Control Optimization of Residential Air-Source Heat Pumps Using Deep Reinforcement Learning

Ben Anton Goff
University of Applied Sciences Augsburg
Email: bengoff16@gmail.com

***Abstract*—This work investigates the application of deep reinforcement learning (DRL) to optimize the control of residential air-source heat pumps (ASHPs). Three state-of-the-art DRL algorithms—Soft Actor-Critic (SAC), Deep Q-Network (DQN), and Advantage Actor-Critic (A2C)—are evaluated in a custom thermal simulation environment incorporating a two-zone RC thermal model and temperature-dependent heat pump efficiency. The agents are trained to minimize a multi-objective cost function balancing thermal comfort, energy consumption, and compressor cycling. Experimental results demonstrate that SAC achieves 26% better performance than a classical PID baseline controller, with 17% energy reduction and 44% fewer comfort violations. This study confirms the potential of deep RL to surpass traditional control methods in building energy systems.**

I. Motivation and Background

Heating and cooling account for a major share of residential energy use, and in many European countries air-source heat pumps (ASHPs) are being installed at record rates to replace fossil-fuel boilers. While these systems are efficient, they are typically operated with fixed rule-based control parameters that cannot adapt to the building's changing thermal dynamics, occupancy, or weather conditions. This results in suboptimal performance, excessive compressor cycling, and unnecessary energy use.

Every dwelling exhibits unique thermal inertia, insulation, and occupant behavior, meaning that a single static control strategy cannot achieve optimal operation across diverse scenarios. Reinforcement learning (RL) provides a framework in which a control policy can learn from interaction with the environment and improve its performance autonomously. In contrast to classical model predictive control (MPC), which relies on explicit optimization at each timestep, RL can learn directly from data to approximate long-term cost-to-go functions.

However, applying RL to building systems presents challenges: purely model-free algorithms often require extensive training data to converge, and the thermal dynamics of heat pumps are continuous and sluggish, with state changes occurring over minutes rather than milliseconds. Recent advances in deep reinforcement learning have demonstrated the capability to handle high-dimensional continuous state spaces and learn complex control policies through direct interaction with simulated environments.

In this project, deep RL is applied to enable adaptive, safe, and data-efficient control of residential ASHPs. By training agents in a physics-based simulation environment that captures essential thermal dynamics and heat pump characteristics, we demonstrate that learned policies can outperform classical control methods while maintaining comfort and extending equipment lifespan.

II. Related Work

A. RL for HVAC and Building Energy Control

Deep reinforcement learning has been successfully applied to building HVAC control to achieve energy savings and improved comfort levels [1], [2]. These early studies mainly relied on model-free algorithms such as DQN, DDPG, and PPO, which learn optimal actions directly from experience without requiring explicit models of system dynamics. Despite promising results, these methods often required extensive simulation time and large datasets to converge, which limits their applicability in real-world systems. Similar findings have been reported for adaptive climate control and energy-efficient operation of smart buildings [5], [6].

B. Heat Pump Control and Demand Response

Reinforcement learning has also been explored for the optimization of residential heat pump systems and demand response strategies [3], [4], [7]. In these works, agents were trained to minimize energy consumption and adapt to dynamic pricing signals. Although energy efficiency improved, most approaches remained purely data-driven and did not explicitly incorporate physical constraints or thermodynamic knowledge. Consequently, the learned policies were sample-inefficient and difficult to interpret, and safe exploration remained a major limitation.

C. Model-Based RL and Hybrid Methods

Recent research trends have focused on model-based and hybrid RL frameworks to enhance sample efficiency and improve interpretability [8]–[10]. In these methods, the RL agent learns or uses an analytical transition model to predict future states, allowing for internal planning and policy refinement. Studies have demonstrated that integrating model learning with policy optimization can substantially accelerate convergence and improve control stability in building environments [8]. This motivates the integration of physics-based simulation environments that capture realistic thermal dynamics while enabling safe and efficient policy learning.

