3526.

[20] S. Touzani, A.K. Prakash, Z. Wang, S. Agarwal, M. Pritoni, M. Kiran, R. Brown, J. Granderson, Controlling distributed energy resources via deep reinforcement learning for load flexibility and energy efficiency, Appl. Energy 304 (2021) 117733.

[21] H. Zhang, S. Seal, D. Wu, F. Bouffard, B. Boulet, Building energy management with reinforcement learning and model predictive control: A survey, IEEE Access 10 (2022) 27853–27862, http://dx.doi.org/10.1109/ACCESS.2022.3156581.

[22] J.A. Rossiter, Model-Based Predictive Control: a Practical Approach, CRC Press, 2017.

[23] J.R. Vázquez-Canteli, Z. Nagy, Reinforcement learning for demand response: A review of algorithms and modeling techniques, Appl. Energy (ISSN: 0306-2619) 235 (2019) 1072–1089, http://dx.doi.org/10.1016/j.apenergy.2018.11.002, URL https://www.sciencedirect.com/science/article/pii/S0306261918317082.

[24] A. Chatterjee, D. Khovalyg, Dynamic indoor thermal environment using reinforcement learning-based controls: Opportunities and challenges, Build. Environ. (ISSN: 0360-1323) 244 (2023) 110766, http://dx.doi.org/10.1016/j.buildenv.2023.110766, URL https://www.sciencedirect.com/science/article/pii/S036013232300793X.

[25] A. Shaqour, A. Hagishima, Systematic review on deep reinforcement learning-based energy management for different building types, Energies 15 (22) (2022) 8663.

[26] Q. Fu, Z. Han, J. Chen, Y. Lu, H. Wu, Y. Wang, Applications of reinforcement learning for building energy efficiency control: A review, J. Build. Eng. 50 (2022) 104165.

[27] Z. Wang, T. Hong, Reinforcement learning for building controls: The opportunities and challenges, Appl. Energy 269 (2020) 115036.

[28] D. Denyer, D. Tranfield, Producing a Systematic Review, Sage Publications Ltd, 2009.

[29] A. Liberati, D.G. Altman, J. Tetzlaff, C. Mulrow, P.C. Gøtzsche, J.P.A. Ioannidis, M. Clarke, P.J. Devereaux, J. Kleijnen, D. Moher, The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration, BMJ (ISSN: 0959-8138) 339 (2009) http://dx.doi.org/10.1136/bmj.b2700, URL https://www.bmj.com/content/339/bmj.b2700.

[30] J. Moos, K. Hansel, H. Abdulsamad, S. Stark, D. Clever, J. Peters, Robust reinforcement learning: A review of foundations and recent advances, Mach. Learn. Knowl. Extr. 4 (1) (2022) 276–315.

[31] A. Naug, M. Quinones-Grueiro, G. Biswas, Deep reinforcement learning control for non-stationary building energy management, Energy Build. (ISSN: 0378-7788) 277 (2022) 112584, http://dx.doi.org/10.1016/j.enbuild.2022.112584, URL https://www.sciencedirect.com/science/article/pii/S0378778822007551.

[32] E. Lecarpentier, Reinforcement Learning in Non-Stationary Environments (Ph.D. thesis), Toulouse, ISAE, 2020.

[33] S.M. Dawood, A. Hatami, R.Z. Homod, Trade-off decisions in a novel deep reinforcement learning for energy savings in HVAC systems, J. Build. Perform. Simul. 15 (6) (2022) 809–831, http://dx.doi.org/10.1080/19401493.2022.2099465.

[34] P. Lissa, C. Deane, M. Schukat, F. Seri, M. Keane, E. Barrett, Deep reinforcement learning for home energy management system control, Energy AI 3 (2021) 100043.

[35] X. Deng, Y. Zhang, H. Qi, Towards optimal HVAC control in non-stationary building environments combining active change detection and deep reinforcement learning, Build. Environ. 211 (2022) 108680.

[36] J. Hao, D.W. Gao, J.J. Zhang, Reinforcement learning for building energy optimization through controlling of central HVAC system, IEEE Open Access J. Power Energy 7 (2020) 320–328, http://dx.doi.org/10.1109/OAJPE.2020.3023916.

