POLITECNICO DI TORINO
Repository ISTITUZIONALE

Toward Feasible Deep Reinforcement Learning for HVAC Control: Training Strategies Compared

Original
Toward Feasible Deep Reinforcement Learning for HVAC Control: Training Strategies Compared / Loffa, M.A., Rando Mazzarino, P., Bertolotti, F., Patti, E., Bottaccioli, L.. - (2026). (26th IEEE EEEIC International Conference on Environment and Electrical Engineering (EEEIC) Lisbon, Portugal 29 June – 02 July 2026).

Availability:
This version is available at: 11583/3018458 since: 2026-11-26T13:17:38Z

Publisher:
IEEE

Published
DOI:

Terms of use:

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

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

28 November 2026

Toward Feasible Deep Reinforcement Learning for HVAC Control: Training Strategies Compared

Maria Adelaide Loffa*, Pietro Rando Mazzarino*, Filippo Bertolotti†, Edoardo Patti*, Lorenzo Bottaccioli*
* Politecnico di Torino, Turin, Italy
name.surname@polito.it
† Messana Hydronic Technologies, Soquel, CA, USA
fbertolotti@messana.tech

***Abstract*—As the power system evolves toward greater electrification and flexibility, buildings are emerging as active participants in ensuring energy efficiency and operational stability. Achieving this integration requires advanced control strategies capable of optimizing building performance while intelligently responding to dynamic conditions. Deep Reinforcement Learning (DRL) provides a powerful framework for developing such adaptive and efficient control solutions. This study aims to identify the most effective training strategies for potential real-world deployment of DRL control algorithms using three state-of-the-art algorithms DDPG, TD3, SAC and comparing their behaviour under varying training experience. Each algorithm is trained within the same simulated building environment using the Energym platform under three distinct training configurations: fully online, fully offline, and offline with online fine-tuning. Their performance is analyzed and benchmarked against a performant Proportional-Integrative-Derivative controller. Beyond algorithmic comparison, the work emphasizes a practical, deployment-oriented assessment of these approaches, providing a unified comparison of online, offline, and hybrid training requirements to evaluate their suitability for real-world intelligent building energy management**

***Index Terms*—Deep Reinforcement Learning, HVAC control, Reinforcement learning training, energy efficiency, thermal comfort, Demand side management**

I. Introduction

In the context of rising electricity demand and increasingly complex power systems, ensuring grid efficiency and stability has become a central challenge. The transition to smarter, electrified infrastructures requires advanced control mechanisms that balance flexibility, reliability, and efficiency. Smart Grid technologies enable sophisticated Demand Response (DR) strategies, dynamically adjusting consumption to support real-time grid balancing and alleviate peak loads. Residential buildings account for roughly 40% of European total energy consumption [1], and with the ongoing electrification of building systems, they are poised to play a pivotal role in managing power grid imbalances. Deep Reinforcement Learning (DRL) has emerged as a powerful method for adaptive building energy management, optimizing performance while interacting dynamically with related systems such as the power grid. In online learning scenarios [2], [3], the DRL agent derives the optimal policy through continuous interaction, though early trial-and-error may cause thermal discomfort or excess energy use. Offline learning [4] relies on pre-collected datasets, but requires substantial, representative data [5]. A practical compromise is hybrid offline-to-online training, where offline pre-training is followed by online fine-tuning [6]. Continuous action space algorithms outperform discrete methods (e.g., DQN [7]) by enabling smoother, more precise control [8]. Key examples include Deep Deterministic Policy Gradient (DDPG) and Twin Delayed DDPG (TD3) as deterministic algorithms, and Soft Actor-Critic (SAC) as a stochastic alternative. Comparative studies in energy systems show varied strengths: DDPG and SAC for PV Maximum Power Point Tracking [9], DDPG, TD3, and SAC for energy storage optimization [10], and TD3 versus multi-agent SAC for DC microgrids [11], highlighting differences in convergence speed, stability, and robustness. This study directly compares DDPG, TD3, and SAC for building HVAC control.

