Comparative Study of Deep Reinforcement Learning Algorithms for Residential HVAC Control

Kunal Shankar and Casey Oliver Dettlaff

PhD Student(s) at SCALE Lab,

Energy Systems Innovation Center,

Electrical Engineering and Computer Science,

Washington State University

Machine Learning Project

fall 2025

WASHINGTON STATE UNIVERSITY

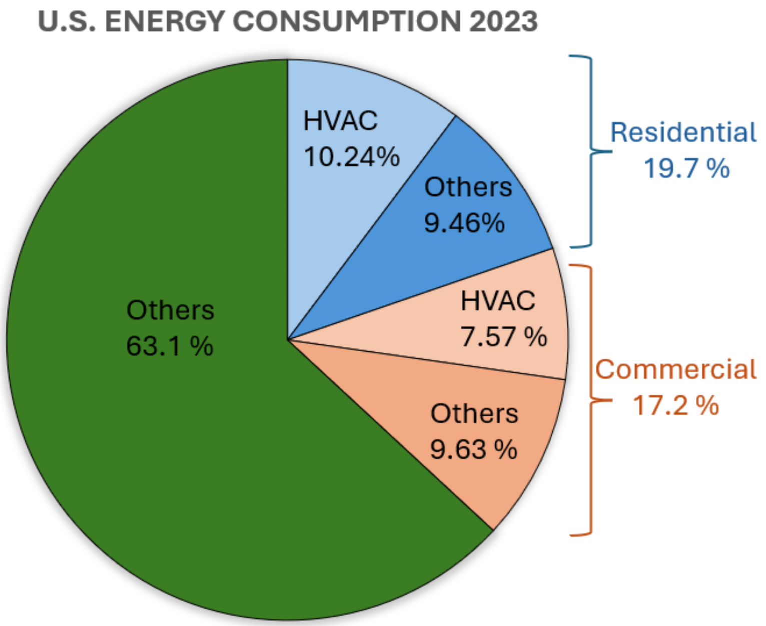
TABLE OF CONTENTS

Washington State University

1. Motivation

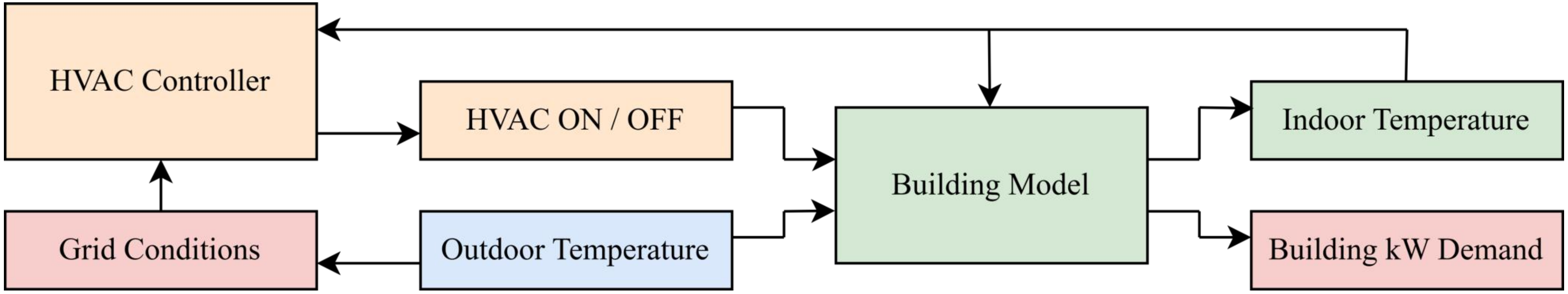
Motivation

- Right now, electricity demand is increasing faster than supply.
- Building new capacity, such as Generators and Transmission Lines, is slow and difficult.
- Energy storage can relieve some stress on the grid, but batteries are still too expensive to deploy at scale.
- Because of their thermal inertia, buildings can serve as energy storage.
- Buildings can adapt their energy consumption to grid needs. This is known as Demand Response (DR).
- Buildings already exist at scale on the grid.
- Buildings can make use of HVAC for DR.



Research Challenge

- Traditional thermostats use basic on/off control at fixed setpoints, functioning only to keep the indoor temperature within a predefined comfort band.
- Many regions are adopting dynamic electricity pricing that changes throughout the day.
- Smart HVAC controller: responds to price signals and shifts energy usage accordingly.
- Model Predictive Control: requires accurate system models, computationally heavy.
- To address these challenges, we use Reinforcement Learning to learn an effective control policy.

