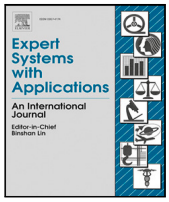
Expert Systems With Applications 238 (2024) 121749

Contents lists available at ScienceDirect

Expert Systems With Applications

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



An online reinforcement learning approach for HVAC control

Francesco M. Solinas [a], Alberto Macii [a], Edoardo Patti [a], Lorenzo Bottaccioli [b,*]

[a] Department of Control and Computer Engineering, Politecnico di Torino, Turin, Italy
[b] Interuniversity Department of Regional and Urban Studies and Planning, Politecnico di Torino, Turin, Italy

ARTICLE INFO

Keywords:
Reinforcement learning
Deep deterministic policy gradient
System identification
Imitation learning

ABSTRACT

Heating, Ventilation and Air Conditioning (HVAC) optimization for energy consumption reduction is becoming ever more a topic of the utmost environmental and energetic concerns. The two most employed methodologies for optimizing HVAC systems are Model Predictive Control (MPC) and Reinforcement Learning (RL). This paper compares three different RL approaches to HVAC optimization: one based on a black-box system identification model trained on historical data, one based on a white-box model of a building and one online method based on an imitation learning pretraining phase on historical data. The three approaches are compared with a literature baseline and an EnergyPlus baseline. Results show that the overall best method in terms of energy consumption reduction (65% decrease) and thermal comfort increase (25% increase) is the approach based on the white-box model. However, the proposed methodology, based on online and imitation learning, demonstrates remarkable efficiency, achieving comparable improvements in energy consumption after just a few months of online training, while maintaining thermal comfort at around the same level as the baseline. These results prove a direct online RL approach, which avoid the use of costly simulations, can provide a reliable and inexpensive solution to the problem of HVAC optimization.

1. Introduction

More than half of the world's population now lives in metropolitan regions. According to United Nations estimates, urban areas would host roughly 68% million people by 2030, with one-third of them living in municipalities with populations of at least half a million (United Nations, 0000b). According to the United Nations Habitat Division (United Nations, 0000a), urbanization is largely energy intensive. Cities consume and create approximately 75% of world's primary energy supply and 50%–60% of world's greenhouse gas emissions. Thanks to the advancement in Information Communication Technologies and with the spread deployment of Internet of Things, it is possible to monitor buildings and to develop new fine-grained algorithms for energy consumption optimization (Wigle, 2014). Heating, Ventilation and Air Conditioning (HVAC) systems are among the highest energy consumption appliances in today's urban scenarios, making them one of the most important objects of research.

On top of traditional methods, based on the control of Proportional Integral Derivative (PID) parameters (Soyguder, Karakose, & Alli, 2009), the two most employed methodologies for HVAC optimization are Model Predictive Control (MPC) (Drgoňa et al., 2020) and Reinforcement Learning (RL) (Wang & Hong, 2020). The two methods differ mainly in the fact that MPC requires a detailed model of the target environment in order to perform the optimization and choose the optimal action. On the contrary, a vast class of RL algorithms are model-free, meaning that they can be applied to any environment, even with scarce or no information regarding its functioning. RL (Vázquez-Canteli & Nagy, 2019) and in general Machine Learning methods have been successfully applied to HVAC optimization and in general building or smart grid control (Tiwari et al., 2022).

This paper thus aims at tackling the problem of reducing energy consumption in HVAC systems. This problem consists in the task of optimizing the overall energy expenses or consumption of the building HVAC systems, maintaining unvaried the internal comfort of its users.

Our goal with the present work is that of providing a practical solution to HVAC control, which can be readily applied to target buildings and use-case scenarios. The main limitation of model-free RL algorithms relies on the training phase because the agent needs to try several actions before converging on the optimal policy. Hence, in the training phase, the agent will apply several actions that may cause an improper state of the environment, that in a building HVAC system might cause an increase in energy consumption or discomfort for the user. To solve this issue, literature solutions propose the usage of a simulated environment to train the model-free RL agent

* Corresponding author.
E-mail addresses: francesco.solinas@polito.it (F.M. Solinas), alberto.macii@polito.it (A. Macii), edoardo.patti@polito.it (E. Patti), lorenzo.bottaccioli@polito.it (L. Bottaccioli).

https://doi.org/10.1016/j.eswa.2023.121749
Received 7 December 2022; Received in revised form 5 August 2023; Accepted 19 September 2023
Available online 26 September 2023

and then apply the agent to the real building. Such solutions rely on costly white-box modelling of the target building or need other auxiliary measures, such as wall temperature monitoring, to develop data-driven system identification models. Both approaches can disrupt the time and cost-effectiveness of the solutions. Our novel methodology, however, is readily applicable to any target use-case scenario with few or none preparatory work. As explained in more detail in this paper, our proposed approach relies on a fully continuous RL agent. The agent undergoes an Imitation Learning (IL) phase, which consists of a training methodology in which the agent is exposed to building historical data, comprehensive of the actions performed by a standard controller and their consequences on the environment, and actively learns without having to directly perform any action itself. After this pre-training phase, the agent is ideally deployed on the target building HVAC system. For simplicity, a simulated model in EnergyPlus (Zhang & Lam, 2019), one of the most renowned building energy simulation programs (Yang & Becerik-Gerber, 2014), is adopted as our real-case building, meaning that all constraints that would apply to a real building have also been considered in our simulated scenario. Results show that our approach is sound and effective, surpassing the standard rule-based control systems already in the very first weeks of adoption.

The paper is structured as follows: Section 2 presents a review of the most recent literature about reinforcement learning approaches regarding optimal control of HVAC systems, also detailing the novelties that our approach brings to the current state of the art. Section 3 describes the three approaches investigated in this paper, namely, the system identification model, the white-box modelling environment and the fully online system. This section also fully describes the reinforcement learning algorithm of choice, which is the Deep Deterministic Policy Gradient (DDPG). Section 5 presents results for each different methodology. At last, in Section 6, an overview of the proposed work is given and final remarks are discussed.

2. State of the art

In what follows, a literature review of the state of the art on HVAC optimization through Reinforcement Learning techniques is presented. As shown, several attempts to adopt different methodologies have been utilized. Most solutions employ model-free algorithms that work on continuous action space for single-zone problems, consisting in optimization tasks in which an agent has to discover the optimal HVAC strategy (i.e. temperature setpoint, air mass flow rate) to maximize one or more target objectives (i.e. thermal comfort, energy cost) in an environment built of one whole thermal zone. The most preferred method for the optimization task is the DDPG algorithm. When the action space is discrete, the most common option is the Deep Q-Network (DQN) algorithm or some kind of variation of it. All works target either the overall energy consumption, the overall energy cost, the thermal comfort or a combination of these.

A work presented in Kou et al. (2021) compares a data-driven approach, that is one in which the optimal strategy arises as a consequence of gathered data, based on a model-free DDPG algorithm, with a rule-based approach and a Model Predictive Control (MPC) methodology. The latter requires the adoption of a detailed model of the building and accurate indoor air temperature predictions, while the data-driven approach requires a smaller amount of inputs. The control actions space consisted of the simple on/off switch of the HVAC system. The model-free algorithm showed good performances in terms of computational speed and cost reduction, however, the MPC method performed the best, despite being more time and resource-consuming. In Du, Zandi et al. (2021), a 2-zone building optimization is performed, comparing a DDPG algorithm with a discrete DQN model and a rule-based approach. As in Du, Li et al. (2021), the agent controls the indoor temperature setpoint. The DDPG method showed the best performance in terms of energy cost reduction and thermal comfort violation. A DDPG algorithm is also used in Rahimpour, Verbič, and Chapman (2020) for heating optimization, and compared with a model-based RL approach consisting of an Approximate Dynamic Programming (ADP) model. The model uses as a control action the on/off switching of the HVAC system. The DDPG compared similarly to the ADP, while not requiring a model of the environment. In Gao, Li, and Wen (2020), a DDPG algorithm is compared with a DQN and simpler RL approaches, for a control task concerning the definition of the HVAC indoor air temperature and relative humidity setpoints. The proposed DDPG method showed an increase in thermal comfort of up to 10% and a decrease in energy cost.