The goal of this paper is to investigate how different training strategies shape the behavior, requirements, and deployability of three Deep Reinforcement Learning (DRL) algorithms, DDPG, TD3, and SAC, for HVAC heating control in an electrically supplied simulated building. While the algorithms are compared against a rule-based Proportional-Integral-Derivative (PID) controller in terms of thermal comfort and energy consumption, the primary focus of the study is on understanding the influence of offline learning, online learning, and hybrid offline-to-online fine-tuning on training data needs, convergence speed, stability, and the occurrence of undesirable control actions during learning. By analyzing these strategies within a unified experimental framework, the work provides practical guidance on which training approaches are most suitable for real-world deployment, addressing key challenges related to data availability, safe exploration, and operational robustness. The findings contribute to advancing DRL-based HVAC control toward reliable, scalable, and grid-responsive building energy management.

II. Proposed Framework

This proposed study aims to benchmark and compare different training approaches and different DRL algorithms in solving a heating control task in a simple building that is electrically supplied due to the integration of a heating pump system as a generation module.

This work is supported by SINEGRA - Sistema INtelligente per l'Efficienza e la Gestione delle Reti Avanzate -, an Italian National project funded by the Ministry of Environment and Energy Security - MASE.

A. Environment development and data

The chosen environment for this study is implemented using Energym [12], an open-source Python interface for Modelica [13] building models. The selected case study is called *SimpleHouseRad* (from Energym documentation [14]) and it consists of a single-zone, poorly insulated residential building equipped with a water-to-water heat pump and a radiator system. This type of heat pump transfers heat between two water loops, a source side and a load side, and is commonly used in both heating and cooling applications. However, for the purpose of this study no cooling is considered.

The building thermal dynamics and the heating system are modeled through the EnergyPlus engine, while weather boundary conditions are imported from the NASA POWER database [15] for the location of Aosta (Italy). Only heating months (January–March and October–December) are considered. The simplicity and physical coherence of this configuration allow reproducible assessments of control strategies. The single-zone heating model was selected to isolate the effect of training paradigms on algorithm performance, thereby avoiding confounding influences from multi-zone dynamics, occupancy, or nonlinear HVAC interactions. This ensures reproducibility of results, which is essential for benchmarking reinforcement learning algorithms, where differences in building models could otherwise obscure comparative outcomes.

B. Control strategies and DRL algorithms

The control problem in this system is designed as a Markov Decision Process in which, at every step, given the boundary and indoor condition, the power supplied from the heatpump is controlled. The typical objective in these kinds of problems relies on finding the right tradeoff between respecting indoor thermal comfort and reducing as much as possible the power consumption. In this context, we present a baseline control approach used for the benchmarking, three DRL algorithms, and three different training strategies.

a) PID Baseline Controller: The baseline is a Proportional-Integral-Derivative (PID) controller, widely used in HVAC systems due to its simplicity and robustness. It regulates the indoor temperature by minimizing the tracking error between the measured temperature and a comfort setpoint. The control input is computed as the sum of three components: a proportional term reacting to the instantaneous error, an integral term accumulating past errors, and a derivative term anticipating future trends. The gains were manually tuned to balance comfort and energy consumption, with Kp = 0.1, Ki = 0, and Kd = 220. This configuration emphasizes fast, anticipative responses through the derivative term while avoiding integral wind-up, which is unnecessary given the slow and near-linear thermal dynamics of the system. Under these conditions, the PID controller achieves performance close to optimal for the selected system. Its strong performance, combined with its widespread adoption in HVAC applications, makes it a meaningful and non-trivial benchmark against which improvements achieved by reinforcement learning agents can be assessed. However, it is well known that PID controllers exhibit limited scalability and rapidly degrading performance as system dynamics become more complex or less predictable, whereas deep reinforcement learning methods are better suited to such environments [16].