[37] L. Yu, Y. Sun, Z. Xu, C. Shen, D. Yue, T. Jiang, X. Guan, Multi-agent deep reinforcement learning for HVAC control in commercial buildings, IEEE Trans. Smart Grid 12 (1) (2021) 407–419, http://dx.doi.org/10.1109/TSG.2020.3011739.

[38] Q. Fu, Z. Li, Z. Ding, J. Chen, J. Luo, Y. Wang, Y. Lu, ED-DQN: An event-driven deep reinforcement learning control method for multi-zone residential buildings, Build. Environ. (ISSN: 0360-1323) 242 (2023) 110546, http://dx.doi.org/10.1016/j.buildenv.2023.110546, URL https://www.sciencedirect.com/science/article/pii/S0360132323005735.

[39] K. He, Q. Fu, Y. Lu, Y. Wang, J. Luo, H. Wu, J. Chen, Predictive control optimization of chiller plants based on deep reinforcement learning, J. Build. Eng. (ISSN: 2352-7102) 76 (2023) 107158, http://dx.doi.org/10.1016/j.jobe.2023.107158, URL https://www.sciencedirect.com/science/article/pii/S2352710223013384.

[40] S. Park, S. Park, M.-i. Choi, S. Lee, T. Lee, S. Kim, K. Cho, S. Park, Reinforcement learning-based BEMS architecture for energy usage optimization, Sensors 20 (17) (2020) 4918.

[41] I. Tahir, A. Nasir, A. Algethami, Optimal control policy for energy management of a commercial bank, Energies 15 (6) (2022) 2112.

[42] S. Brandi, M.S. Piscitelli, M. Martellacci, A. Capozzoli, Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings, Energy Build. 224 (2020) 110225.

[43] C. Correa-Jullian, E.L. Droguett, J.M. Cardemil, Operation scheduling in a solar thermal system: A reinforcement learning-based framework, Appl. Energy 268 (2020) 114943.

[44] Y. Du, H. Zandi, O. Kotevska, K. Kurte, J. Munk, K. Amasyali, E. Mckee, F. Li, Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning, Appl. Energy 281 (2021) 116117.

[45] G. Pinto, D. Deltetto, A. Capozzoli, Data-driven district energy management with surrogate models and deep reinforcement learning, Appl. Energy 304 (2021) 117642.

[46] C. Gao, D. Wang, Comparative study of model-based and model-free reinforcement learning control performance in HVAC systems, J. Build. Eng. (ISSN: 2352-7102) 74 (2023) 106852, http://dx.doi.org/10.1016/j.jobe.2023.106852, URL https://www.sciencedirect.com/science/article/pii/S2352710223010318.

[47] W. Li, Y. Zhao, J. Zhang, C. Jiang, S. Chen, L. Lin, Y. Wang, Indoor temperature preference setting control method for thermal comfort and energy saving based on reinforcement learning, J. Build. Eng. (ISSN: 2352-7102) 73 (2023) 106805, http://dx.doi.org/10.1016/j.jobe.2023.106805, URL https://www.sciencedirect.com/science/article/pii/S2352710223009841.

[48] D. Zhuang, V.J. Gan, Z. Duygu Tekler, A. Chong, S. Tian, X. Shi, Data-driven predictive control for smart HVAC system in IoT-integrated buildings with time-series forecasting and reinforcement learning, Appl. Energy (ISSN: 0306-2619) 338 (2023) 120936, http://dx.doi.org/10.1016/j.apenergy.2023.120936, URL https://www.sciencedirect.com/science/article/pii/S0306261923003008.

[49] C. Blad, S. Bøgh, C. Kallesøe, P. Raftery, A laboratory test of an offline-trained multi-agent reinforcement learning algorithm for heating systems, Appl. Energy (ISSN: 0306-2619) 337 (2023) 120807, http://dx.doi.org/10.1016/j.apenergy.2023.120807, URL https://www.sciencedirect.com/science/article/pii/S030626192300171X.