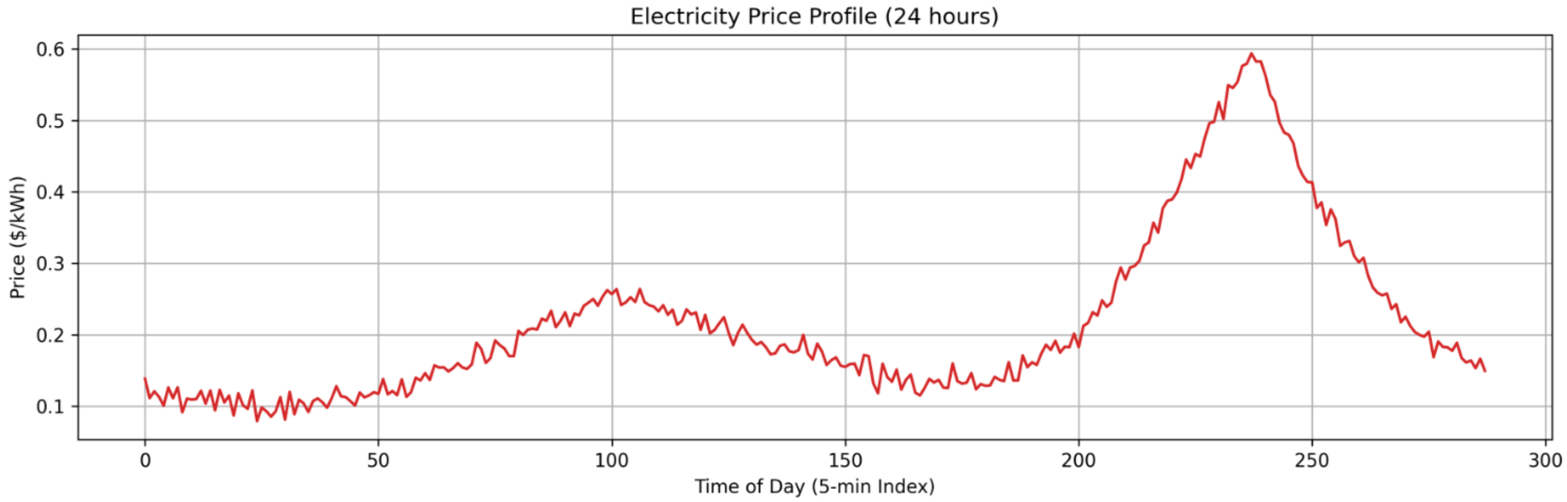


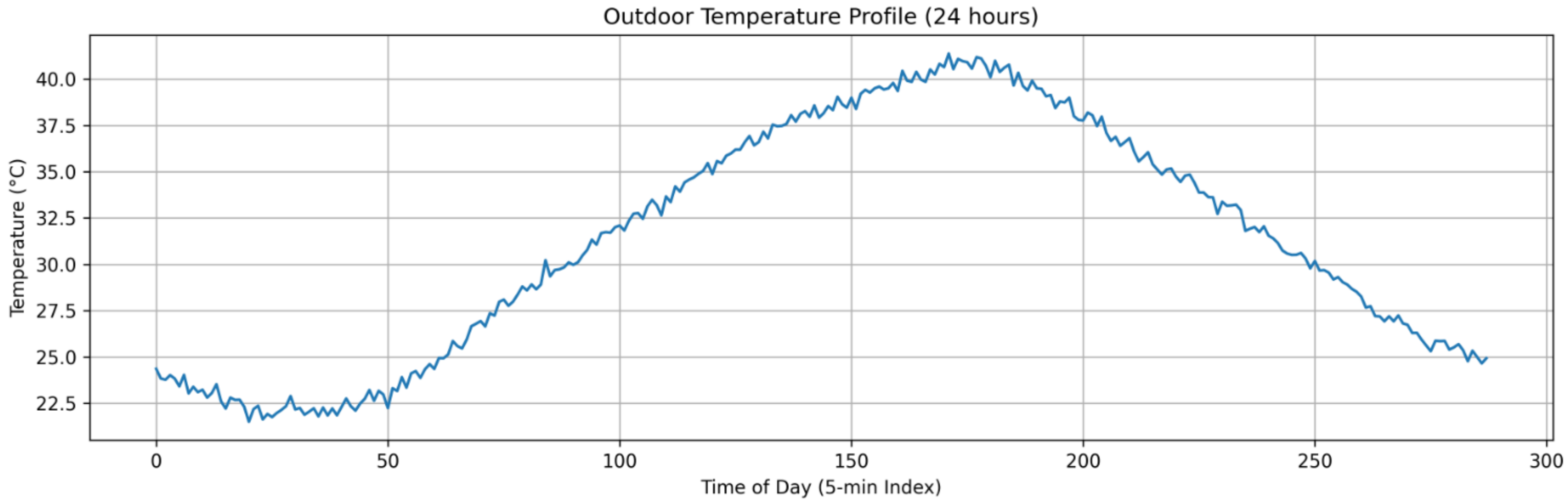
Why Deep Reinforcement Learning?