Authors of Ding, Du, and Cerpa (2019) propose a holistic approach, in which an innovative Deep Reinforcement Learning algorithm, the Branching Duelling Q-Network, is employed to optimize over the energy consumption and the user thermal comfort. It is then compared with a traditional Dual Deep Q-Network (DDQN) and a rule-based approach. The action space consists in controlling the HVAC traditional systems and several other building subsystems, such as lights, blinds and window systems. The proposed method obtains a reduction up to 12% in energy consumption. In Du, Li et al. (2021), a Deep Deterministic Policy Gradient (DDPG) algorithm is trained on both cooling and heating seasons and compared with a standard rule-based approach. The DDPG algorithm controls the indoor air temperature setpoint on a continuous scale, and shows a reduction in overall electricity cost and consumption with only modest comfort temperature violations (around 0.6 °C). In Brandi, Piscitelli, Martellacci, and Capozzoli (2020), a DQN algorithm is used to optimize the control over the supply water temperature of a building. The action space is discrete and consists of a set of water temperature increments. The proposed method achieves a significant reduction in energy consumption, setpoint violation and overall thermal comfort, compared with a traditional rule-based approach. In Barrett and Linder (2015), an intelligent thermostat based on a DQN algorithm is developed. The action space is discrete and consists of the on/off switching of both cooling and heating HVAC systems. Slight cost reduction and improved thermal comfort are obtained, compared to a traditional scheduled thermostat. In Zhang, Kuppannagari, Kannan, and Prasanna (2019), a black box approach based on a neural network is employed, in which the network learns the system identification by continuous interaction with the target building modelled in EnergyPlus. Then, an MPC approach is compared with a model-free Proximal Policy Optimization (PPO) algorithm, showing that the model-based approach can achieve up to 10x improvements in energy consumption and thermal comfort with respect to the RL algorithm.

In Wei, Wang, and Zhu (2017), a DQN algorithm is employed for discrete action-space, consisting of several multiple choices of air flow rate control, over a multi-zone HVAC system. The exponentially large action space is dealt with the help of custom heuristics for computational reduction purposes. Reduction in energy cost and maximization of thermal comfort are the goals of the algorithm, which is tested on an EnergyPlus environment. In Fu and Zhang (2021) an innovative Twin Delayed Deep Deterministic Policy Gradient algorithm and Model Predictive Control (TD3-MPC) is introduced and compared with a standard DDPG algorithm for continuous control over indoor air temperature setpoint, aimed at energy cost reduction and thermal comfort maximization. The novel TD3-MPC scores around 15% better than its DDPG counterpart. In Solinas, Bellagarda, Macii, Patti, and Bottaccioli (2021) a hybrid approach to demand side management optimization is presented, in which a system identification model is adopted to realistically represent a target building thermodynamic response, based on only historical data gathered from a standard controller. Then, three methods are compared: a rule-based approach, an MPC and a DQN algorithm, with the objective of reducing energy consumption, and respecting thermal comfort. The action space consists in a discrete set of supply air temperature increments. This approach showed clear advantages in terms of overall results and real-world applications, as

historical data are the only requirements for training the proposed agent.

Table 1 reports the above-mentioned, most relevant works in the literature on this topic. The table details the target of the optimization, the control actions, the kind of algorithm or methodology employed, the kind and size of the dataset over which the system is trained, the kind of environment used for training and testing, if the RL agent utilized is based on a model-free or model-based approach and finally if the target use-case scenario consists of a single or a multi-zone building (and the corresponding number of different thermal zones). Most of the studies focus on optimizing energy consumption or cost, while maintaining thermal comfort, revealing a strong trend towards sustainability and user comfort. The control actions primarily involve manipulating HVAC settings, emphasizing their key role in energy management. The table shows a strong preference for model-free RL methods, with DDPG and DQN being commonly used. However, a few studies employ both model-free and model-based approaches, indicating a potential benefit of combining both techniques. The duration and focus of the studies vary, reflecting the diversity of applications and climates. Most studies apply to single-zone scenarios, but a few extend to multi-zone, showing an increasing complexity in the research field.

This paper expands on the work of Solinas et al. (2021) by implementing a more complex use-case scenario and introducing three distinct approaches to system modelling. In the cited work, the authors focused on using a system identification model to represent system dynamics. Our first approach also employs a system identification model for simulating the building's thermal response, providing a useful comparison on a common baseline. Our second approach uses detailed white-box modelling to simulate the target building, an approach often adopted in the discussed literature. While this method can yield more accurate thermal responses, it often incurs high costs due to the need for comprehensive modelling. The third and most innovative approach is the application of a fully online agent to the target building without any preceding model-based simulation. We use a building model simulated in EnergyPlus as a proxy for a real building, subject to the same constraints such as thermal comfort and energy costs. The agent undergoes an Imitation Learning pre-training phase, a novel method in the field of HVAC system optimization that we believe significantly improves the practicality and efficiency of the solution. To the best of our knowledge, this practical-centred approach, coupled with an imitation learning pre-training phase, has never been proposed before in the literature and represents a strong novelty in the field of optimization for HVAC systems.

3. Methodology

The present work targets the optimization of the heating consumption of an HVAC system. The heating flow $Q_{heating}$ of an HVAC system, and its related energy consumption, can be expressed with the following equation:

$$Q_{heating} = m_{dot} \cdot c_p \cdot (T_{air} - T_{in}) \tag{1}$$

where m_{dot} is the air mass flow rate, c_p is the specific heat capacity, T_{air} is the air temperature from the HVAC system and T_{in} is the indoor air temperature. Hence, w.r.t to the reviewed literature, the action space is extended to include air mass flow rate control, in addition to air temperature control. Three main methodologies are then explored: the first one, similar to Solinas et al. (2021), is based on a black-box system identification model which simulates the building thermodynamics. In the second one, a white-box model based on the EnergyPlus building energy simulation program is adopted to provide a more realistic building response. The third methodology presents a novel online learning approach, adopted to overcome the shortcomings of both previous methods. The rest of this section will describe these approaches. Due to the lack of real buildings in which we could test the three different approaches we have chosen to use the software EnergyPlus (Crawley, Lawrie, Pedersen, & Winkelmann, 2000). Such software is a building simulation program commonly used by researchers and designers and is applied in this work to compare and test the three different methods and to generate a historical dataset of an expert system for the control of HVACs.

The task of finding the best action pair of indoor air temperature increase and air mass flow rate can be described as a Markov Decision Process (MDP) and thus tackled through reinforcement learning algorithms. The optimal control problem is then modelled similarly to our previous work (Solinas et al., 2021), in which an agent has to perform a control action u in a state s, to receive a reward r, transitioning to a new state s'. The goal of the agent is that of finding the best control action in every state, maximizing the expected total reward.

The reward r is defined as the distance of the indoor air temperature x_t of the building to the predetermined setpoint $x_{setpoint}$, plus the cost c_t of performing a particular control action pair $u_{temp,t}, u_{mdot,t}$, respectively the control actions over the temperature and the mass flow rate of the air that is provided by the HVAC system to the building.

$$r = -(\beta \cdot (x_t - x_{setpoint})^2 + \rho \cdot c_t) \tag{2}$$

where

$$c_t = u_{temp,t} \cdot u_{mdot,t} \tag{3}$$

and where β and ρ are variable to weigh the two sides of the equation, the respect of the temperature setpoint on one side, and the energy consumption reduction on the other one. The choice of the cost function is based on the direct proportionality of each of these control actions to the system's energy consumption. Specifically, an increase in the supply air temperature or the air mass flow rate leads to a corresponding increase in the energy consumption by the HVAC system coil and the ventilation system, respectively. This allows the agent to optimize the reward quantity by minimizing the use of the control actions as much as possible, thereby reducing the HVAC system's energy consumption. As in Solinas et al. (2021) state s includes: the difference $x_{setpoint,t} - x_t$ (°C) between the setpoint and the indoor air temperature, with a 4 periods lag; the difference $x_{setpoint,t} - x_{outdoor,t}$ (°C) between the setpoint and the outdoor air temperature; the outdoor relative humidity $RH_{outdoor}$ (%); the wind speed $Wind_{speed}$ (m/s); the wind direction $Wind_{dir}$ $(degrees)$; the diffracted solar radiation Rad_{diff} (W/m^2); the direct solar radiation Rad_{direct} (W/m^2); the number of hours before the start or the end of the next or the ongoing occupancy period, Occ_{start} and Occ_{end} (h) respectively. All inputs are pre-processed according to the $MinMax$ normalization rule. The set of control actions u_t, which the DDPG agent can perform, consists in the supply indoor air temperature and the corresponding air mass flow rate.

Differently to what had been presented in Solinas et al. (2021), however, the action space is now more complex as it includes not only the air temperature control but also the air mass flow rate. Because of this increased complexity, it has been deemed necessary to implement an algorithm for continuous action spaces, namely the Deep Deterministic Policy Gradient (Lillicrap et al., 2015). The DDPG algorithm can be considered a DDQN (Mnih et al., 2015) for continuous action space. As shown in Algorithm 1, the DDPG is composed of two main networks, a critic network W_ω, which is just a Q network as in the DDQN model, and an actor-network μ_θ, which is responsible for sampling actions around a certain mean μ. Both networks have their respective target networks for training stabilization purposes. The parameter γ plays in the DDPG model a similar role than in a DDQN, as already noted in Solinas et al. (2021). γ is employed to mediate between the two terms in the target Q^* value calculation (line 11 in Algorithm 1), namely, the immediate reward r_t and the expected long term reward $Q_\omega(s_{t+1}, argmax_u Q_{\omega'}(s_{t+1}, \mu_{\theta'}(s_{t+1})))$. Higher values for γ, close to 1, favour the maximization of long-term reward, while lower values of γ, around 0.8, favour the maximization of immediate reward.