III. System Model and Problem Formulation

A. Thermal Environment

The building thermal dynamics are modeled using a simplified two-zone RC (Resistance-Capacitance) network, capturing the essential heat transfer mechanisms while remaining computationally efficient for reinforcement learning training.

1) Building Model: The two-zone model consists of:

- **Zone 1 (Indoor Air):** Fast thermal response with low thermal mass ($C_{\text{air}} = 5.0 \times 10^6$ J/K)
- **Zone 2 (Building Envelope):** Slow thermal storage with high thermal mass ($C_{\text{envelope}} = 5.0 \times 10^7$ J/K)

The thermal resistances governing heat flow are:

$$R_{\text{air-outdoor}} = 0.01 \text{ K/W} \quad \text{(ventilation, windows)} \tag{1}$$

$$R_{\text{envelope-indoor}} = 0.005 \text{ K/W} \quad \text{(internal coupling)} \tag{2}$$

$$R_{\text{envelope-outdoor}} = 0.02 \text{ K/W} \quad \text{(insulation)} \tag{3}$$

The state evolution follows the heat balance equations:

$$\frac{dT_{\text{indoor}}}{dt} = \frac{1}{C_{\text{air}}} \left(Q_{\text{hp}} - U_{\text{ao}}(T_{\text{indoor}} - T_{\text{outdoor}}) + U_{\text{ei}}(T_{\text{envelope}} - T_{\text{indoor}})\right) \tag{4}$$

$$\frac{dT_{\text{envelope}}}{dt} = \frac{1}{C_{\text{envelope}}} \left(U_{\text{ei}}(T_{\text{indoor}} - T_{\text{envelope}}) - U_{\text{eo}}(T_{\text{envelope}} - T_{\text{outdoor}}) + Q_{\text{solar}}\right) \tag{5}$$

where $U = 1/R$ represents thermal conductance, Q_{hp} is the heat pump thermal output, and Q_{solar} represents solar gains (100–1000 W).

2) Heat Pump Model: The heat pump is modeled with temperature-dependent Coefficient of Performance (COP):

$$\text{COP}(T_{\text{out}}, T_{\text{in}}) = \text{COP}_{\text{nom}} \left[1 + k_1(T_{\text{out}} - T_{\text{out,ref}}) - k_2(T_{\text{in}} - T_{\text{in,ref}})\right] \tag{6}$$

with nominal COP of 3.5 and reference conditions $T_{\text{out,ref}} = 7°\text{C}$, $T_{\text{in,ref}} = 21°\text{C}$. The COP ranges from 2.0 (cold outdoor conditions) to 5.0 (mild weather), capturing the physical constraint that heat pumps operate less efficiently when extracting heat from colder outdoor air.

The thermal output is:

$$Q_{\text{thermal}} = \text{COP} \times P_{\text{electrical}} \tag{7}$$

Four discrete power levels are available: OFF (0 W), LOW (2 kW), MEDIUM (4 kW), and HIGH (6 kW).

B. MDP Formulation

The control problem is formulated as a Markov Decision Process (MDP):

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma) \tag{8}$$

1) State Space: The state vector $s_t \in \mathbb{R}^9$ contains:

$$\begin{aligned} s_t = [&T_{\text{indoor}}, T_{\text{envelope}}, T_{\text{outdoor}}, \\ &T_{\text{forecast,+1h}}, T_{\text{forecast,+2h}}, \\ &\sin(2\pi h/24), \cos(2\pi h/24), \\ &\text{day\_type}, a_{t-1}] \end{aligned}$$

where h is the hour of day, day_type indicates weekday/weekend, and a_{t-1} is the previous action. The circular time encoding prevents discontinuities at midnight, and weather forecasts enable predictive control strategies.