The PID controller was also used to generate datasets for offline training and to serve as a reference point for evaluating relative improvements in comfort and energy efficiency.

b) Deep Reinforcement Learning Algorithms: Three different DRL algorithms are investigated:

- **Deep Deterministic Policy Gradient (DDPG)**: introduced in al. [17], is a model-free, off-policy actor–critic algorithm for continuous action spaces. Uses experience replay and target networks for stability. Actor output is perturbed with noise for exploration. Sensitive to hyperparameters and prone to overestimation bias.
- **Twin Delayed DDPG (TD3)**: it is an extension of DDPG, specifically designed to address its main limitations, namely overestimation bias in the critic and instability during training [18]. It reduces overestimation and improves stability via twin critics, delayed policy updates, and target policy smoothing. Well-suited for HVAC control, where thermal inertia and external disturbances can otherwise degrade the performance of deterministic policies.
- **Soft Actor–Critic (SAC)**: it is an off-policy actor-critic algorithm grounded in the maximum entropy reinforcement learning framework [19]. It learns a stochastic policy outputting action distributions, balancing exploration and exploitation. SAC maximizes reward and policy entropy, improving robustness to modeling errors and disturbances. Like TD3, it employs two critics and uses the minimum Q-value to reduce overestimation bias.

These algorithms represent the state of the art in continuous control and provide a comparison between deterministic (DDPG, TD3) and stochastic (SAC) policies.

For a fair comparison, all agents share the same environment, action space, state representation, and reward function. The action space corresponds to the normalized heat pump power in the range [0, 1], mapping to 0–5000 W. This action variable is directly related to the capacity of heating the water in the emission system, and the electrical power consumed by this action is an output of the heating pump model. The state received by the agents is represented in Equation 1 and consists of the heatpump power consumption $P\ [W]$, the temperature supplied to the radiators $T_s\ [K]$, the outdoor air temperature $T_o\ [K]$ and the temperature difference between the room setpoint and the actual room indoor temperature $\Delta T\ [K]$.

$$S_t = \big(P\ [W],\quad T_s\ [K],\quad T_o\ [K],\quad \Delta T\ [K]\big) \tag{1}$$

Finally, Equation 2 represents the used reward function. The reward at time step t, denoted as r_t, is expressed as a weighted penalty that combines thermal comfort and energy consumption. The first term, $\alpha\,|T_{in,t} - T_{set}|$, represents the absolute deviation of the indoor air temperature $T_{in,t}$ from

the setpoint T_{set}, scaled by a comfort weight α. This term penalizes violations of the thermal comfort constraint, with larger deviations incurring higher penalties. The second term, $\beta\, P_t$, corresponds to the heating power consumption P_t at the same timestep, multiplied by the energy weight β. This component ensures that the agent is discouraged from excessive energy use even when comfort is satisfied.

$$r_t = -\left(\alpha\ |T_{in,t} - T_{set}| + \beta\, P_t\right) \tag{2}$$

C. Training approaches and Evaluation

For each agent type, three training approaches are considered: (i) online learning, (ii) offline learning, and (iii) offline-to-online learning. In online learning, agents interact directly with the environment and continuously update their policies. In offline learning, agents are trained exclusively on datasets generated using the PID controller. These datasets span different lengths (1–12 months) and weather conditions to ensure diversity. The offline-to-online approach initializes agents through offline training and then performs a limited online fine-tuning phase lasting 1–12 weeks.

Across these methods, a quasi-optimal choice of the hyperparameters, as well as the reward weights, is conducted using the Optuna library [20]. Each trial consists of one year of training followed by evaluation over a six-month test period with unseen weather data. The objective function captures the trade-off between energy consumption and thermal comfort, ensuring balanced performance. The search space includes standard parameters such as learning rates, discount factor, batch size, replay buffer settings, and target update coefficients, as well as algorithm-specific parameters (e.g., target smoothing for TD3 and entropy temperature for SAC). Optimization is performed separately for each training strategy, and the best configurations are selected after convergence. Optimization was conducted separately for the online, offline, and hybrid offline-to–online training strategies. For the hybrid offline-to–online approach, the offline pre-training phase was initialized using the weights of the best-performing offline models, with the same hyperparameters as in the offline experiments. The only differences concern the update interval and the number of updates, for which smaller values were selected to reflect the shorter online fine-tuning phase. Specifically, the best performing offline configurations over the fixed evaluation period were selected: the models offline trained with one month of data for DDPG and TD3, and the model offline trained with two months of data for SAC. The selected hyperparameters are shown in Table I. The weights of the reward function, expressed in 2, were also optimized using *Optuna* to identify the combination that enables the agent to achieve the best trade-off between minimizing energy consumption and maintaining thermal comfort. The estimated values are $\beta = 0.6$ and $\alpha = 2.4$. This systematic and consistent hyperparameter tuning procedure enables a fair comparison across algorithms and training strategies, ensuring that observed performance differences are attributable to the learning approach rather than suboptimal parameter choices.