[50] X. Lin, D. Yuan, X. Li, Reinforcement learning with dual safety policies for energy savings in building energy systems, Buildings (ISSN: 2075-5309) 13 (3) (2023) http://dx.doi.org/10.3390/buildings13030580, URL https://www.mdpi.com/2075-5309/13/3/580.

[51] X. Zhong, Z. Zhang, R. Zhang, C. Zhang, End-to-end deep reinforcement learning control for HVAC systems in office buildings, Designs (ISSN: 2411-9660) 6 (3) (2022) http://dx.doi.org/10.3390/designs6030052, URL https://www.mdpi.com/2411-9660/6/3/52.

[52] X. Fang, G. Gong, G. Li, L. Chun, P. Peng, W. Li, X. Shi, X. Chen, Deep reinforcement learning optimal control strategy for temperature setpoint real-time reset in multi-zone building HVAC system, Appl. Therm. Eng. (ISSN: 1359-4311) 212 (2022) 118552, http://dx.doi.org/10.1016/j.applthermaleng.2022.118552, URL https://www.sciencedirect.com/science/article/pii/S1359431122005038.

[53] A. Dmitrewski, M. Molina-Solana, R. Arcucci, CntrlDA: A building energy management control system with real-time adjustments. Application to indoor temperature, Build. Environ. (ISSN: 0360-1323) 215 (2022) 108938, http://dx.doi.org/10.1016/j.buildenv.2022.108938, URL https://www.sciencedirect.com/science/article/pii/S0360132322001810.

[54] Z. Li, Z. Sun, Q. Meng, Y. Wang, Y. Li, Reinforcement learning of room temperature set-point of thermal storage air-conditioning system with demand response, Energy Build. (ISSN: 0378-7788) 259 (2022) 111903, http://dx.doi.org/10.1016/j.enbuild.2022.111903, URL https://www.sciencedirect.com/science/article/pii/S0378778822000743.

[55] D. Xu, Learning efficient dynamic controller for HVAC System, in: M.A. Jan (Ed.), Mob. Inf. Syst. (ISSN: 1574-017X) 2022 (2022) 4157511, http://dx.doi.org/10.1155/2022/4157511.

[56] Y. Du, F. Li, K. Kurte, J. Munk, H. Zandi, Demonstration of intelligent HVAC load management with deep reinforcement learning: Real-world experience of machine learning in demand control, IEEE Power Energy Mag. 20 (3) (2022) 42–53, http://dx.doi.org/10.1109/MPE.2022.3150825.

[57] C. Blad, S. Bøgh, C. Kallesøe, A multi-agent reinforcement learning approach to price and comfort optimization in HVAC-systems, Energies (ISSN: 1996-1073) 14 (22) (2021) http://dx.doi.org/10.3390/en14227491, URL https://www.mdpi.com/1996-1073/14/22/7491.

[58] B. Huchuk, S. Sanner, W. O'Brien, Development and evaluation of data-driven controls for residential smart thermostats, Energy Build. (ISSN: 0378-7788) 249 (2021) 111201, http://dx.doi.org/10.1016/j.enbuild.2021.111201, URL https://www.sciencedirect.com/science/article/pii/S0378778821004850.

[59] M. Biemann, F. Scheller, X. Liu, L. Huang, Experimental evaluation of model-free reinforcement learning algorithms for continuous HVAC control, Appl. Energy (ISSN: 0306-2619) 298 (2021) 117164, http://dx.doi.org/10.1016/j.apenergy.2021.117164, URL https://www.sciencedirect.com/science/article/pii/S0306261921005961.

[60] T. Wei, S. Ren, Q. Zhu, Deep reinforcement learning for joint datacenter and HVAC load control in distributed mixed-use buildings, IEEE Trans. Sustain. Comput. 6 (3) (2021) 370–384, http://dx.doi.org/10.1109/TSUSC.2019.2910533.