- Reinforcement Learning acts as a natural framework for sequential decision-making.
- The agent discovers optimal HVAC policies through interaction with the building environment.
- Handles multi-objective tradeoffs between cost, comfort, and equipment wear.
- Leverages deep RL with neural networks to scale to continuous state spaces.
- To evaluate different RL algorithms, we need computationally efficient models and a Gymnasium for a standardized environment.

2. Setup

Generated Data



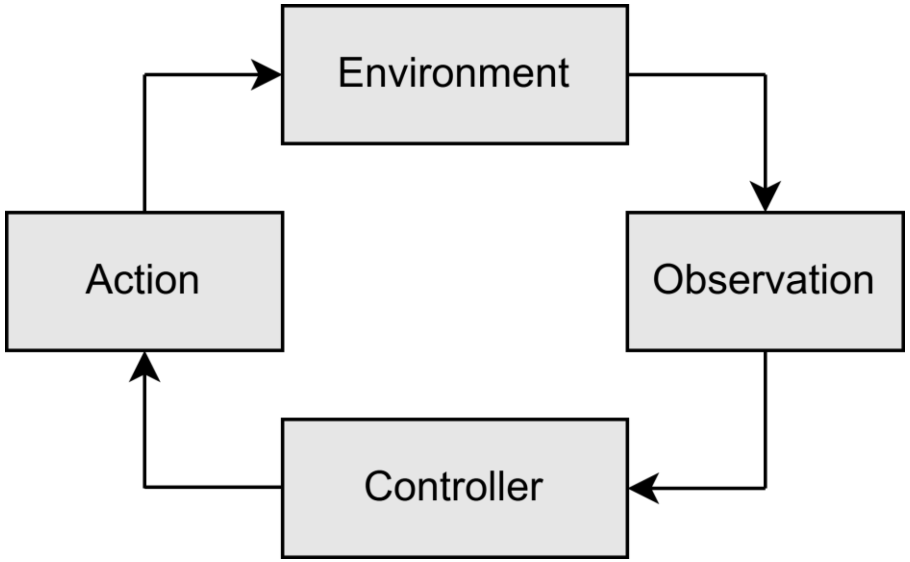


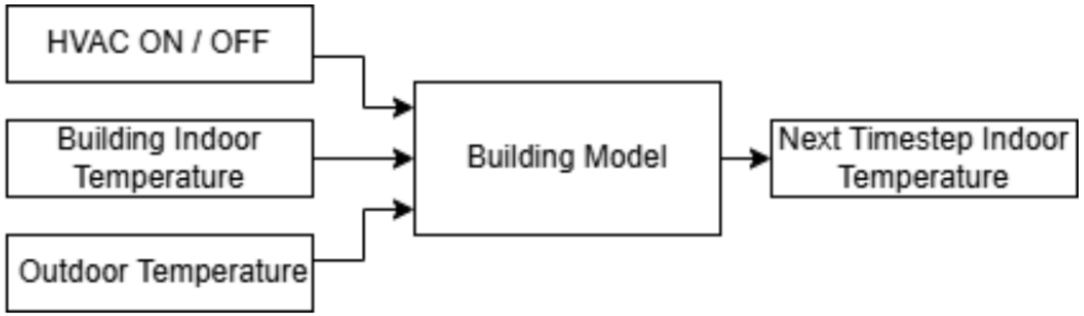
WASHINGTON STATE UNIVERSITY

3. Environment

Gymnasium Environment Design

- Gymnasium provides agent-environment interaction loop for RL training.
- Agent observes indoor temperature, outdoor temperature, electricity price, and time of day.
- Agent controls HVAC with binary actions: turn ON or turn OFF.
- Each episode runs for 24 hours at 5-minute intervals, totaling 288 timesteps.
- The building model is implemented as part of the environment.