TABLE I
OPTIMIZED HYPERPARAMETERS SELECTED VIA OPTUNA FOR EACH ALGORITHM AND TRAINING STRATEGY.

	DDPG			TD3			SAC		
Parameter	Online	Offline	Off–On	Online	Offline	Off–On	Online	Offline	Off–On
Discount factor γ	0.95	0.88	0.95	0.90	0.88	0.88	0.88	0.88	0.88
Actor learning rate	7.0e–4	9.3e–5	7.0e–4	6.0e–4	3.9e–4	9.9e–4	9.9e–4	1.3e–4	9.9e–4
Critic learning rate	2.0e–4	8.6e–4	2.0e–4	1.3e–4	2.4e–4	2.7e–5	2.7e–5	1.2e–5	2.7e–5
Batch size	2048	256	2048	2048	128	2048	2048	512	2048
Soft update coefficient τ	0.05	0.05	0.05	0.05	0.001	0.02	0.02	0.05	0.02
Replay buffer size	50k	–	50k	50k	–	50k	50k	–	50k
Update interval	8	–	4	9	–	2	8	–	2
Number of updates	2	–	1	1	–	1	1	–	1
Target smoothing σ	–	–	–	0.2	0.5	0.2	–	–	–
Target smoothing clip	–	–	–	0.5	0.3	0.5	–	–	–
Entropy temperature lr	–	–	–	–	–	–	2.0e–4	1.1e–5	2.0e–4

The performance of all control strategies is quantified using a set of Key Performance Indicators. Standard metrics include the average indoor temperature error (absolute deviation from setpoint, in Kelvin) and the average thermal power consumption (W) from which a percentage of savings with respect to the baseline is extracted. In addition, a stepwise classification scheme is applied during training, categorizing each timestep into four operational quadrants (comfort satisfied/violated × low/high energy). This additional metric is used to evaluate the operability of agents during the learning phase, providing insight into the suitability of each training paradigm for deployment. A systematic simulation campaign ensures comparability. Weather data from four years (2020–2024) are used for training and interaction, while testing is conducted on an unseen year (2019). Offline datasets vary in length (1, 2, 3, 6, and 12 months), with 5-minute resolution and multiple weather scenarios to increase robustness. Each configuration is evaluated over multiple random seeds, and results are averaged and compared against the PID baseline to assess both steady-state performance and convergence behavior.

III. RESULTS

This section presents RL results under different strategies, focusing on indoor temperature error and thermal power consumption (Figures 1 and 2), and evaluates real-world deployment potential. All algorithms were trained under fully online and fully offline regimes. These experiments guided a hybrid approach, initializing each agent from its best offline model and applying brief online fine-tuning for rapid, reliable deployment.

a) Online learning: The online training results (Fig. 1, top row, circular markers) show that DDPG and TD3 quickly reduce temperature deviation below 1 K within one month, improving further over longer training. SAC exhibits higher early variability, with deviations up to 4 K in the first two months, but gradually stabilizes to levels comparable with DDPG and TD3 after about six months, highlighting the need for extended interaction for convergence. All deviations initially exceed the PID baseline, which performs strongly over the full evaluation period. A similar pattern appears in power consumption (Fig. 1, bottom row). DDPG and TD3 remain stable across training durations, slightly increasing with time but consistently below the PID controller (dash-dotted line); for example, TD3 consumes 1719 W after one month and 1738 W after twelve months. SAC shows a more