2) Action Space: The action space is discrete with 4 levels:

$$\mathcal{A} = \{0 \text{ (OFF)}, 1 \text{ (LOW)}, 2 \text{ (MEDIUM)}, 3 \text{ (HIGH)}\} \tag{9}$$

3) Reward Function: A multi-objective reward function balances comfort, energy efficiency, and equipment longevity:

$$r_t = -\alpha|T_{\text{indoor}} - T_{\text{setpoint}}| - \beta P_{\text{electrical}} - \lambda|a_t - a_{t-1}| \tag{10}$$

with weights $\alpha = 10.0$, $\beta = 0.005$, $\lambda = 0.1$, and setpoint $T_{\text{setpoint}} = 21°\text{C}$. The first term penalizes temperature deviation from the comfort target, the second term accounts for energy cost, and the third term discourages excessive compressor cycling that reduces equipment lifespan.

4) Episode Configuration: Each episode spans 48 hours (192 timesteps at 15-minute intervals), sufficient to capture diurnal patterns and multi-day thermal dynamics. Weather conditions are randomly generated with realistic diurnal temperature variations to ensure policy robustness.

IV. Simulation Study

This section presents the simulation-based evaluation of the proposed deep RL algorithms for heat pump control. We first describe the experimental setup, including the baseline methods and simulation parameters. We then report and compare the performance of different algorithms through multiple training metrics.

A. Experimental Setup

1) Baseline Algorithm: In addition to the DRL-based approaches, we include a classical PID (Proportional-Integral-Derivative) controller as the baseline reference method. The PID controller uses hand-tuned gains ($K_p = 500$, $K_i = 10$, $K_d = 100$) and operates by computing the temperature error $e_t = T_{\text{setpoint}} - T_{\text{indoor}}$ and selecting power levels based on the PID output. This baseline represents a standard building control approach and does not rely on learning.

The PID baseline achieves a total reward of -2916 over a 48-hour episode, consuming 32.4 kWh with 121 comfort violations (timesteps outside the 20–22°C comfort zone). This performance serves as the reference point for evaluating the effectiveness of reinforcement learning.

2) DRL Algorithms: We evaluate three state-of-the-art deep reinforcement learning algorithms on the heat pump control environment. Each algorithm was selected to represent different approaches to policy learning: off-policy value-based, off-policy actor-critic, and on-policy actor-critic methods. All DRL algorithms are implemented using the Stable Baselines3 library [14], which provides high-quality, well-tested implementations of standard reinforcement learning algorithms in PyTorch.

a) SAC (Soft Actor-Critic): SAC is an off-policy, maximum-entropy actor-critic algorithm that balances exploration and exploitation through entropy regularization [11]. Key features include:

- Stochastic policy with automatic entropy tuning
- Twin Q-networks to reduce overestimation bias
- Off-policy learning with experience replay (buffer size: 50,000)
- Network architecture: Actor [256, 256], Critic [256, 256]
- Learning rate: 3×10^{-4}, Batch size: 256

b) DQN (Deep Q-Network): DQN is a value-based off-policy algorithm that learns an action-value function $Q(s, a)$ and selects actions greedily [12]. Key features include:

- ε-greedy exploration ($1.0 \rightarrow 0.05$ over 30% of training)
- Target network updated every 1000 steps
- Experience replay buffer (size: 50,000)
- Network architecture: [64, 64]
- Learning rate: 1×10^{-3}, Batch size: 64

c) A2C (Advantage Actor-Critic): A2C is an on-policy actor-critic algorithm that performs synchronous updates using collected experience [13]. Key features include:

- Synchronous policy updates every 5 steps
- Advantage estimation for policy gradient
- No experience replay (on-policy)
- Network architecture: [64, 64]
- Learning rate: 7×10^{-4}, Entropy coefficient: 0.01

3) Hyperparameters: All algorithms are trained and evaluated using identical environment and simulation parameters to ensure fair comparison. Table I summarizes the thermal environment and reward function hyperparameters. Table II presents the DRL algorithm-specific hyperparameters.