[61] Z. Jiang, M.J. Risbeck, V. Ramamurti, S. Murugesan, J. Amores, C. Zhang, Y.M. Lee, K.H. Drees, Building HVAC control with reinforcement learning for reduction of energy cost and demand charge, Energy Build. (ISSN: 0378-7788) 239 (2021) 110833, http://dx.doi.org/10.1016/j.enbuild.2021.110833, URL https://www.sciencedirect.com/science/article/pii/S0378778821001171.

[62] B. Liu, M. Akcakaya, T.E. Mcdermott, Automated control of transactive HVACs in energy distribution systems, IEEE Trans. Smart Grid 12 (3) (2021) 2462–2471, http://dx.doi.org/10.1109/TSG.2020.3042498.

[63] H. Zhao, J. Zhao, T. Shu, Z. Pan, Hybrid-model-based deep reinforcement learning for heating, ventilation, and air-conditioning control, Front. Energy Res. (ISSN: 2296-598X) 8 (2021) http://dx.doi.org/10.3389/fenrg.2020.610518, URL https://www.frontiersin.org/articles/10.3389/fenrg.2020.610518.

[64] X. Yuan, Y. Pan, J. Yang, W. Wang, Z. Huang, Study on the application of reinforcement learning in the operation optimization of HVAC system, Build. Simul. (ISSN: 1996-8744) 14 (1) (2021) 75–87, http://dx.doi.org/10.1007/s12273-020-0602-9.

[65] D. Azuatalam, W.-L. Lee, F. de Nijs, A. Liebman, Reinforcement learning for whole-building HVAC control and demand response, Energy AI (ISSN: 2666-5468) 2 (2020) 100020, http://dx.doi.org/10.1016/j.egyai.2020.100020, URL https://www.sciencedirect.com/science/article/pii/S2666546820300203.

[66] K. Kurte, J. Munk, O. Kotevska, K. Amasyali, R. Smith, E. McKee, Y. Du, B. Cui, T. Kuruganti, H. Zandi, Evaluating the adaptability of reinforcement learning based HVAC control for residential houses, Sustainability (ISSN: 2071-1050) 12 (18) (2020) http://dx.doi.org/10.3390/su12187727, URL https://www.mdpi.com/2071-1050/12/18/7727.

[67] L. Yu, W. Xie, D. Xie, Y. Zou, D. Zhang, Z. Sun, L. Zhang, Y. Zhang, T. Jiang, Deep reinforcement learning for smart home energy management, IEEE Internet Things J. 7 (4) (2020) 2751–2762, http://dx.doi.org/10.1109/JIOT.2019.2957289.

[68] K.U. Ahn, C.S. Park, Application of deep Q-networks for model-free optimal control balancing between different HVAC systems, Sci. Technol. Built Environ. 26 (1) (2020) 61–74, http://dx.doi.org/10.1080/23744731.2019.1680234.

[69] Z. Zhang, A. Chong, Y. Pan, C. Zhang, K.P. Lam, Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning, Energy Build. (ISSN: 0378-7788) 199 (2019) 472–490, http://dx.doi.org/10.1016/j.enbuild.2019.07.029, URL https://www.sciencedirect.com/science/article/pii/S0378778818330858.

[70] Y. Chen, L.K. Norford, H.W. Samuelson, A. Malkawi, Optimal control of HVAC and window systems for natural ventilation through reinforcement learning, Energy Build. (ISSN: 0378-7788) 169 (2018) 195–205, http://dx.doi.org/10.1016/j.enbuild.2018.03.051, URL https://www.sciencedirect.com/science/article/pii/S0378778818302184.

[71] Y. Wang, K. Velswamy, B. Huang, A long-short term memory recurrent neural network based reinforcement learning controller for office heating ventilation and air conditioning systems, Processes (ISSN: 2227-9717) 5 (3) (2017) http://dx.doi.org/10.3390/pr5030046, URL https://www.mdpi.com/2227-9717/5/3/46.

[72] S. Baghaee, I. Ulusoy, User comfort and energy efficiency in HVAC systems by Q-learning, in: 2018 26th Signal Processing and Communications Applications Conference, SIU, 2018, pp. 1–4, http://dx.doi.org/10.1109/SIU.2018.8404287.