In the following three proposed implementations, the only changes to the illustrated algorithm regards the state-transition function f_ϕ and the number of episodes for which the agent is allowed to train.

Table 1
RL optimization studies.

Study	Target	Control actions	Methodologies	Data and season period	Model free/based	Single/Multi zone (n. of zones)
Kou et al. (2021)	Energy cost, thermal comfort, and utility-level load violation	Switching the HVAC on/off (continuous)	DDPG, Alternating Direction Method of Multipliers (ADMM), Rule-based	30 Summer Days, Cooling	Model-Free	Single-Zone
Du, Zandi et al. (2021)	Energy cost, thermal comfort	Temperature setpoint (continuous)	DDPG, DQN, Rule-based	1 month, Heating	Model-Free	Multi-Zone (2)
Du, Li et al. (2021)	Energy cost, thermal comfort	Temperature setpoint (continuous)	DDPG, Rule-based	1 month, Heating and Cooling	Model-Free	Multi-Zone (2)
Rahimpour et al. (2020)	Energy cost, thermal comfort	Switching the HVAC on/off (continuous)	DDPG, Approximate Dynamic Programming (ADP), Rule-based	294 days, Heating	Model-Free, Model-Based	Single-Zone
Gao et al. (2020)	Energy cost, thermal comfort	HVAC temperature and humidity setpoints (continuous)	DDPG, Q-Learning, SARSA, DQN	10,000 h, Heating and Cooling	Model-Free	Single-Zone
Ding et al. (2019)	Energy consumption, thermal comfort, visual comfort and indoor air quality	HVAC setpoint, lighting, blind slat angle and open % of window (discrete)	Branching Duelling Q-Network (BDQ), DDQN, Rule-based	10 years, Heating and Cooling	Model-Free	Single-Zone
Brandi et al. (2020)	Energy consumption, thermal comfort	Supply water temperature (discrete)	DQN, Rule-based	3 months, Heating	Model-Free	Single-Zone
Barrett and Linder (2015)	Energy cost, thermal comfort	Switching Heating and Cooling on/off (discrete)	Q-Learning, Rule-based	150 days, Heating	Model-Free	Single-Zone
Zhang et al. (2019)	Energy consumption, thermal comfort	HVAC setpoint temperature, air mass flow rate (continuous)	Model-Based RL, PPO, Rule-based	65 days, Heating	Model-Free, Model-Based	Single-Zone
Wei et al. (2017)	Energy cost, thermal comfort	Air flow rate (discrete)	Deep Q Learning, Q Learning, Rule-based	100 months, Heating and Cooling	Model-Free	Single-Zone, Multi-Zone (up to 5)
Fu and Zhang (2021)	Energy cost, thermal comfort	Indoor air temperature setpoint (continuous)	TD3-MPC, DDPG	100 days, Heating	Model-based, Model-Free	Multi-Zone (5)
Chen, Cai, and Bergés (2019a)	Energy consumption, thermal comfort	Supply indoor air temperature (discrete)	PPO-MPC	100 days, Heating	Model-based, Model-Free	Single-Zone
Solinas et al. (2021)	Energy consumption, thermal comfort	Supply indoor air temperature (discrete)	DDQN, MPC, Rule-based	3 months, Heating	Model-Free	Single-Zone
Proposed solution	Energy consumption, thermal comfort	Supply air temperature, air flow rate (continuous)	DDPG, DDQN, Rule-based	3 months, Heating	Model-Free	Single-Zone

The state-transition function f_ϕ represents the system dynamics and maps a set of inputs y_t into the next state s_{t+1}:

$$f_\phi(y_t) \rightarrow s_{t+1} \tag{4}$$

f_ϕ is the core of the proposed environment as it is necessary in order to provide accurate feedback to the learning agent regarding the effects of its action on the world.

It will be shown how different configurations provide different advantages and shortcomings for HVAC systems optimization. As shown in Fig. 1, three configurations are presented differing from one another by the way they treat the f_ϕ state-transition function. In the first approach, a black-box model based on a neural network is used to learn f_ϕ from historical data and predict new states accordingly. The black-box model is then used as the environment to receive agent actions and output corresponding rewards and next states. In a second approach, a simulated building modelled in EnergyPlus is used to approximate f_ϕ and provide virtual data to the agent. The training process thus happens directly on the simulated building. For both the black-box and the white-box models, the testing has been performed on a simulated building taking the role of the actual real building. In the last configuration, a simulated building is treated in all concerns *as if* it was a real-world building, providing ground-truth values for the f_ϕ state-transition function. In this last approach, first an Imitation Learning phase based on historical data is adopted, acting as expert data for this pre-training phase. Then the agent is directly deployed in an online fashion on the real building.

3.1. Black-box model

The most evident advantage of performing a system identification based on historical data from the target building is that of creating a fast and reliable simulated model which can provide a good approximation of the state-transition function f_ϕ.

Compared to Solinas et al. (2021), the environment under study is more complex as the air mass flow rate control action has been introduced. For this reason, a more powerful system identification method has been deemed necessary for better capturing the system dynamics.

More specifically, a neural network model has been implemented for approximating the non-linear state-transition function, mapping the initial internal temperature of the building x_t, the set of control actions u_t, namely the distributed air temperature and the air mass flow rate, and a set of disturbances d_t, consisting in the outdoor temperature and the occupancy status of the building, into the following internal temperature $\hat{x}_{t+1}$

$$f_\phi(x_t, u_t, d_t) \rightarrow \hat{x}_{t+1} \tag{5}$$

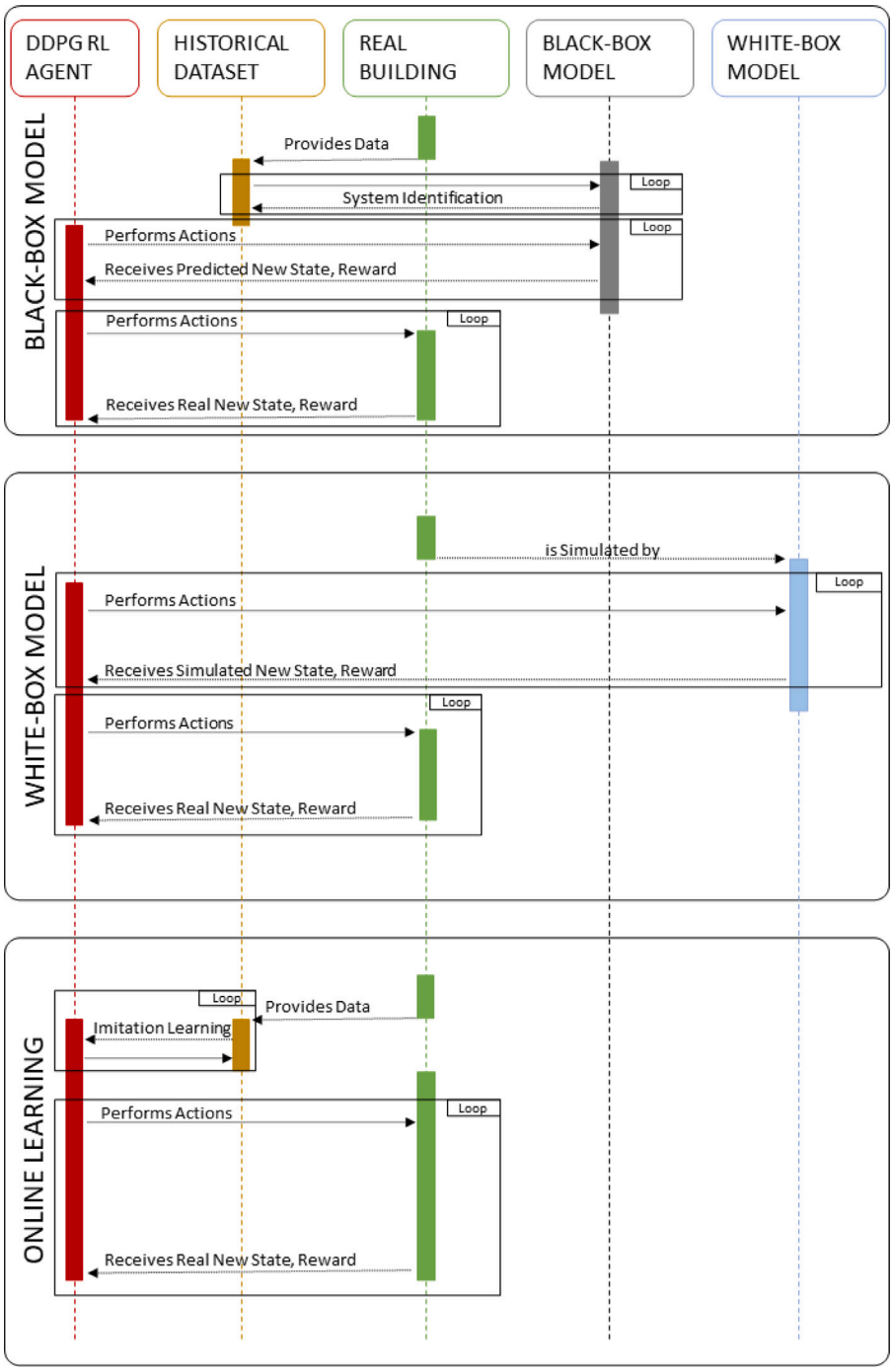


Fig. 1. Sequence diagram of the three proposed approaches.