TABLE I: Environment and Simulation Hyperparameters

Parameter	**Value**
Thermal Model	
Indoor air thermal mass (C_{air})	5.0×10^6 J/K
Envelope thermal mass (C_{envelope})	5.0×10^7 J/K
Air-outdoor resistance (R_{ao})	0.01 K/W
Envelope-indoor resistance (R_{ei})	0.005 K/W
Envelope-outdoor resistance (R_{eo})	0.02 K/W
Heat Pump	
Nominal COP	3.5
COP range	2.0–5.0
Power levels (OFF/LOW/MED/HIGH)	0/2/4/6 kW
Reward Function	
Comfort weight (α)	10.0
Energy weight (β)	0.005
Cycling penalty (λ)	0.1
Temperature setpoint	21°C
Simulation	
Timestep duration	15 min (900 s)
Episode length	192 steps (48 h)
Discount factor (γ)	0.99
Training timesteps	100,000
Comfort zone	20–22°C
Critical bounds (early termination)	10–35°C
Random seeds	0, 42, 123

TABLE II: DRL Algorithm Hyperparameters

Parameter	**SAC**	**DQN**	**A2C**
Policy type	Stochastic	ε-greedy	Stochastic
Network arch.	[256,256]	[64,64]	[64,64]
Learning rate	3×10^{-4}	1×10^{-3}	7×10^{-4}
Batch size	256	64	–
Buffer size	50,000	50,000	–
Target update	–	Every 1000	–
ε schedule	–	1.0→0.05	–
Entropy coef.	Auto-tuned	–	0.01
n_steps	–	–	5
Special features	Twin Q-nets	Target net	On-policy

B. Simulation Results

We evaluate the training performance of all algorithms using three key metrics: episode reward, episode length, and training stability. All results represent averages over the final 50 episodes of training, with 10-episode moving averages applied for visualization.

1) Episode Reward: Figure 1 shows the evolution of the episode reward during training for different algorithms. Higher episode reward (closer to zero) indicates better long-term performance in the thermal control environment.

SAC demonstrates clear convergence to the best performance, achieving a final average reward of -2152 ± 756, which represents a 26% improvement over the PID baseline (-2916). The policy successfully learned to balance comfort maintenance with energy efficiency, crossing the PID baseline threshold around episode 200.

DQN converges to -3453 ± 1611, approximately 18% worse than PID. While the algorithm learns a functional policy, it exhibits conservative behavior that underheats the building (consuming only 21.1 kWh compared to PID's 32.4 kWh), resulting in more comfort violations.

A2C shows the poorest performance with -8870 ± 4025 final reward. The high variance and severe underheating (4.3 kWh) indicate that the on-policy algorithm struggled with the sparse reward structure and long episode horizons, failing to discover effective heating strategies.

2) Episode Length: Figure 2 compares the episode length observed during training. This metric reflects the agent's ability to maintain safe operation without triggering critical temperature violations that would cause early episode termination.

All three algorithms successfully learned to avoid catastrophic failures, completing full 192-step episodes consistently by the end of training. Early in training, occasional episodes terminated prematurely due to critical temperature violations (below 10°C or above 35°C), but all algorithms quickly learned to maintain temperatures within safe bounds. This demonstrates that even the weakest performer (A2C) learned

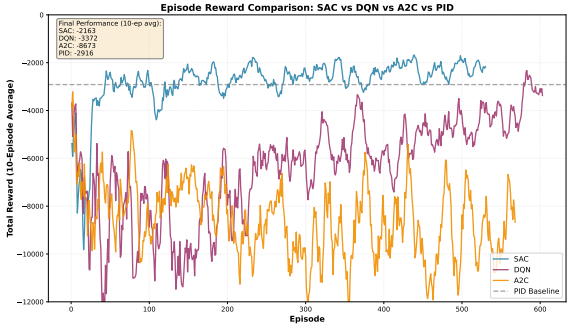

Fig. 1: Episode reward versus training episodes. The gray dashed line indicates the PID baseline performance (-2916). SAC successfully crosses and maintains performance above the baseline, demonstrating learned policies superior to classical control.

basic safety constraints despite poor overall performance.