[73] T.M. Hansen, E.K.P. Chong, S. Suryanarayanan, A.A. Maciejewski, H.J. Siegel, A partially observable Markov decision process approach to residential home energy management, IEEE Trans. Smart Grid 9 (2) (2018) 1271–1281, http://dx.doi.org/10.1109/TSG.2016.2582701.

[74] D.B. Crawley, L.K. Lawrie, F.C. Winkelmann, W.F. Buhl, Y.J. Huang, C.O. Pedersen, R.K. Strand, R.J. Liesen, D.E. Fisher, M.J. Witte, et al., EnergyPlus: creating a new-generation building energy simulation program, Energy Build. 33 (4) (2001) 319–331.

[75] W.A. Beckman, L. Broman, A. Fiksel, S.A. Klein, E. Lindberg, M. Schuler, J. Thornton, TRNSYS the most complete solar energy system modeling and simulation software, Renew. Energy 5 (1–4) (1994) 486–488.

[76] M. Wetter, W. Zuo, T.S. Nouidui, X. Pang, Modelica buildings library, J. Build. Perform. Simul. 7 (4) (2014) 253–270, http://dx.doi.org/10.1080/19401493.2013.765506.

[77] D. Blum, J. Arroyo, S. Huang, J. Drgoňa, F. Jorissen, H.T. Walnum, Y. Chen, K. Benne, D. Vrabie, M. Wetter, et al., Building optimization testing framework (BOPTEST) for simulation-based benchmarking of control strategies in buildings, J. Build. Perform. Simul. 14 (5) (2021) 586–610.

[78] IESVE, IES virtual enviroment (IESVE), 2023, [Online] Available at:, https://www.iesve.com/products.

[79] J.R. Vázquez-Canteli, J. Kämpf, G. Henze, Z. Nagy, CityLearn v1.0: An openai gym environment for demand response with deep reinforcement learning, in: Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, BuildSys '19, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450370059, 2019, pp. 356–357, http://dx.doi.org/10.1145/3360322.3360998.

[80] D. Holmberg, M. Burns, S. Bushby, A. Gopstein, NIST Transactive Energy Modeling and Simulation Challenge Phase II Final Report, Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD, 2019, http://dx.doi.org/10.6028/NIST.SP.1900-603.

[81] K. Al Sayed, A. Boodi, R. Sadeghian Broujeny, K. Beddiar, Reinforcement learning for optimal HVAC control: From theory to real-world applications, in: IECON 2023- 49th Annual Conference of the IEEE Industrial Electronics Society, 2023, pp. 1–6, http://dx.doi.org/10.1109/IECON51785.2023.10312131.