The model consists of three stacked fully connected layers with Rectified Linear Unit (ReLU) activation function, modelled in PyTorch.

3.2. White-box model

Compared to a black-box modelling tool such as the system identification described in the previous section, white-box modelling has the advantage of being based on an analytical thermodynamics simulation of the system. The white-box model, in fact, simulates the real building in its most relevant thermal characteristics and behaviours. The main advantage is that such kind of modelling is more accurate in approximating the real-world state-transition function f_ϕ. However, the development of such a model is a time-consuming and costly process, which involves a thorough investigation of the target premises and a careful analysis of its structure, which has to be accurately replicated in the chosen simulator.

When a white-box model is available, it is however the best option to approximate the building's thermal response, without actually having to connect the learning agent to the target structure. For the purpose of this project, the model has been implemented in EnergyPlus (Crawley et al., 2000), following the structure of Gym-like environments (Brockman et al., 2016).

3.3. Online learning

Due to the difficulties regarding the system identification model for the presented environment with a more complex action space, and the

Algorithm 1 The DDPG algorithm for HVAC optimization.

Random initialize parameters ω, θ of critic network Q_ω and actor-network μ_θ, parameters $\omega' \leftarrow \omega$ of target critic network $Q_{\omega'}$ and parameters $\theta' \leftarrow \theta$ of target actor network $\mu_{\theta'}$
Initialize replay memory D, learning rate α and target network update parameter τ
while $Episode < EP_{max}$ **do**
  while $Steps < Steps_{max}$ **do**
    Observe state s_t and sample action $u_{t+1} = \mu_\theta(s_t) + \epsilon$ where $\epsilon \sim OU$
    Perform action u_{t+1} in the environment, get the reward r_{t+1}, calculate the new temperature $x_{t+1} = f_\phi(s_t, u_t)$ and observe next state $s_{t+1} = [x_{t+1}, d_{t+1}]$
    Store (s_t, u_t, r_t, s_{t+1}) in the replay memory D
  end while
  while $Update\_Steps < Update\_Steps_{max}$ **do**
    Sample a batch (s_t, u_t, r_t, s_{t+1}) of size Bs from memory D
    Compute target Q value: $Q^*(s_t, u_t) = r_t + \gamma \cdot Q_\omega(s_{t+1}, argmax_u Q_{\omega'}(s_{t+1}, \mu_{\theta'}(s_{t+1})))$
    Update Q_ω performing gradient descent step on $(Q^*(s_t, u_t) - Q_\omega(s_t, u_t))^2$
    Update μ_θ performing gradient ascent step on $Q_\omega(s_t, \mu_\theta(s))$
    Update $Q_{\omega'}$ parameters: $\omega' \leftarrow \tau \cdot \omega + (1 - \tau) * \omega'$
    Update μ_θ parameters: $\theta' \leftarrow \tau \cdot \theta + (1 - \tau) * \theta'$
  end while
end while

impracticability of simulating a white-box model for every new test building, a different approach to HVAC optimization is presented here.

This approach is based on an online learning technique, in which the RL agent responsible for choosing the correct action pair of temperature and air mass flow rate is directly connected to the target building, can perform an action on the actual system and receive real-time feedback from it. By doing so, it is possible to gather the real thermodynamic response from the environment, overcoming the difficulties of a system identification model and avoiding the need of costly and time-consuming white-box modelling. On the other hand, this approach brings several shortcomings to be appropriately tackled. The agent, indeed, needs to quickly learn an effective strategy to at least match the performance of the existing baseline controller. If this does not happen quickly enough, the training cost regarding energy consumption and thermal comfort might be too high in the very first weeks or months for practical implementation. This would be an expected outcome as in the very first stages of learning RL agents are basically clueless regarding what the optimal action is and tend to explore a lot of possibly inefficient strategies.

In a nutshell, compared to the two other kinds of methodologies, during online learning, the agent has to reach a satisfactory level of performance in a very short time frame compared to the offline training methods described previously.

To overcome such difficulties, the online approach is integrated with a pre-training phase, where the agent is exposed to historical data regarding the existing controller of the building and the system dynamics. This is based on an Imitation Learning methodology (Hussein, Gaber, Elyan, & Jayne, 2017) which has been proven useful in robotics (Schaal, 1999), 3D navigation tasks (Hussein, Elyan, Gaber, & Jayne, 2018) and many other applications. At first, an expert system (a standard rule-based HVAC controller, in our case) is recorded in order to generate a historical dataset representing a first-hand experience of the real-world dynamics for the learning agent. More specifically, the main feature of Imitation Learning is that during this training phase, the agent cannot choose its own actions, but these are determined by what the existing controller did during the recorded period. The agent is thus fed with expert actions and their real consequences on the environment as recorded in the historical dataset, that is, the action set u_t and consequent state s'_t are predetermined at each timestep by the historical real-world process. Once this pre-training phase is completed, the agent acts and learns in the simulated environment just as before.

4. Experimental setup

To rigorously evaluate the efficiency of HVAC control models, a comprehensive testing methodology is employed. In this section, we present the simulation environment, training data, evaluation metrics, and specific setup for each of the three models under examination: Black-box model, White-box model, and the Online Learning model.

4.1. Simulation environment

The experimental setup hinges on the EnergyPlus (E+) simulator, a top-tier building energy modelling (BEM) software. For all testing, an EnergyPlus model, retrievable here (Chen, Cai, & Bergés, 2019b) and employed in Chen et al. (2019a), is used. This model is composed of a five thermal zones building, which is treated as a unique thermal zone for the purpose of our work, that is, at each timestep t the average of the indoor air temperature of the five zones is computed and employed as x_t, one of the primary component of state s_t. The proposed methods will be tested against an EnergyPlus rule-based approach, a standard controller which performs predefined actions based on a fixed schedule, depending on the time of the day, the day of the week and the season of the year. The EnergyPlus (Crawley et al., 2000) baseline was chosen due to its comprehensive and accurate simulation capabilities, being a state-of-the-art tool, developed by the U.S. Department of Energy for simulating buildings' thermal dynamics. It provides a valuable benchmark for comparing and evaluating the performance of our proposed HVAC control approach.

In all cases, two possible control actions are available for the proposed DDPG agent: continuous supply air temperature increases, given the constraints of a minimum value of 0 ° C and a maximum of 6 ° C, and continuous air mass flow rate provision, from a minimum of 0 m^3/s to a maximum of 3.5 m^3/s. When, for comparison purposes, the tests involve the DDQN algorithm, the control actions set is constructed as a combination of seven possible temperature increments [0,1,2,3,4,5,6] and four air mass flow rate actions [0,1,2,3.5], making a total action space of 28 different action pairs.

4.2. Key performance indicators

To compare the results achieved by the three different approaches between them and w.r.t to the EnergyPlus baseline, we adopted the following KPIs that are (i) the Percentage People Dissatisfied (PPD), (ii) Coil Power and (iii) HVAC Power. Such KPIs are widely adopted in the literature presented in Section 2 to compare and validate the algorithms w.r.t. baseline e.g. Chen et al. (2019a), Ding et al. (2019). PPD is a metric that indicates the average percentage of people that would be dissatisfied in certain thermal conditions. Anything below 20% is considered to be acceptable by the American Society of Heating, Refrigerating and Air Conditioning Engineers (ASHRAE), which sets the standards on thermal comfort (American Society of Heating, Refrigerating and Air Conditioning Engineers (Atlanta, Georgia), 2017). Coil Power expresses the energy consumed by the heating coil in terms of kWh, while HVAC Power indicates the overall energy consumption of the whole system. In a nutshell, a strategy that tends to supply warmer air temperature at a lower mass flow rate will proportionally increase the Coil Power consumption, and vice versa a strategy that mainly relies on the supply air mass flow rate is going to keep coil consumption at lower levels.

4.3. Black-box model setup

The Black-box model is a data-driven approach that learns the thermal dynamics of the building based on historical data. For training and testing, one year of hourly data from E+ simulations was utilized. This dataset includes a wide range of variables, from external and internal temperatures to solar radiation. Performance was evaluated using two

Table 2
Results comparison for black-box model and energyPlus (E+) baseline.