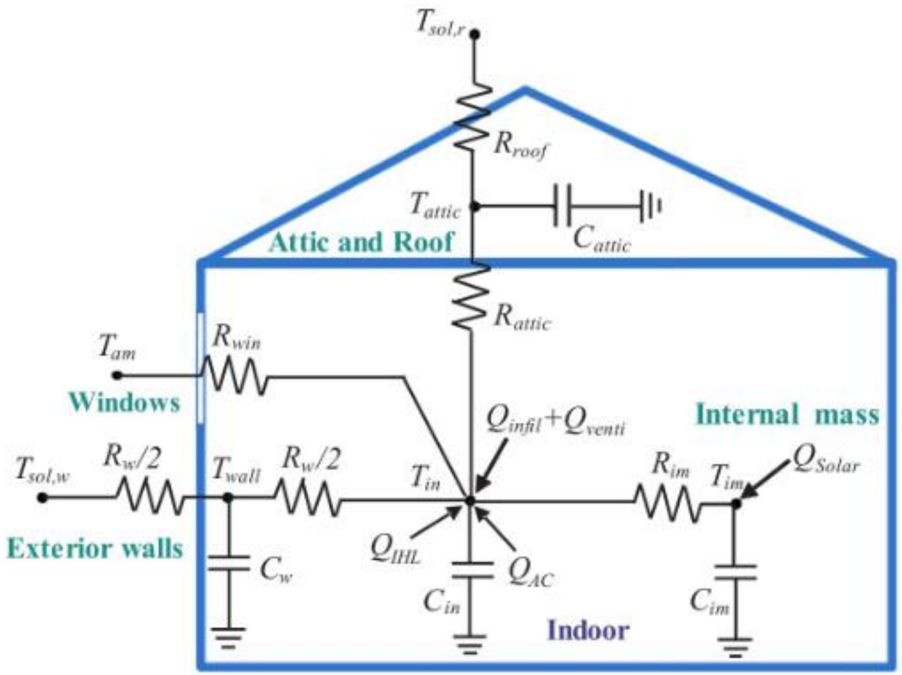
WASHINGTON STATE UNIVERSITY

RC Modelling

- We model Building Internal Dynamics by the principle of “Heat-flow”.
- Resistors represent resistance to heat flow.
- Capacitors represent the building capacity to store thermal energy.

$$C_z \frac{dT_z}{dt} = \sum_{i=1}^{N} \frac{T_{w_i} - T_z}{R_{zw_i}} + \frac{T_a - T_z}{R_{za}} + A_z Q_{HVAC} + B_z Q_{Int} + D_z Q_{Solar},$$

$$C_{w_i} \frac{dT_{w_i}}{dt} = \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{T_{w_j} - T_{w_i}}{R_{w_{ij}}} + \frac{T_z - T_{w_i}}{R_{zw_i}} + \frac{T_{am} - T_{w_i}}{R_{wa_i}} + B_{w_i} Q_{Int} + D_{w_i} Q_{Solar}.$$



$$\dot{\underline{x}} = A(\underline{\theta})\underline{x} + B(\underline{\theta})\underline{u} + D(\underline{\theta})\underline{w}$$
$$y = C\underline{x}.$$

4. Algorithms

RL– Algorithms

DQN

- Learns Q-values: How good each action is in a given state
- Uses replay buffer to break data correlation, improving stability.
- Explores with ε-greedy strategy: Random actions early, optimal actions later.
- Ideal for discrete action spaces like binary HVAC control

PPO

- Directly learns a policy: which actions to take.
- Uses clipped objective to prevent overly large policy updates.
- Collects experience batches and performs multiple training passes.
- Known for stability and robustness in control problems.

SAC

- Actor-critic method: learns both policy and value function.
- Encourages exploration through entropy regularization.
- Stays uncertain early in training to avoid local optima.
- Typically used for continuous control.