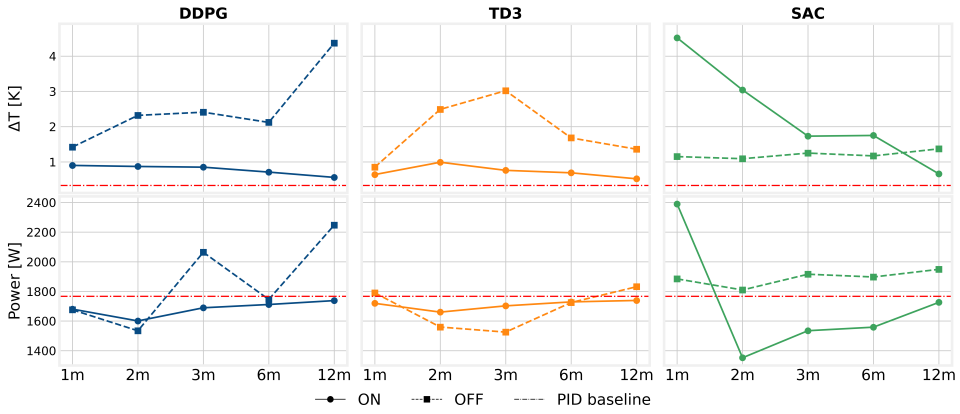


Fig. 1. Mean indoor temperature error (top) and average thermal power consumption (bottom) for DDPG, TD3, and SAC. Circles denote online training and squares offline training, The dash-dotted line indicates PID baseline controller.

complex trend: performance improves significantly in the first two months, then power consumption gradually rises for 3, 6, and 12 months, yet it remains below the PID benchmark.

b) Offline learning: Under exclusive offline training, the three DRL algorithms show markedly different behavior in thermal comfort and energy performance (Fig. 1, square markers). SAC consistently outperforms DDPG and TD3, maintaining average temperature deviations near 1 K with just two months of data and remaining stable as dataset size increases. TD3 performs well with limited data, reaching low deviations after one month, making it suitable for data-scarce scenarios. However, its stability decreases with larger datasets, with deviations rising to 3 K after three months, suggesting sensitivity to dataset composition and potential overfitting. Energy-wise, TD3 initially reduces power consumption, but usage increases over time, eventually exceeding the PID baseline at twelve months. DDPG shows the most erratic behavior, with large deviations up to 4 K and power peaks exceeding 4.8 kW, resulting in high and inconsistent energy consumption.

In summary, under exclusive offline training, SAC demonstrates the strongest robustness and data efficiency among the evaluated algorithms, while TD3 remains competitive only for small datasets. Nevertheless, all offline-trained agents underperform relative to both online-trained policies and the PID benchmark, indicating that offline data alone are insufficient to ensure reliable building control performance.

c) Offline-to-Online learning: The hybrid offline-to-online training strategy demonstrates the most promising results, combining the data efficiency of offline pre-training with the adaptability of online fine-tuning. The online phase was conducted over short durations, 1 and 2 weeks ans 1,2 and 3 months, reflecting a practical scenario for real-world deployment, where extended training periods of 6 to 12 months are often infeasible.

Figure 2 presents KPIs for temperature deviation and energy consumption across varying lengths of online fine-tuning. DDPG and TD3 converge rapidly, achieving average temperature deviations below 0.7 K after only one to two weeks of online training. SAC, while initially less stable during the first week, benefits significantly from offline pre-training,

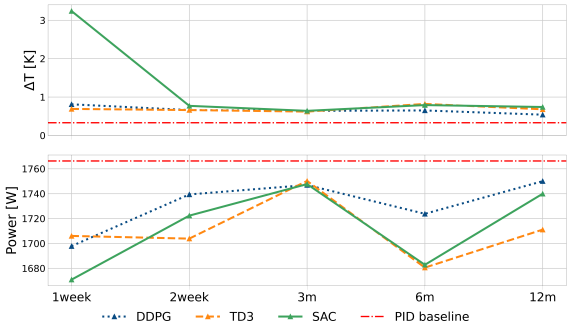


Fig. 2. Mean indoor temperature error (top) and average thermal power consumption (bottom) for DDPG, TD3, and SAC trained with the offline-to-online method.The dash-dotted line indicates PID baseline controller.

reaching ΔT values below 0.8 K after a brief adjustment period. In terms of power consumption, the trends exhibit an initial low point, followed by a gradual increase up to approximately one month of deployment, after which values fluctuate slightly, likely due to seasonal variation. However, all three algorithms maintain average values below the PID baseline, demonstrating energy-efficient behavior. DDPG, TD3 adn SAC show consistent performance in optimizing thermal comfort while balancing energy efficiency, with their average power consumption stabilizing around 1.7 kW.