	E+	*Proposed DDPG*		
		$\gamma = 0.8$	$\gamma = 0.9$	$\gamma = 0.99$
PPD	**19.33%**	45.76%	41.01%	42.89%
Coil Power	6317 kWh	**1209 kWh**	1320 kWh	1660 kWh
HVAC Power	10076 kWh	**5309 kWh**	6471 kWh	6850 kWh

metrics: Percentage of People Dissatisfied (PPD) and HVAC system's energy consumption (in kWh). Lower values in both metrics indicate better performance. The training data consists in the historical dataset as created by a standard EnergyPlus scheduled rule-based controller, which acts as the baseline and the expert system here. The neural network is a fully connected three layers model, of 32 neurons each with Rectified Linear Unit (ReLU) activation function, modelled in PyTorch. Learning rate is set to 0.001 and batch size to 32.

4.4. White-box model setup

The White-box model makes use of physical laws and architectural details of the building to simulate the building's thermal behaviour. It employs the same training and testing data, which can be retrieved at Chen et al. (2019b), as the Black-box model, obtained from one year of EnergyPlus simulations. Performance evaluation is identical to the Black-box model, employing PPD and energy consumption as metrics.

The proposed algorithm (DDPG) is tested against the model (DDQN) presented in Solinas et al. (2021) which acts on discrete action spaces only. For a fair comparison, models are trained and tested on the same 3-months data used in Solinas et al. (2021). Both are tested on various levels of the γ hyper-parameter. As before, the EnergyPlus scheduled controller is adopted as the baseline. This literature baseline was chosen because it represents our previous work in the development of control strategies for a similar, albeit less complex, HVAC system. This allows us to evaluate the progress and improvements made in our current work compared to our previous methodologies.

Two scenarios are presented in which internal loads are modelled differently. Internal loads are all those endogenous factors, apart from the HVAC system itself, that contribute to the heating of the building, such as the electrical appliances and the presence of people inside the building's areas. In the first scenario, no internal loads are modelled, showing the algorithm's ability in learning how to optimize the building energy distribution without additional factors. In the second scenario, internal loads are present and strongly influence the internal thermodynamics of the system, allowing the RL agent to exploit their presence by consuming significantly less energy than in the previous case.

4.5. Online learning model setup

The Online Learning model employs an iterative process to learn and improve its performance over time.

For simplicity, the white-box EnergyPlus model acts here as the real building on which the methods are tested online. In a real world scenario, the building simulated with EnergyPlus can be easily replaced by a real world building equipped with Internet of Things devices that monitor the environmental conditions and report the needed data.

Two scenarios are proposed in which the chosen algorithm (DDPG), is tested in an online setting without previous training and with previous offline imitation learning training. In the latter case, the RL algorithm is shown one year of baseline EnergyPlus controller actions, which enables it to start the online training phase with previous, though limited, knowledge of the environment.

Compared to tested Black-box and White-box models, the whole year is considered in this experimental campaign. This has been proven useful for increasing the learning capabilities of the online RL agent. Despite this, the action space is the same as the other tested approaches, for the whole year. As only increments in the supply indoor air temperature are possible in the proposed methodology, cooling actions were not allowed for neither the rule-based approach nor the RL agent, resulting in poorer performances in terms of thermal comfort for both approaches during the summer months. For this reason, despite showing results for the whole period, we will focus the discussion of the results on the heating season.

5. Results

In this section, results are presented for the three proposed approaches, namely the offline learning based on the system identification black-box model, the offline learning based on a white-box model and an online and Imitation Learning approach.

5.1. Black-box model

In this section, results are presented for the offline learning methods based on the newly developed system identification model (see Section 3.1).

The neural network is capable of quickly learning from the provided historical dataset. Indeed, after just 30 epochs, the loss on the training- and the test-set gets as low as 0.34 °C. No hyperparameters optimization campaign has been carried out, because the model performs in a satisfactory way on the testing dataset, while lacking to transpose these results when employed as the state-transition function f_ϕ during the training of the RL agent.

In fact, despite the good performance on the historical dataset as observable in Fig. 2 for both the training and test loss, the system identification seems to not be able to accurately represent the state-transition function f_ϕ for all action pairs that the agent implements during its training phase, as shown in Table 2. The agent is completely not able to match the performance of the rule-based controller in terms of thermal comfort, meaning that the state-transition function, upon which the thermal response and thus its reward is calculated, led it strongly off the path. On the contrary, the energy consumption term in the reward function does not depend on the state-transition function, but it is simply represented as the cost of performing a certain action proportionally to its intensity, making the agent much more able to perform efficiently in that regard.

The poor performances of the system identification model are most likely due to the fact that adding a second control variable, namely the mass air flow rate, made the environment significantly more complex and thus harder to encompass with a system identification model. On top of that, the historical data upon which the model is trained exhibits a very narrow action space. This is due to the rule-based controller routine scheduling, which often chooses the same action pairs for getting the building's internal temperature to the required setpoint, and thus does not allow the system identification model to really encompass all possible action pairs and their effects on the system thermal dynamics. A larger historical dataset, consisting of a much greater variety of control action pairs and their direct consequences on the building thermal response would most likely allow significant improvements to the system identification model introduced here. This greater variety of actions is however seldom available and, in most cases, unfeasible to collect as it would require, in real-case scenarios, that the standard controller performs many potentially inefficient control action pairs.

5.2. White-box model

This section presents results where the RL algorithm is directly trained on a white-box model such as that of EnergyPlus (see Section 3.2). This is a perfectly viable approach whenever detailed modelling of the target building is present.

When no internal loads are present, the agent can fulfil its task only by relying on temperature increments in the supplied air, as no

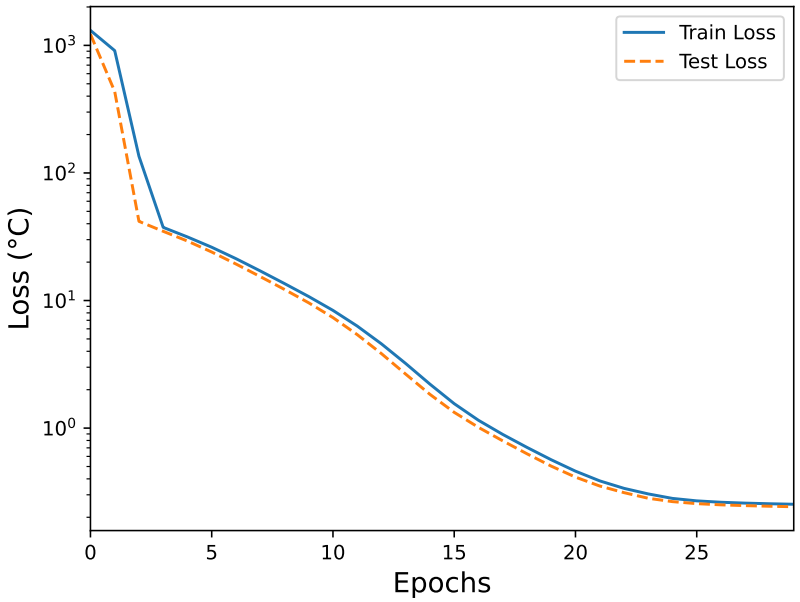


Fig. 2. Training loss on historical data for the system identification model.

Table 3
Results comparison for white-box model, energyPlus (E+) baseline and DDQN baseline, no internal loads.

	E+	*Proposed DDPG*			*DDQN* (Solinas et al., 2021)		
		$\gamma = 0.8$	$\gamma = 0.9$	$\gamma = 0.99$	$\gamma = 0.8$	$\gamma = 0.9$	$\gamma = 0.99$
PPD	25.11%	26.97%	23.85%	**21.89%**	27.12%	30.40%	27.82%
Coil Power	12765 kWh	**5672 kWh**	6617 kWh	9400 kWh	8311 kWh	9582 kWh	12816 kWh
HVAC Power	15132 kWh	**7508 kWh**	8471 kWh	11278 kWh	10509 kWh	11835 kWh	14838 kWh

Table 4
Results comparison for white-box model, energyPlus (E+) baseline and DDQN baseline, with internal loads.