Washington State University

Hyperparameter tuning

DQN

1. Learning Rate: 0.001
2. Discount Factor: 0.98
3. Better Buffer Size: 100,000
4. Exploration Fraction: 0.05

PPO

1. Learning Rate: 5e-5
2. Discount Factor: 0.95
3. Batch Size: 256
4. Rollout Horizon: 256

SAC

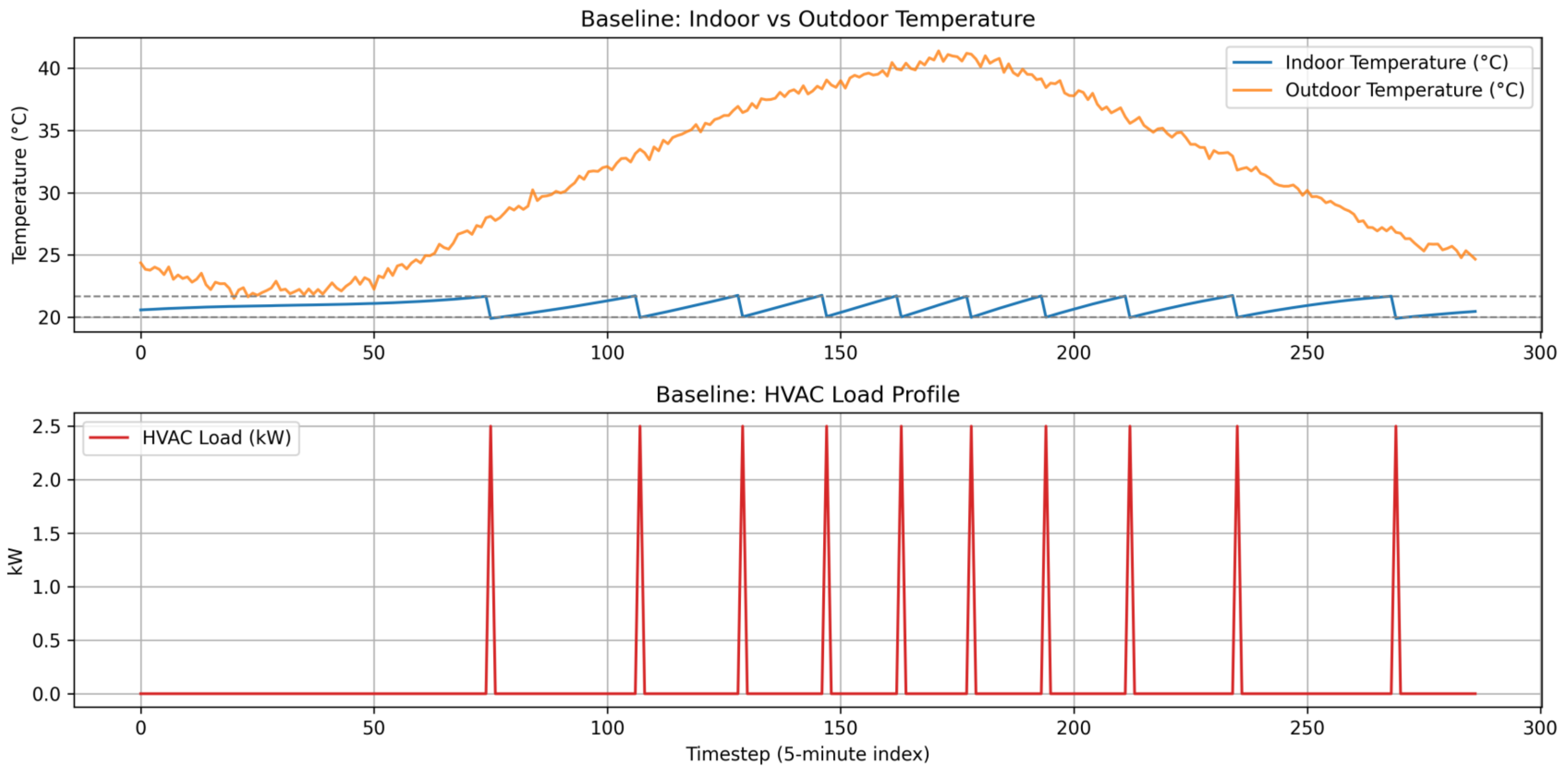
1. Learning Rate: 3e-4
2. Discount Factor: 0.98
3. Batch Size: 128
4. Polyak smoothing coefficient: 0.01
5. Entropy coefficient setting: auto

- Parameter Optimization was performed for each algorithm.
- The models were trained for 50,000 timesteps with different hyperparameters.
- We chose the hyperparameters resulting in the lowest loss.