A. Deployability implications

The primary goal of this work was to identify a training strategy that enables RL-based HVAC controllers to be deployed without compromising thermal comfort or energy efficiency. Safe real-world training is generally infeasible, as extensive exploration can degrade comfort and raise costs. Consequently, RL controllers are typically trained in virtual environments simulating building dynamics. However, constructing accurate simulators is complex, labor-intensive, and error-prone, and discrepancies between simulated and real behavior can limit policy transferability. This raises the question: to what extent can RL agents be trained for deployment without highly accurate simulators or long, unsafe exploration phases?

Our analysis shows that neither purely online nor purely offline learning provides a practical path. Online training requires months of interaction with the real system, during which comfort violations and inefficiencies are inevitable. Offline training suffers from dataset limitations and generalization challenges, and collecting sufficiently rich datasets demands extensive sensors and prolonged data campaigns. The hybrid offline-to-online strategy offers a compromise between the safety of model-based initialization and the adaptability of online learning. Figure 3 shows the quality of the control actions in the offline-to-online scenario during the first weeks of online fine-tuning. The characterization of actions accounts for the two concurrent objectives: temperature deviation and energy savings. An action is classified as good control when the average temperature deviation remains below 0.5,°C. For the RL algorithms, good actions are further divided depending on whether the controller consumed less or more energy than the PID benchmark ("savings" or "waste"). The PID controller is also reported; however, for the PID, the terms savings and waste do not apply, and only temperature deviation is considered. The PID controller is a strong benchmark and achieves around 90% good control even under the strict 0.5 °C threshold. Nevertheless, it still exhibits a small portion of bad control actions. This is due to the physical limitations of the environment, which prevent the system from instantly reaching the desired temperature when the setpoint changes abruptly (e.g., in the early morning when it increases from 16 °C to 20 °C).

Looking at the RL agents' performance in Figure 3, we observe the benefits of online fine-tuning within just a few weeks, as reflected by the increasing percentages of good actions and the gradual reduction of the intrinsic bad-control actions also present under the performant PID strategy. For deterministic algorithms such as DDPG and TD3, one month of offline data is sufficient to learn a competent initial policy. As reported in Figure 3, the strict 0.5 °C threshold causes many actions to fall under the "bad control with savings" category. However, in practice, all algorithms except SAC already in the first week of online tuning (i.e., still mostly governed by the offline-learned policy) never experience maximum temperature deviations exceeding 2 °C, a range that is fully acceptable and almost unnoticeable to occupants, while still achieving higher energy savings than the PID controller. The SAC algorithm exhibits some peak deviations of around 6 °C in the first week of online fine-tuning, but it quickly stabilizes afterward, falling back within the same deviation ranges as the other agents. In summary, deterministic algorithms with offline pre-training followed by short online tuning enable rapid adaptation to real buildings. Stochastic algorithms like SAC follow a similar trend but need more interactions to stabilize. The hybrid approach reduces both offline data requirements and risky online exploration, providing safe initialization, rapid convergence, and strong long-term performance, making it the most practical strategy for real-world HVAC control examined in this study.

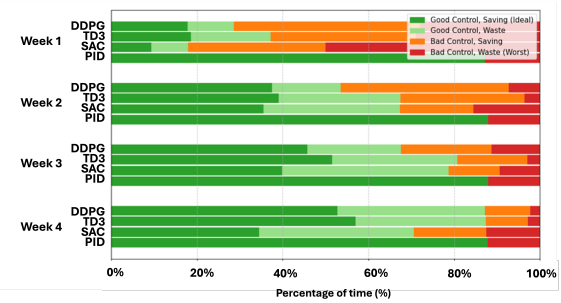


Fig. 3. Quality percentages of the control actions during the first 4 weeks of online fine-tuning for each algorithm and benchmarked with the PID control behaviour

IV. Lesson Learned

In this section, we provide a summary of the insights gathered throughout this study, with the aim of offering a practical and concise reference on key considerations for RL-based HVAC control in buildings.