	E+	*Proposed DDPG*			*DDQN* (Solinas et al., 2021)		
		$\gamma = 0.8$	$\gamma = 0.9$	$\gamma = 0.99$	$\gamma = 0.8$	$\gamma = 0.9$	$\gamma = 0.99$
PPD	19.33%	**14.58%**	16.38%	16.95%	21.13%	20.90%	19.89%
Coil Power	6317 kWh	**931 kWh**	1511 kWh	3109 kWh	2090 kWh	932 kWh	4890 kWh
HVAC Power	10076 kWh	**3265 kWh**	4200 kWh	5975 kWh	5378 kWh	4296 kWh	7589 kWh

other internal factor helps raising the temperature. Results for different values of γ are shown in Table 3: higher γ favours thermal comfort, lower γ favours energy saving, as observed and discussed more extensively in Solinas et al. (2021). The observed results clearly indicate how the DDPG algorithm is doing a better job at both respecting the thermal comfort of users and reducing energy consumption than the rule-based approach and the DDQN algorithm. More specifically, Table 3 shows how the proposed method can reach up to 50% overall energy consumption reduction with respect to the EnergyPlus standard controller, and up to 29% reduction with respect to the DDQN literature baseline, respectively from 15132 kWh for the EnergyPlus controller and from 10509 kWh for the DDQN baseline to 7508 kWh. These results are obtained by keeping the thermal comfort (PPD) increase under 8% with respect to the rule-based controller, from a PPD of 25.11% to 26.97%, while even slightly improving thermal comfort over the literature baseline. On top of that, if thermal comfort is considered as the main goal of the optimization, thus selecting a higher value for γ, a PPD increase of almost 10% is achievable with respect to the baseline controller, from 25.11% to 21.89%, while still reducing the energy consumption of around 26%.

More surprisingly, when internal loads are present, as shown in Table 4, the agent is actually capable of making use of this variable by drastically reducing the coil energy consumption by 85% with respect to the EnergyPlus baseline, from 6317 kWh to a mere 931 kWh, while also increasing the thermal comfort by around 25% with respect to the EnergyPlus and the literature baseline, from a PPD of respectively 19.33% and 20.90% to 14.58%. The proposed DDPG agent is capable of achieving such results by exploiting the fact that, in the specific use-case scenario, electrical appliances and people present in the building already provide all the heat required to warm the building up to the target setpoint, and not much additional heating power is needed. The agent mainly relies on supply air mass flow rate, in order to maintain good thermal conditions and account for possible overheating, therefore reducing the overall HVAC's consumption less drastically by around 65% and 24% with respect to the two baselines, from respectively 10076 kWh for EnergyPlus and 5378 kWh for the DDQN to 3265 kWh for the proposed DDPG method.

In general, it can be observed that the DDQN results are generally worse than that of DDPG, considering each different set of values for the hyperparameter γ. This might be due to the fact that the added control action variable, the supply air mass flow rate, severely increased the discrete action-space employed by this kind of RL agents, thus making the optimization harder to achieve with the same level of performance. Also, discrete action spaces allow for less fine-grained control over the HVAC system, reducing the overall effectiveness of the agent.

5.3. Online learning

This section presents the results achieved by the proposed online learning (see Section 3.3), directly connected to the target building.

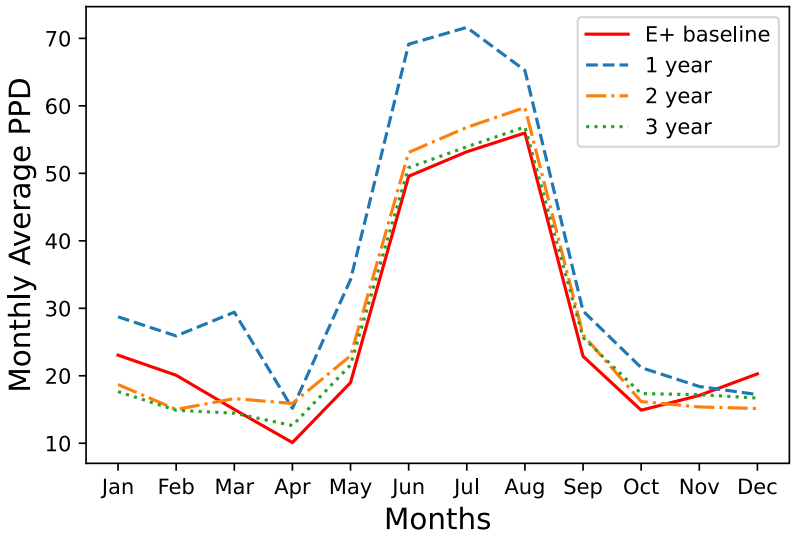


Fig. 3. Three years comparison between the EnergyPlus (E+) baseline and the proposed approach without pretraining, in terms of monthly average PPD.

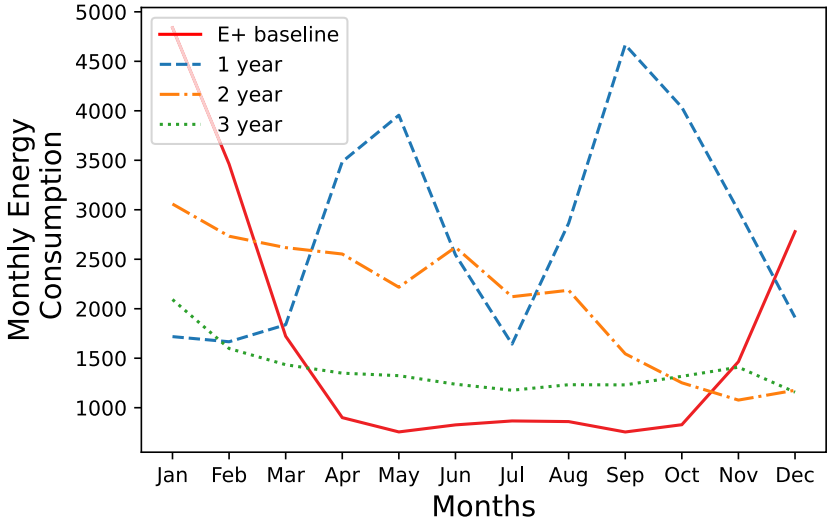


Fig. 4. Three years comparison between the EnergyPlus (E+) baseline and the proposed approach without pretraining, in terms of monthly total energy consumption.

In Fig. 3 and 4, it is possible to observe the results for the online training process without previous Imitation Learning. It is evident how the online DDPG agent performs worse than the baseline in the whole first year in terms of PPD (see Fig. 3), and for the most part of the year in terms of energy consumption (see Fig. 4). Only during the third year of training, the RL agent gets closer to the performance of the rule-based approach in terms of energy consumption, while the training has to get to its second year to match the performance in terms of thermal comfort.

When considering the agent that went through a previous Imitation Learning phase on historical data, it is immediately evident how this greatly enhances the agent performance since the very first moments of action. As shown in Fig. 5 the DDPG agent performs slightly worse than the baseline in terms of PPD, in the heating period, especially during its first month of training, while improving from the second month onward. Fig. 6 shows significantly more promising results as the agent is able to drastically reduce the energy consumption from the very first day of action and especially so in the whole heating season.

Table 5 reports the overall results after the first year of training for the rule-based EnergyPlus controller, the online DDPG algorithm without pre-training and with Imitation Learning pre-training. It is evident how the Imitation Learning technique provides significant benefits in terms of both PPD reduction and especially energy consumption with respect to the same online algorithm without the pre-training phase. Similarly, when compared with the EnergyPlus controller, the proposed DDPG and Imitation Learning approach is able to reduce energy consumption from the very first year by around 35%, from 20056 kWh to 13033 kWh, while only slightly increasing the PPD by less than 10%, from 26.92% to 29.43%.

Table 5
Results for the RL online training scenario on 1 year data.

	E+	*Proposed DDPG*	
		No Imitation Learning	*Imitation Learning*
PPD	**26.92%**	35.85%	29.43%
Coil Power	11102 kWh	18954 kWh	**5821 kWh**
HVAC Power	20056 kWh	33319 kWh	**13033 kWh**

Fig. 7 displays the monthly rewards achieved by the DDPG agent, both with and without the Imitation Learning phase. Fig. 8 shows the cumulative rewards over time for both agents. The cumulative reward is a measure of the total reward that the agent has accumulated over time, and it can provide an understanding of the overall performance of the agent. The reward metric has been presented in Section 3 in

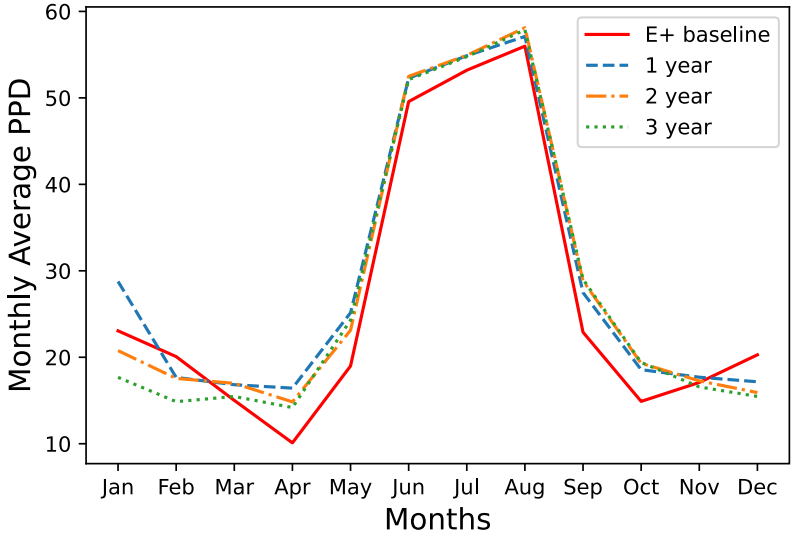


Fig. 5. Three years comparison between the EnergyPlus (E+) baseline and the proposed approach with Imitation Learning pretraining, in terms of monthly average PPD.