6. Results

RESULTS

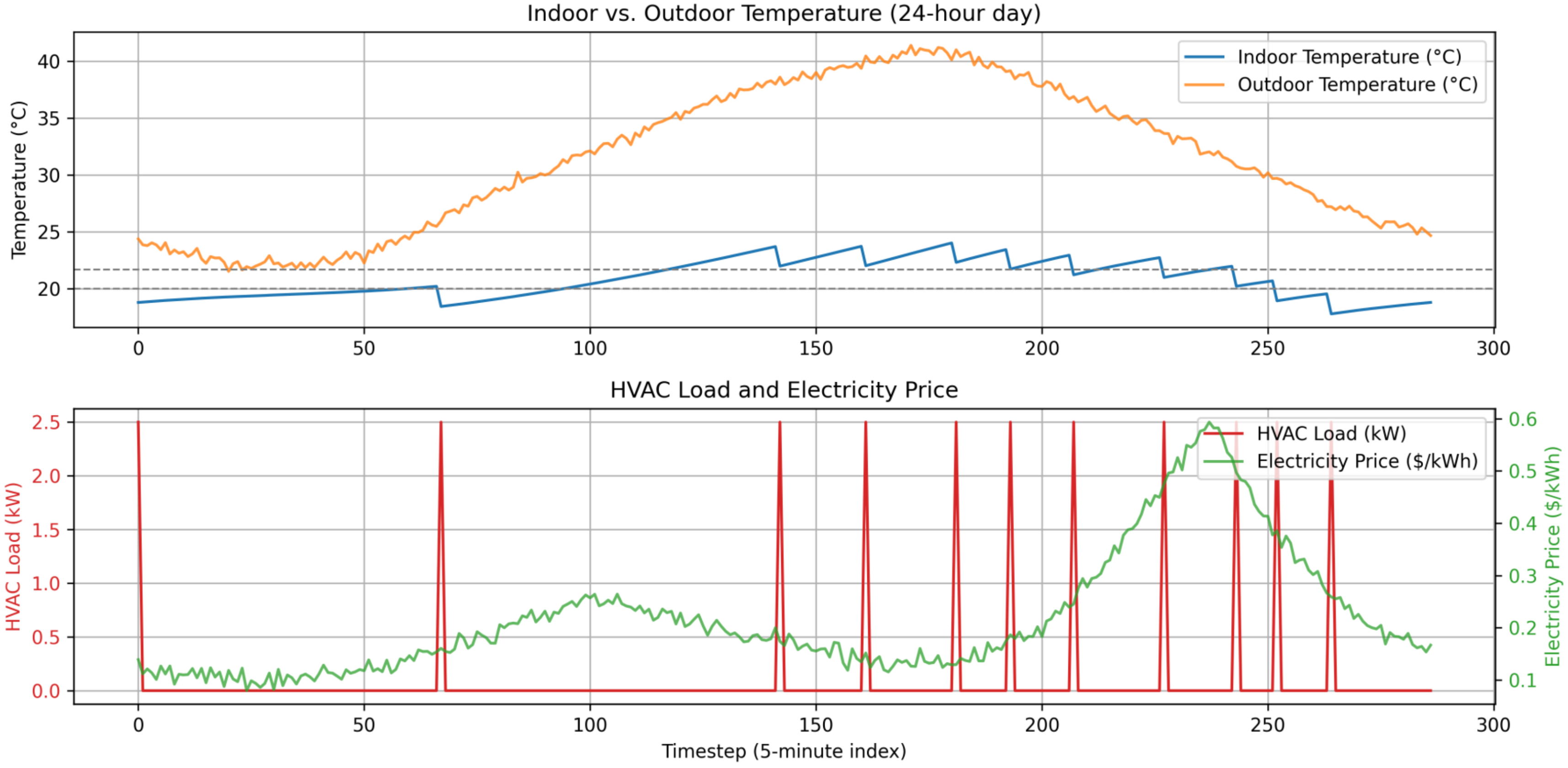
Baseline Controller



WASHINGTON STATE UNIVERSITY

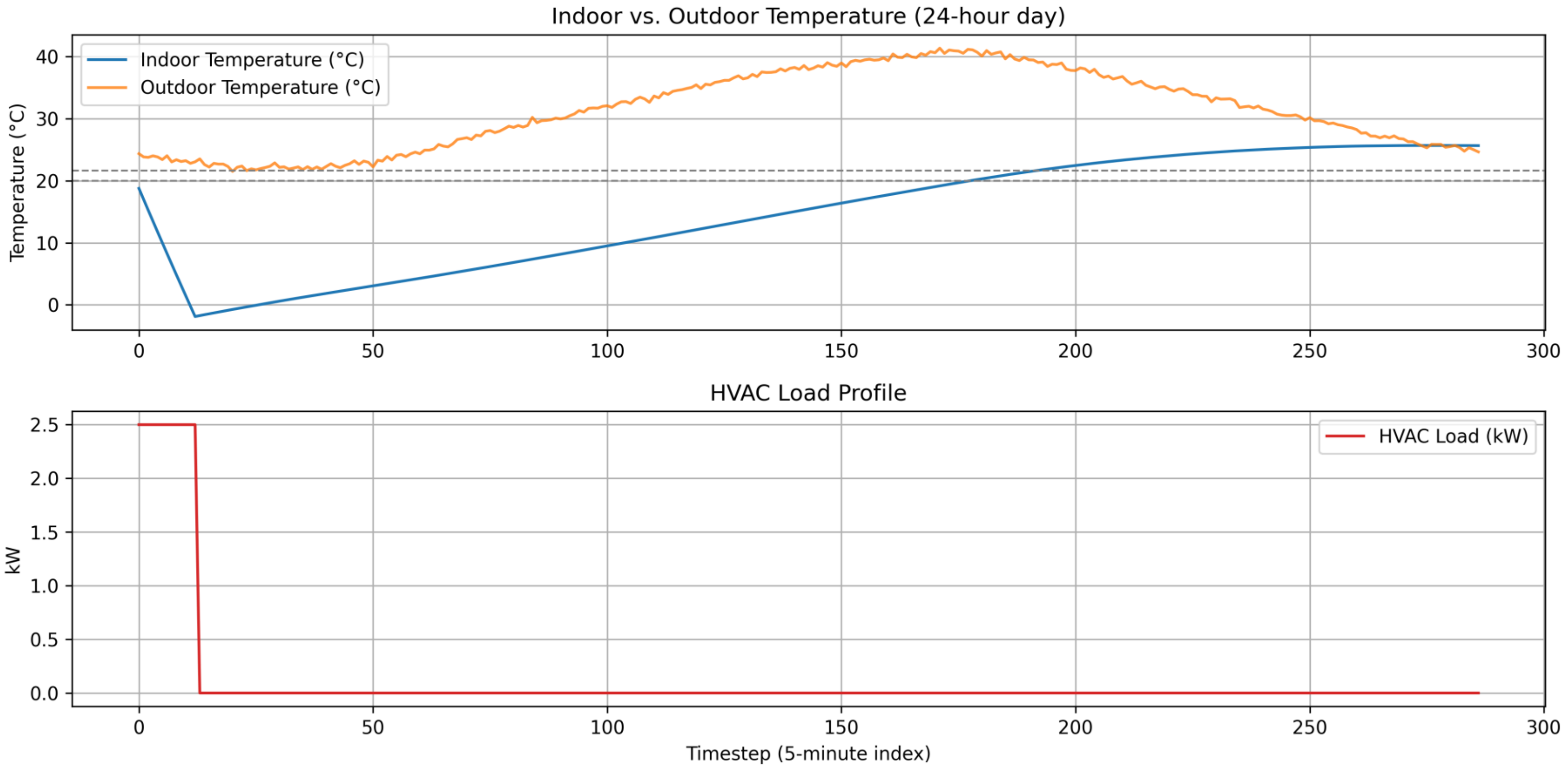
RESULTS

DQN Controller



RESULTS

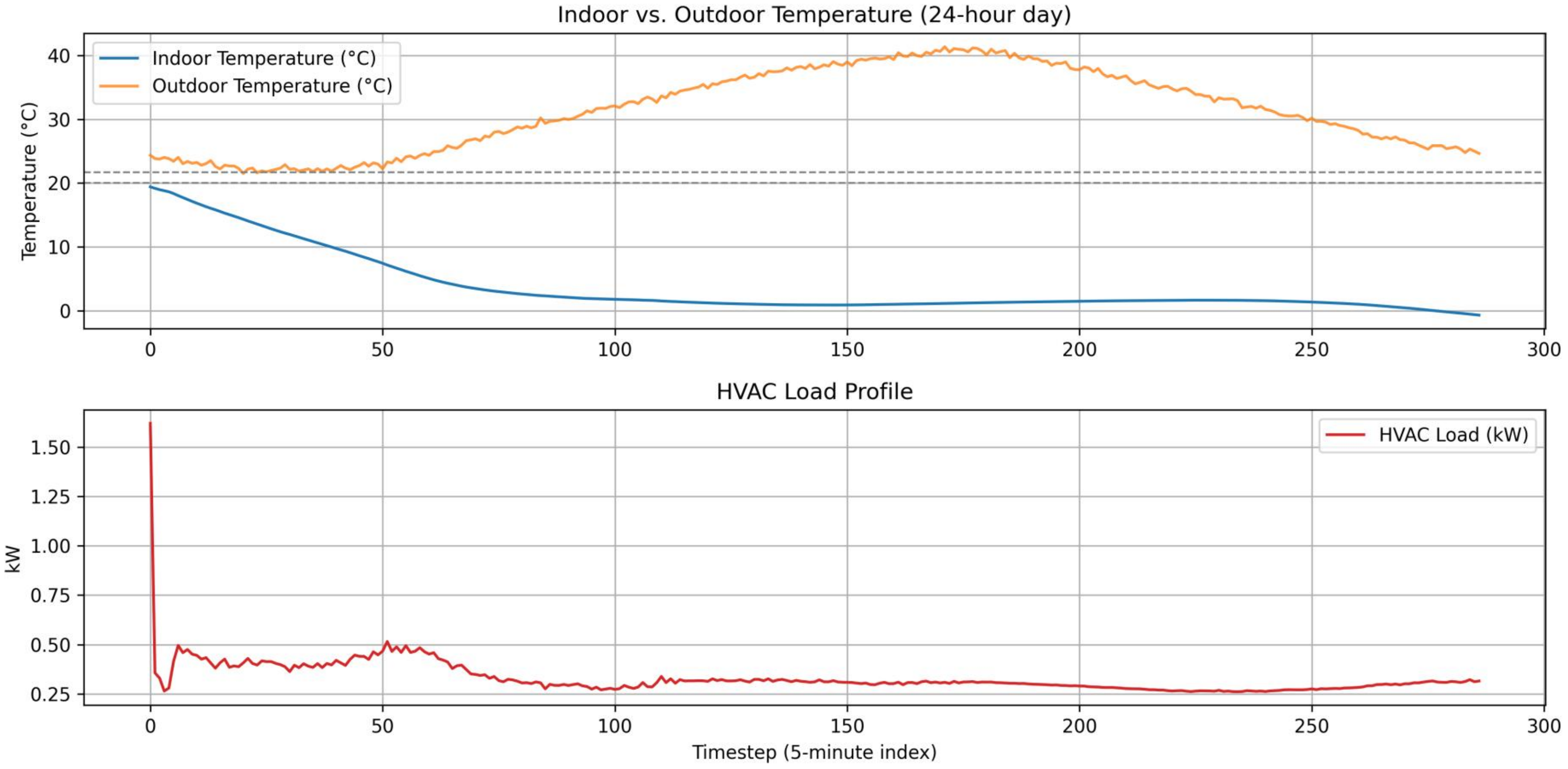
PPO Controller



WASHINGTON STATE UNIVERSITY

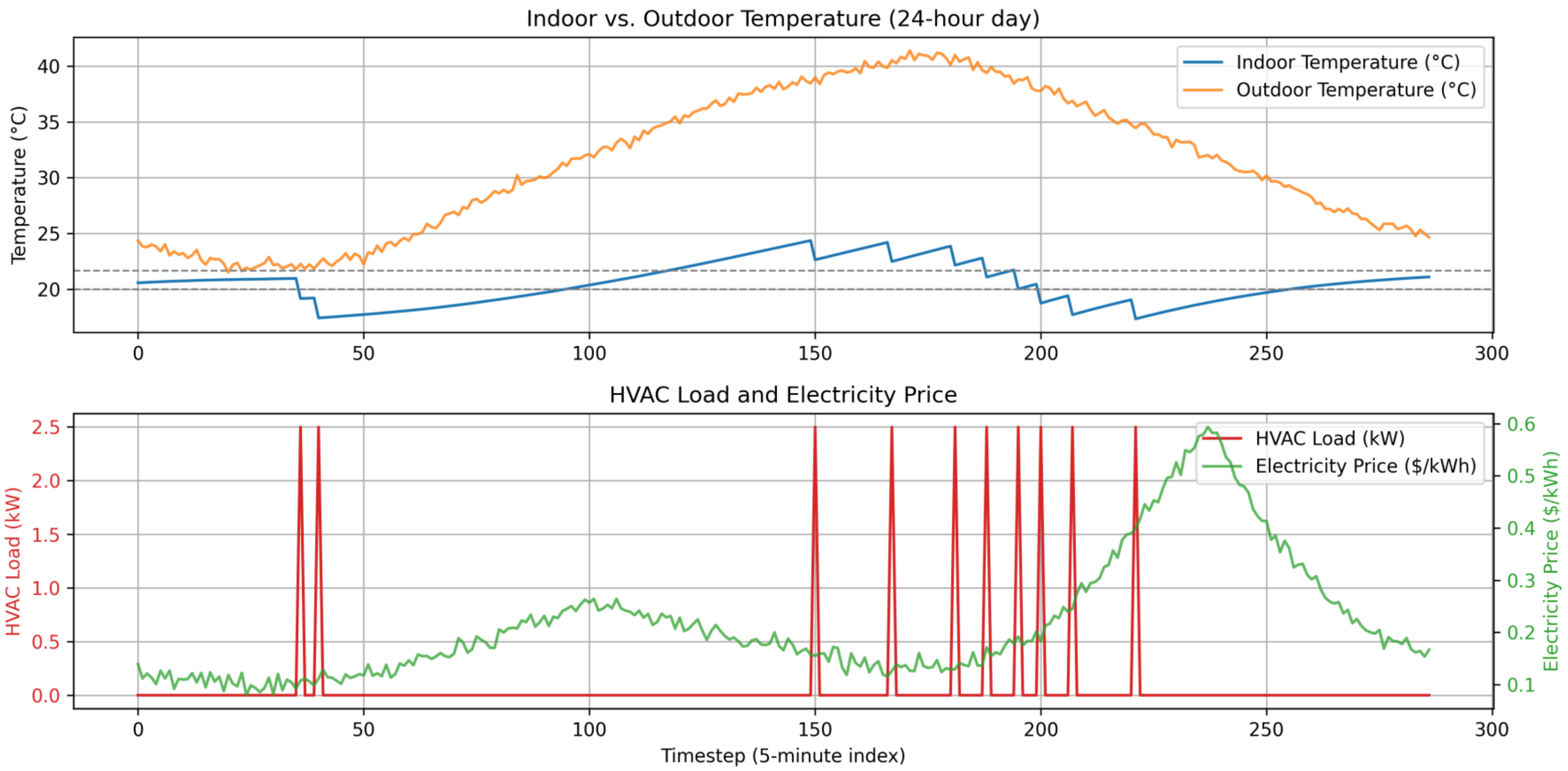
RESULTS

SAC Controller



RESULTS

Cost-Aware DQN Controller



WASHINGTON STATE UNIVERSITY

6. Conclusion and Future Work

Conclusion

- DQN successfully learned duty-cycle HVAC control after 500,000 timesteps with optimized hyperparameters.
- DQN outperformed PPO and SAC because it naturally matches discrete ON/OFF actions and handles low-dimensional state spaces.
- DQN learned to pre-cool buildings before price peaks, demonstrating cost-aware load shifting.
- Deep RL is feasible for residential HVAC control and opens new paths for grid-interactive buildings.
- Future Work: Extend to multi-day training, increased environmental variability, and refined reward tuning for improved comfort reliability.

THANK YOU