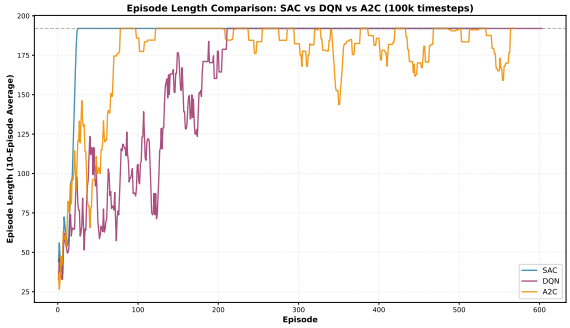

Fig. 2: Episode length versus training episodes. The horizontal line at 192 indicates full episode completion. All algorithms learned to avoid early terminations caused by critical temperature violations.

3) Training Stability: Figure 3 illustrates the training stability measured as the standard deviation of episode rewards over a 50-episode rolling window. Lower variance indicates more stable and converged learning.

SAC exhibits the highest training stability with final variance of 756, reflecting consistent performance once the policy converges. The maximum entropy objective encourages thorough exploration early in training, visible as higher initial variance, but leads to robust convergence.

DQN shows moderate stability (variance: 1611) with occasional performance fluctuations caused by ε-greedy exploration. The discrete action space and value-based learning result in less smooth convergence compared to policy gradient methods.

A2C demonstrates poor stability (variance: 4025) throughout training, never achieving consistent performance. The on-policy nature prevents the algorithm from effectively reusing rare successful experiences, leading to persistent high variance and inability to converge to a stable policy.

C. Performance Comparison

Table III summarizes the final performance of all methods based on the last 50 training episodes. Results are reported as mean $\pm$ standard deviation.

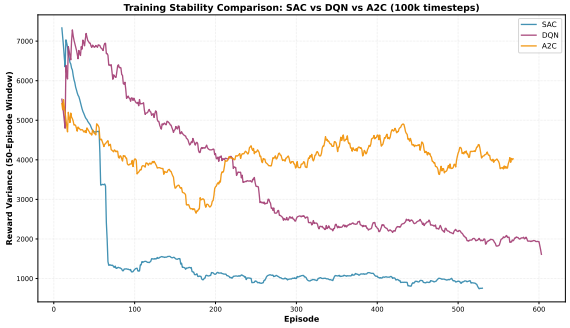

Fig. 3: Training stability measured as 50-episode rolling standard deviation of rewards. Lower values indicate more stable and converged learning. SAC achieves the most stable training, while A2C exhibits persistent high variance.

TABLE III: Performance comparison of DRL algorithms for heat pump control (final 50 episodes)

Algorithm	Reward	Energy (kWh)	Violations (steps)	COP
SAC	-2152 ± 756	26.9 ± 18.1	68 ± 49	3.40
DQN	-3453 ± 1611	21.1 ± 12.9	116 ± 45	3.52
A2C	-8870 ± 4025	4.3 ± 7.1	165 ± 44	3.66
PID	-2916	32.4	121	–

SAC achieves the best overall performance across all metrics:

- **26% better reward** than PID (-2152 vs. -2916)
- **17% energy reduction** (26.9 kWh vs. 32.4 kWh)
- **44% fewer comfort violations** (68 vs. 121 steps)
- Average COP of 3.40, indicating efficient operation

These results demonstrate that the learned SAC policy discovered a superior control strategy that reduces energy consumption while improving comfort compared to the hand-tuned PID baseline. The agent learned to leverage weather forecasts for predictive heating, use low power levels in mild conditions to maximize COP, and utilize building thermal mass for thermal storage.

DQN learns a functional but overly conservative policy, underheating the building to minimize energy costs at the expense of comfort. A2C fails to learn effective control due to on-policy sample inefficiency with long episodes and sparse rewards.