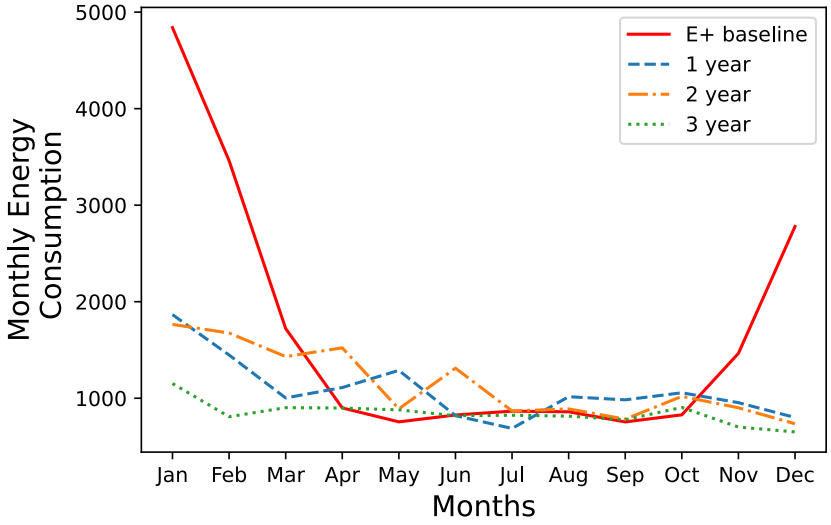


Fig. 6. Three years comparison between the EnergyPlus (E+) baseline and the proposed approach with Imitation Learning pretraining, in terms of monthly total energy consumption.

Eq. (2). The blue line represents the agent that was pre-trained with Imitation Learning, while the red line represents the agent without any pre-training. It can be again observed from the monthly rewards that the agent performing Imitation Learning achieves higher rewards especially in the very first months of training, compared to the agent without Imitation Learning. This further highlights the impact of the Imitation Learning on helping the agent grasp a good understanding of the dynamics of the HVAC system. Again, it is possible to note the seasonal nature of the data, with the agent being constrained to perform poorly during the summer months (from May to September) because no cooling action can be performed.

It is thus clear how the proposed method allows for a quick and efficient HVAC control of buildings, where energy consumption is reduced from the first few weeks of operation with respect to the traditional scheduled controllers. Furthermore, in this approach, the RL agent only requires a historical dataset in order to perform an offline Imitation Learning phase, and is capable of being deployed online on the target real-world building without the need for costly energy white-box modelling of the structure.

Table 6 presents an overall comparison of all the approaches presented in this work, including an EnergyPlus baseline and a literature baseline (Solinas et al., 2021). Results have been gathered by a testing phase on 3 months of data, in order to be consistent with the literature baseline. The online learning algorithm has been pre-trained following the Imitation Learning methodology and then directly deployed on the building. Its results correspond then to its very first 3 months of actual learning. Despite being the best performer in terms of both thermal comfort and energy consumption reduction, the approach based on a white-box modelling of the target building is costly, in terms of both computational resources and time, and thus not always feasible in all scenarios. The last proposed approach, based on an Imitation Learning pre-training phase on historical expert data and online learning on the target building is less performing overall but carries significant advantages in terms of being able to be almost immediately deployed on the target real building with negligible loss in terms of performance.

6. Concluding remarks

This paper has presented three different approaches to HVAC optimization through Reinforcement Learning techniques. In the first approach, a system identification model is adopted in which a neural network is trained on a 3-months historical dataset, and a DDPG agent is trained receiving thermal response feedback on the actions taken directly by the trained system identification model. This approach did

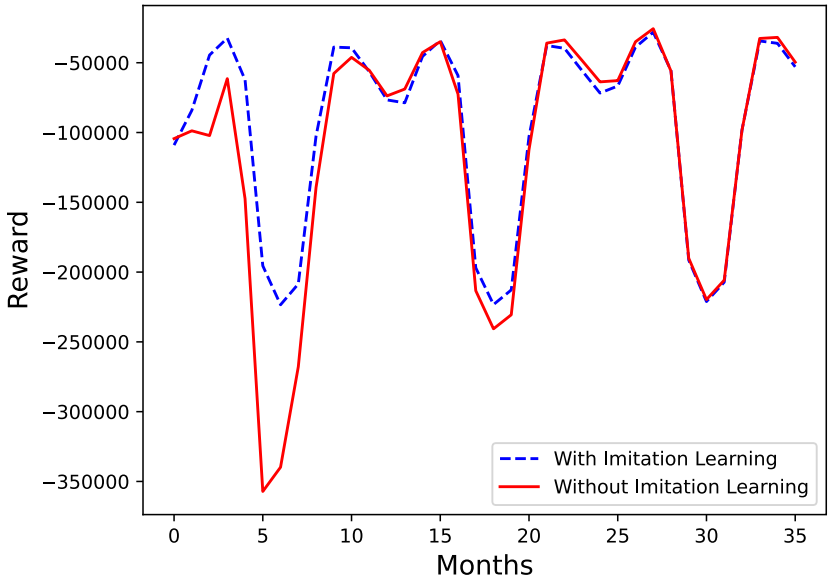


Fig. 7. Monthly Rewards for the DDPG agent with and without Imitation Learning.

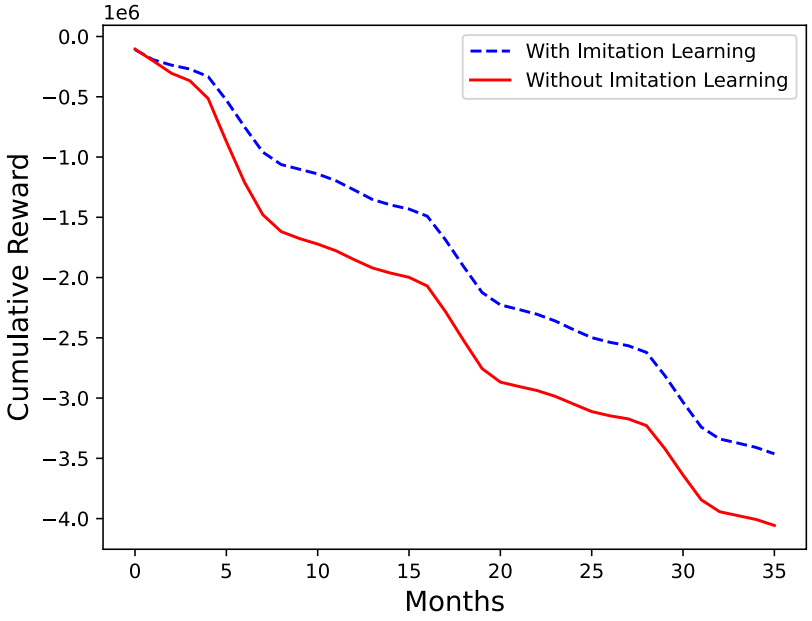


Fig. 8. Cumulative Rewards over Time for the DDPG agent with and without Imitation Learning.

not provide satisfactory results, as opposite to what is shown in Solinas et al. (2021). This is most likely due to the increased complexity of the action space in the proposed use-case scenario and the possible overfitting of the historical training dataset.

In the second approach, a white-box model is used to simulate the building thermal response and train the DDPG agent. As anticipated, this method led to superior outcomes as the simulation accurately represents the state-transition function that governs the system's thermodynamics. Theoretically, the white-box model holds a clear edge due to its close approximation of the physical system, which enables optimal energy consumption and comfort outcomes. However, despite this theoretical superiority, it is often impractical in real-world applications due to the extensive costs and time demands. Thorough modelling of each building and their HVAC systems—necessary for control and optimization—can be prohibitively expensive and time-consuming, limiting the approach's overall feasibility. Thus, while the white-box model approach provides the best performance in a controlled setting, its implementation costs highlight the necessity for alternative, more practical solutions.

Finally, an approach is presented in which the DDPG agent is trained directly on the target building, following an offline phase of imitation learning where it is exposed to a historical dataset gathered from a traditional rule-based controller operating on the building. This online approach demonstrates that, after only a few days of training, the proposed agent can match and even surpass the performances of the baseline controller in terms of energy consumption reduction, while maintaining user thermal comfort. Our proposed methodology addresses a key gap in the current state-of-the-art HVAC optimization research. Traditional RL methods are limited by the need to use extensive and costly simulations for training before practical application to mitigate the risk of, when applied to real-world systems, taking numerous suboptimal actions that might impair system efficiency and

Table 6
Comparison of results for the three proposed approaches and the two baselines on 3 months data.