[82] C. Szepesvári, Algorithms for reinforcement learning, in: Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 4, (1) Morgan & Claypool Publishers, 2010, pp. 1–103.
[83] S. Whiteson, et al., Adaptive Representations for Reinforcement Learning, vol. 291, Springer, 2010.
[84] Z. Rahimpour, G. Verbič, A.C. Chapman, Energy management of buildings with phase change materials based on dynamic programming, in: 2019 IEEE Milan PowerTech, IEEE, 2019, pp. 1–6.
[85] D. Silver, Lectures on reinforcement learning, 2015, URL https://www.davidsilver.uk/teaching/.
[86] M.G. Bellemare, W. Dabney, R. Munos, A distributional perspective on reinforcement learning, in: International Conference on Machine Learning, PMLR, 2017, pp. 449–458.
[87] A.L. Strehl, M.L. Littman, An analysis of model-based interval estimation for Markov decision processes, J. Comput. System Sci. (ISSN: 0022-0000) 74 (8) (2008) 1309–1331, http://dx.doi.org/10.1016/j.jcss.2007.08.009, Learning Theory 2005. URL https://www.sciencedirect.com/science/article/pii/S0022000008000767.
[88] A.-M. Farahmand, A. Barreto, D. Nikovski, Value-Aware Loss Function for Model-based Reinforcement Learning, in: A. Singh, J. Zhu (Eds.), Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, vol. 54, PMLR, 2017, pp. 1486–1494, URL https://proceedings.mlr.press/v54/farahmand17a.html.
[89] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, Playing atari with deep reinforcement learning, 2013, arXiv preprint arXiv:1312.5602.
[90] R. Rubinstein, The cross-entropy method for combinatorial and continuous optimization, Methodol. Comput. Appl. Probab. (ISSN: 1573-7713) 1 (2) (1999) 127–190, http://dx.doi.org/10.1023/A:1010091220143.
[91] T. Salimans, J. Ho, X. Chen, S. Sidor, I. Sutskever, Evolution strategies as a scalable alternative to reinforcement learning, 2017, arXiv:1703.03864.
[92] M. Janner, J. Fu, M. Zhang, S. Levine, When to trust your model: Model-based policy optimization, 2021, arXiv:1906.08253.
[93] M. Deisenroth, C. Rasmussen, PILCO: A model-based and data-efficient approach to policy search, 2011, pp. 465–472.
[94] S. Levine, V. Koltun, Guided policy search, in: S. Dasgupta, D. McAllester (Eds.), Proceedings of the 30th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 28, (3) PMLR, Atlanta, Georgia, USA, 2013, pp. 1–9, URL https://proceedings.mlr.press/v28/levine13.html.
[95] V. Mnih, A.P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, K. Kavukcuoglu, Asynchronous methods for deep reinforcement learning, in: International Conference on Machine Learning, PMLR, 2016, pp. 1928–1937.
[96] T.P. Lillicrap, J.J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning, 2015, arXiv preprint arXiv:1509.02971.
[97] S. Fujimoto, H. Hoof, D. Meger, Addressing function approximation error in actor-critic methods, in: International Conference on Machine Learning, PMLR, 2018, pp. 1587–1596.
[98] T. Haarnoja, A. Zhou, P. Abbeel, S. Levine, Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, in: International Conference on Machine Learning, PMLR, 2018, pp. 1861–1870.
[99] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, 2017, arXiv preprint arXiv:1707.06347.
[100] T. Weber, S. Racaniere, D.P. Reichert, L. Buesing, A. Guez, D.J. Rezende, A.P. Badia, O. Vinyals, N. Heess, Y. Li, et al., Imagination-augmented agents for deep reinforcement learning, 2017, arXiv preprint arXiv:1707.06203.
[101] J. Schulman, S. Levine, P. Abbeel, M. Jordan, P. Moritz, Trust region policy optimization, in: International Conference on Machine Learning, PMLR, 2015, pp. 1889–1897.
[102] T. Kurutach, I. Clavera, Y. Duan, A. Tamar, P. Abbeel, Model-ensemble trust-region policy optimization, 2018, arXiv:1802.10592.
[103] V. Feinberg, A. Wan, I. Stoica, M.I. Jordan, J.E. Gonzalez, S. Levine, Model-based value estimation for efficient model-free reinforcement learning, 2018, arXiv:1803.00101.
[104] P. D'Oro, A.M. Metelli, A. Tirinzoni, M. Papini, M. Restelli, Gradient-aware model-based policy search, Proceedings of the AAAI Conference on Artificial Intelligence, (ISSN: 2159-5399) vol. 34 (04) (2020) 3801–3808, http://dx.doi.org/10.1609/aaai.v34i04.5791.
[105] Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, N. de Freitas, Sample efficient actor-critic with experience replay, 2017, arXiv:1611.01224.
[106] J. Beck, R. Vuorio, E.Z. Liu, Z. Xiong, L. Zintgraf, C. Finn, S. Whiteson, A survey of meta-reinforcement learning, 2023, arXiv preprint arXiv:2301.08028.
[107] N. Schweighofer, K. Doya, Meta-learning in reinforcement learning., Neural Netw. Official J. Int. Neural Netw. Soc. (ISSN: 0893-6080) 16 (1) (2003) 5–9, http://dx.doi.org/10.1016/s0893-6080(02)00228-9, arXiv:12576101.
[108] K. Rakelly, A. Zhou, D. Quillen, C. Finn, S. Levine, Efficient off-policy meta-reinforcement learning via probabilistic context variables, 2019, arXiv:1903.08254.