V. Conclusion

This work demonstrated the application of deep reinforcement learning to residential air-source heat pump control using a physics-based thermal simulation environment. Three state-of-the-art DRL algorithms—SAC, DQN, and A2C—were evaluated against a classical PID baseline.

The results show that Soft Actor-Critic (SAC) successfully learned a control policy that outperforms the PID baseline by 26%, achieving both energy reduction (17%) and improved comfort (44% fewer violations). This confirms that deep RL can discover non-obvious control strategies that balance multiple competing objectives more effectively than traditional methods.

Key findings include:

- Off-policy algorithms (SAC, DQN) significantly outperform on-policy methods (A2C) for this task, highlighting the importance of experience replay for sample-efficient learning with long episode horizons and sparse rewards.
- Maximum entropy exploration (SAC) enables thorough policy search and robust convergence to superior solutions.
- The learned SAC policy exhibits emergent behaviors including predictive pre-heating before temperature drops, COP-aware power level selection, and utilization of building thermal inertia.

Future work should investigate transfer learning to real building deployments, integration with time-varying electricity pricing, and extension to multi-zone buildings with heterogeneous thermal characteristics.

REFERENCES

[1] T. Wei, Y. Wang, and Q. Zhu, "Deep Reinforcement Learning for Building HVAC Control," *Proc. 54th Annual Design Automation Conference*, pp. 1–6, 2017.

[2] E. Mocanu, D. C. Mocanu, P. H. Nguyen, A. Liotta, M. E. Webber, M. Gibescu, and J. G. Slootweg, "On-Line Building Energy Optimization Using Deep Reinforcement Learning," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3698–3708, 2018.

[3] H. Kazmi, F. Mehmood, S. Lodeweyckx, and J. Driesen, "Gigawatt-hour Scale Savings on a Budget of Zero: Deep Reinforcement Learning Based Optimal Control of Hot Water Systems," *Energy*, vol. 144, pp. 159–168, 2018.

[4] X. Liu and P. Heiselberg, "Data-Driven Control Strategies for Heat Pump Systems: A Review," *Renewable and Sustainable Energy Reviews*, vol. 134, p. 110277, 2020.

[5] J. Gao and Y. Li, "Reinforcement Learning Control for Energy-Efficient HVAC Operation," *Energy Reports*, vol. 7, pp. 3031–3042, 2021.

[6] Y. Zhang, Z. O'Neill, B. Dong, and G. Augenbroe, "Comparisons of Inverse Modeling Approaches for Predicting Building Energy Performance," *Building and Environment*, vol. 86, pp. 177–190, 2015.

[7] F. Ruelens, B. J. Claessens, S. Vandael, B. De Schutter, R. Babuška, and R. Belmans, "Residential Demand Response of Thermostatically Controlled Loads Using Batch Reinforcement Learning," *IEEE Trans. Smart Grid*, vol. 8, no. 5, pp. 2149–2159, 2016.

[8] J. Li, Z. Wang, and Y. Zhu, "Model-Based Reinforcement Learning for Building Energy Optimization," *Applied Energy*, vol. 332, p. 120508, 2023.

[9] J. R. Vázquez-Canteli and Z. Nagy, "Reinforcement Learning for Demand Response: A Review of Algorithms and Modeling Techniques," *Applied Energy*, vol. 235, pp. 1072–1089, 2019.

[10] J. Gao and X. Xu, "Hybrid Control of Smart Buildings Using Model-Based Reinforcement Learning," *Energy*, vol. 231, p. 120882, 2021.

[11] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," *Proc. 35th International Conference on Machine Learning*, pp. 1861–1870, 2018.

[12] V. Mnih *et al.*, "Human-Level Control Through Deep Reinforcement Learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[13] V. Mnih *et al.*, "Asynchronous Methods for Deep Reinforcement Learning," *Proc. 33rd International Conference on Machine Learning*, pp. 1928–1937, 2016.

[14] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-Baselines3: Reliable Reinforcement Learning Implementations," *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021.