	E+	*Proposed DDPG*			*DDQN* (Solinas et al., 2021)
		System Identification	*White Box*	*Online Imitation Learning*	
PPD	19.33%	45.76%	**14.58%**	21.09%	20.90%
Coil Power	6317 kWh	1209 kWh	**931 kWh**	1229 kWh	932 kWh
HVAC Power	10076 kWh	7508 kWh	**3265 kWh**	4316 kWh	4296 kWh

user comfort. Conversely, our approach incorporates an offline imitation learning phase, enabling the agent to start immediate interaction with the target real-world environment with good efficiency, without the need for complex modelling simulations. Although this method's performance may be slightly lower than the white-box model due to the inherent limitations in directly capturing the physical system's nuances, this minor trade-off in accuracy is more than offset by the substantial gains in efficiency and applicability. This approach is therefore viable for real-case scenarios where only historical data concerning the target building is typically available, and the agent needs to rapidly learn an optimal strategy to minimize time spent exploring ineffective and costly strategies.

In future work, we plan to refine our system identification model by investigating the impact of various hyperparameters, such as the number of connected layers in the neural network and learning rates, on the model's performance. This examination will contribute to better understanding why our current model fell short of expectations and will provide a basis for improvement. Additionally, we aim to explore more complex scenarios involving multiple thermal zones, which would enhance the realism and applicability of our approach. Moreover, through the utilization of Internet of Things capabilities for environmental variable monitoring and control, we intend to conduct a practical application of our approach to a real-world building. This would offer a more robust validation of its effectiveness and help identify any potential shortcomings. We believe that this iterative and critical approach to refining our methodology will result in substantial improvements in HVAC optimization, paving the way for more efficient and comfortable building environments.

CRediT authorship contribution statement

Francesco M. Solinas: Writing – original draft, Writing – review & editing, Methodology, Software, Validation, Investigation. **Alberto Macii:** Writing – original draft, Writing – review & editing. **Edoardo Patti:** Writing – original draft, Writing – review & editing, Supervision, Methodology, Validation. **Lorenzo Bottaccioli:** Writing – original draft, Writing – review & editing, Supervision, Conceptualization, Methodology, Software, Validation.

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.

Acknowledgement

This publication is part of the project NODES which has received funding from the MUR –M4C2 1.5 of PNRR with grant agreement no. ECS00000036

References

American Society of Heating, Refrigerating and Air Conditioning Engineers (Atlanta, Georgia) (2017). ANSI/ASHRAE standard 55-2017: thermal environmental conditions for human occupancy. *ASHRAE standard*, ASHRAE.

Barrett, E., & Linder, S. (2015). Autonomous hvac control, a reinforcement learning approach. In *Proc. of ECML PKDD 2015* (pp. 3–19). Springer.

Brandi, S., Piscitelli, M. S., Martellacci, M., & Capozzoli, A. (2020). Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings. *Energy and Buildings*, *224*, Article 110225.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., et al. (2016). Openai gym. arXiv preprint arXiv:1606.01540.

Chen, B., Cai, Z., & Bergés, M. (2019a). *Gnu-RL: A Precocial Reinforcement Learning Solution for Building HVAC Control Using a Differentiable MPC Policy* (pp. 316–325). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3360322.3360849.

Chen, B., Cai, Z., & Bergés, M. (2019b). Gnu-RL. GitHub repository, GitHub, https://github.com/INFERLab/Gnu-RL.

Crawley, D. B., Lawrie, L. K., Pedersen, C. O., & Winkelmann, F. C. (2000). Energy plus: energy simulation program. *ASHRAE Journal*, *42*(4), 49–56.

Ding, X., Du, W., & Cerpa, A. (2019). OCTOPUS: Deep reinforcement learning for holistic smart building control. In *Proc. of buildsys '19* (pp. 326–335). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3360322.3360857.

Drgoňa, J., Arroyo, J., Figueroa, I. C., Blum, D., Arendt, K., Kim, D., et al. (2020). All you need to know about model predictive control for buildings. *Annual Reviews in Control*.

Du, Y., Li, F., Munk, J., Kurte, K., Kotevska, O., Amasyali, K., et al. (2021). Multi-task deep reinforcement learning for intelligent multi-zone residential HVAC control. *Electric Power Systems Research*, *192*, Article 106959. http://dx.doi.org/10.1016/j.epsr.2020.106959, URL https://www.sciencedirect.com/science/article/pii/S0378779620307574.

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

Fu, C., & Zhang, Y. (2021). Research and application of predictive control method based on deep reinforcement learning for HVAC systems. *IEEE Access*, *9*, 130845–130852. http://dx.doi.org/10.1109/ACCESS.2021.3114161.

Gao, G., Li, J., & Wen, Y. (2020). DeepComfort: Energy-efficient thermal comfort control in buildings via reinforcement learning. *IEEE Internet of Things Journal*, *7*(9), 8472–8484. http://dx.doi.org/10.1109/JIOT.2020.2992117.

Hussein, A., Elyan, E., Gaber, M. M., & Jayne, C. (2018). Deep imitation learning for 3D navigation tasks. *Neural Computing and Applications*, *29*(7), 389–404.

Hussein, A., Gaber, M. M., Elyan, E., & Jayne, C. (2017). Imitation learning: A survey of learning methods. *ACM Computing Surveys*, *50*(2), 1–35.

Kou, X., Du, Y., Li, F., Pulgar-Painemal, H., Zandi, H., Dong, J., et al. (2021). Model-based and data-driven HVAC control strategies for residential demand response. *IEEE Open Access Journal of Power and Energy*, *8*, 186–197. http://dx.doi.org/10.1109/OAJPE.2021.3075426.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., et al. (2015). Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Rahimpour, Z., Verbič, G., & Chapman, A. C. (2020). Actor-critic learning for optimal building energy management with phase change materials. *Electric Power Systems Research*, *188*, Article 106543. http://dx.doi.org/10.1016/j.epsr.2020.106543, URL https://www.sciencedirect.com/science/article/pii/S0378779620303473.

Schaal, S. (1999). Is imitation learning the route to humanoid robots? *Trends In Cognitive Sciences*, *3*(6), 233–242.

Solinas, F. M., Bellagarda, A., Macii, E., Patti, E., & Bottaccioli, L. (2021). An hybrid model-free reinforcement learning approach for HVAC control. In *2021 IEEE international conference on environment and electrical engineering and 2021 ieee industrial and commercial power systems europe* (pp. 1–6). http://dx.doi.org/10.1109/EEEIC/ICPSEurope51590.2021.9584805.

Soyguder, S., Karakose, M., & Alli, H. (2009). Design and simulation of self-tuning PID-type fuzzy adaptive control for an expert HVAC system. *Expert Systems with Applications*, *36*(3, Part 1), 4566–4573. http://dx.doi.org/10.1016/j.eswa.2008.05.031, URL https://www.sciencedirect.com/science/article/pii/S0957417408002200.

Tiwari, S., Jain, A., Ahmed, N. M. O. S., Alkwai, L. M., Dafhalla, A. K. Y., & Hamad, S. A. S. (2022). Machine learning-based model for prediction of power consumption in smart grid-smart way towards smart city. *Expert Systems*, *39*(5), Article e12832.

United Nations (0000a). Energy, UN-Habitat, URL https://unhabitat.org/urban-themes/energy/.

United Nations (0000b). World Urbanization Prospects, Population Division, URL https://population.un.org/wup/.

Vázquez-Canteli, J. R., & Nagy, Z. (2019). Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied Energy*, *235*, 1072–1089.

Wang, Z., & Hong, T. (2020). Reinforcement learning for building controls: The opportunities and challenges. *Applied Energy*, *269*, Article 115036.

Wei, T., Wang, Y., & Zhu, Q. (2017). Deep reinforcement learning for building HVAC control. In *Proc. of DAC 2017* (pp. 1–6).

Wigle, L. (2014). How the Internet of Things will enable vast new levels of efficiency alan rose, intel corporation Dr. Subramanian Vadari, modern grid solutions.

Yang, Z., & Becerik-Gerber, B. (2014). Coupling occupancy information with HVAC energy simulation: A systematic review of simulation programs. In *Proceedings of the winter simulation conference 2014* (pp. 3212–3223). http://dx.doi.org/10.1109/WSC.2014.7020157.

Zhang, C., Kuppannagari, S. R., Kannan, R., & Prasanna, V. K. (2019). Building HVAC scheduling using reinforcement learning via neural network based model approximation. In *Proc. of BuildSys '19* (pp. 287–296). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3360322.3360861.

Zhang, Z., & Lam, K. P. (2019). Gym-eplus. GitHub repository, GitHub, https://github.com/zhangzhizza/Gym-Eplus.