1) **Hybrid offline-to-online advantages**:Hybrid training is the most effective strategy for real-world deployment. Purely online learning requires months of unstable exploration and risky interactions, while purely offline learning often fails to reach optimal policies. The hybrid approach leverages historical data for a safe starting policy and adapts quickly to actual building dynamics through brief online fine-tuning.
2) **Multi-objective rewards and different learning speed**: A key insight is that different control objectives can be learned at markedly different speeds. For example, when considering conflicting goals such as comfort and energy use, comfort-related regulation tends to stabilize relatively quickly, whereas learning to minimize energy consumption while still maintaining acceptable comfort requires a much longer period of exposure to varying environmental conditions, in order for the system to fully understand and balance the underlying trade-offs. This difference depends on the nature of the objective itself and can be mitigated through proper reward function engineering and weighting.
3) **Learning Dip**: Learning progress is not always linear. For DDPG and TD3, a temporary performance dip was observed around one month, caused by distribution shifts such as seasonal weather changes or replay buffer drift. Older data can temporarily misguide policy updates until the agent adapts to new conditions. Understanding data seasonality, outliers, and structural differences in objectives is therefore essential.
4) **Algorithm Specialization and Trade-offs**: The choice of algorithm should align with the deployment strategy. Deterministic methods like DDPG and TD3 adapt quickly and deliver meaningful results within the first few weeks of online fine-tuning. Stochastic methods such as SAC provide stronger long-term robustness and excel in offline training, but they may exhibit early instability during real-world adaptation. This highlights

the trade-off between rapid initial adaptation and reliable long-term performance, which should guide algorithm selection based on the specific problem.

5) **Structural Limits and Benchmark Validity**:It is important to distinguish limitations of the AI controller from inherent physical constraints of the system. In our study, a "bad control" rate of 12–15% was due to system dynamics, not agent failure. Recognizing these structural limits is crucial when evaluating RL agents. Furthermore, in settings where strong benchmarks already exist, even modest improvements achieved by RL methods can be practically significant.

V. Conclusion

This study conducted a systematic comparison of three DRL algorithms, DDPG, TD3, and SAC, for HVAC control under Online, Offline, and hybrid Offline-to-Online learning paradigms. The results demonstrate that the adopted training strategy critically affects both the learning speed and stability and the practical feasibility of DRL-based approaches for real-world deployment in building energy management systems.

Among the tested configurations, the hybrid Offline-to-Online learning scheme emerged as the most effective and deployable. By combining offline pre-training on historical data with a limited period of online fine-tuning, this approach achieved rapid convergence, robust performance, and reduced operational risks. Specifically, TD3 and DDPG demonstrated stable regulation and energy savings within two weeks of online adaptation, reaching performance comparable to or exceeding that of the PID benchmark. SAC exhibited strong generalization capability in offline scenarios but required extended fine-tuning to achieve similar robustness and efficiency.

Beyond algorithmic performance, the findings highlight the potential of DRL-based HVAC controllers to serve as flexible, grid-responsive resources. By autonomously adjusting heating demand in response to dynamic conditions, these controllers can support demand-side management and contribute to power system flexibility while preserving thermal comfort. The hybrid training paradigm, in particular, offers a practical pathway for safe and efficient deployment in operational environments, minimizing the data and interaction requirements traditionally associated with reinforcement learning.

Future research will focus on extending this framework toward multi-zone and multi-agent control architectures, incorporating occupant-aware adaptation, price-based or frequency-based demand response, and coordination across building aggregations. These developments will further consolidate the role of DRL as a key enabler for energy-efficient, flexible, and grid-interactive buildings within next-generation smart power systems.

References

[1] "Directive (eu) 2024/1275 of the european parliament and of the council," 2024.

[2] M. A. Loffa, A. J. Donato, P. R. Mazzarino, E. Macii, A. Osello, M. Del Giudice, E. Patti, and L. Bottaccioli, "A bimtobem framework for building retrofit and hvac smart control assessment," in *2025 IEEE International Conference on Environment and Electrical Engineering and 2025 IEEE Industrial and Commercial Power Systems Europe (EEEIC / ICPS Europe)*, 2025, pp. 1–6.

[3] S. Xu, Y. Fu, Y. Wang, Z. Yang, Z. O'Neill, Z. Wang, and Q. Zhu, "Accelerate online reinforcement learning for building hvac control with heterogeneous expert guidances," in *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 2022, pp. 89–98.

[4] C. Blad, S. Bøgh, and C. S. Kallesøe, "Data-driven offline reinforcement learning for hvac-systems," *Energy*, vol. 261, p. 125290, 2022.

[5] V. M. Nik and K. Javanroodi, "Adaptive reinforcement learning for energy management–a progressive approach to boost climate resilience and energy flexibility," *Advances in Applied Energy*, vol. 17, p. 100213, 2025.

[6] H.-Y. Liu, B. Balaji, R. Gupta, and D. Hong, "Adaptive policy regularization for offline-to-online reinforcement learning in hvac control," in *Proceedings of the 11th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 2024, pp. 1–10.

[7] E. McKee, "Deep reinforcement learning for real-time residential hvac control," 2019.

[8] G. Gao, J. Li, and Y. Wen, "Deepcomfort: Energy-efficient thermal comfort control in buildings via reinforcement learning," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8472–8484, 2020.

[9] S. E. Nwachukwu, K. A. Folly, and K. O. Awodele, "A comparative study between soft actor-critic (sac) and deep deterministic policy gradient (ddpg) algorithms for solar pv mppt control under partial shading conditions," *IEEE Access*, 2025.

[10] H. Noh and Y. Lim, "Optimizing energy storage systems using deep reinforcement learning in smart grids." *Journal of Information Processing Systems*, vol. 21, no. 2, 2025.

[11] M. S. Khanghah, F. Moghaddam, A. Suratgar, and M. B. Menhaj, "Sac-based distributed optimal control of generation cost in dc microgrids," in *2025 IEEE Seventh International Conference on DC Microgrids (ICDCM)*. IEEE, 2025, pp. 1–7.

[12] P. Scharnhorst, B. Schubnel, C. Fernandez Bandera, J. Salom, P. Taddeo, M. Boegli, T. Gorecki, Y. Stauffer, A. Peppas, and C. Politi, "Energym: A building model library for controller benchmarking," *Applied Sciences*, vol. 11, no. 8, p. 3518, 2021.

[13] P. Fritzson and V. Engelson, "Modelica—a unified object-oriented language for system modeling and simulation," in *European Conference on Object-Oriented Programming*. Springer, 1998, pp. 67–90.

[14] P. Scharnhorst, B. Schubnel, C. Fernández Bandera, J. Salom, P. Taddeo, M. Boegli, T. Gorecki, Y. Stauffer, A. Peppas, and C. Politi, "Energym: A building model library for controller benchmarking," *Applied Sciences*, vol. 11, no. 8, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/8/3518

[15] N. L. R. Center, "Nasa power data access viewer," 2025, nASA Langley Atmospheric Science Data Center (ASDC). [Online]. Available: https://power.larc.nasa.gov/data-access-viewer/

[16] A. Gharbi, M. Ayari, N. Albalawi, Y. E. Touati, and Z. Klai, "Intelligent hvac control: Comparative simulation of reinforcement learning and pid strategies for energy efficiency and comfort optimization," *Mathematics*, vol. 13, no. 14, p. 2311, 2025.

[17] T. P. Lillicrap, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[18] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International conference on machine learning*. PMLR, 2018, pp. 1587–1596.

[19] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft actor-critic algorithms and applications," 2019. [Online]. Available: https://arxiv.org/abs/1812.05905

[20] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631.