CORE-EMS: A Reinforcement Learning Approach for Intelligent Building Energy Management

Abdellatif Zeghib, Ahmad Arif, Jawad Almuttawa, Nawaf Almalki, Yousef Jamal-Allail

August 2025

Abstract

CORE-EMS is a reinforcement-learning framework for smart building energy management. We deploy a centralized Soft Actor–Critic (SAC) controller in CityLearn to issue continuous actions for cooling equipment and thermal storage across multiple Saudi climates. The reward balances three operational objectives—thermal comfort, smooth operation (low ramping), and reliability under outages—while the environment enforces device limits and grid metrics. We benchmark against a rule-based controller (RBC) and a practical model predictive control (MPC) baseline. Results show that SAC achieves baseline-level grid performance while substantially improving comfort and smoothness, demonstrating a viable, data-driven pathway for practical building control under flat tariffs.

1 Introduction

The building sector in Saudi Arabia accounts for approximately 30% of the total energy consumption in the Kingdom [1], with the majority driven by cooling demands due to the hot climate. These high levels of energy consumption increase operational costs for households, and place a significant burden on the national power grid, especially during peak hours in summer. Therefore, improving the efficiency of building energy management is a critical priority in achieving the Kingdom's sustainability goals, as outlined in Vision 2030.

Building Energy Management Systems (BEMS) have emerged as a key technology for addressing these challenges. It works by monitoring and controlling devices such as HVAC systems, lighting, and energy storage units (ESUs). BEMS aim to improve energy efficiency, reduce operational costs, and maintain acceptable levels of occupant comfort. In addition, modern BEMS are increasingly expected to support grid-interactive features such as load shifting, demand response, and the integration of renewable energy sources, making them central to the transition towards smart and sustainable cities.

Despite these promises, most current BEMS implementations remain relatively simplistic. They are usually based on fixed schedules, rule-based control (RBC), or short-horizon predictive models such as model predictive control (MPC). While these approaches can reduce consumption to some extent, they often fall short when dealing with long-term dependencies, nonlinear dynamics, or the stochastic nature of real-world building environments.

Another important limitation lies in their inability to address multiple, and often conflicting, objectives simultaneously. For example, strategies that minimize operational cost may compromise thermal comfort, while comfort-oriented strategies may increase energy use and carbon emissions. Balancing these trade-offs is critical for ensuring that BEMS do not simply optimize one dimension at the expense of others. This gap highlights the need for more advanced, data-driven approaches capable of capturing system dynamics, learning from historical patterns, and coordinating multiple objectives in real time.

Recent advancements in Reinforcement Learning (RL) have been applied to building energy manage-

ment, where agents learn control policies through interaction with the environment. RL agents can learn the control policy, and manage the devices of the building, by interacting with the environment.

In this paper, we introduce CORE-EMS, a Reinforcement Learning framework for intelligent building energy management. CORE-EMS incorporates four key dimensions; Comfort, Operational cost, Resilience, and Energy Intelligence, into the reward function that guides the agent's behavior. The framework leverages a Transformer-based dynamics model for accurate prediction of building thermal behavior, coupled with a reinforcement learning agent for adaptive decision-making. To benchmark performance, we compare CORE-EMS with rule-based and model predictive control baselines on simulated Saudi residential buildings.

2 Literature Review

2.1 Rule-Based Control (RBC)

Early approaches to building energy management have relied heavily on rule-based control (RBC), where predefined heuristics or simple if-then rules are used to regulate heating, ventilation, and air-conditioning (HVAC) systems, lighting, and ESUs. The main advantage of RBC is its simplicity: rules are easy to design, require minimal computational resources, and can be implemented without detailed system models. However, RBC strategies are inherently rigid and unable to adapt to dynamic operating conditions such as variations in occupancy, weather, or electricity prices. As a result, while RBC can achieve modest reductions in energy use, its effectiveness is limited in modern smart building applications.

2.2 Model Predictive Control (MPC)

Model predictive control (MPC) has emerged as a more advanced strategy for building energy management. MPC optimizes control actions by solving a constrained optimization problem over a finite prediction horizon, using forecasts of disturbances such as weather and occupancy. Its main strengths are the ability to incorporate system constraints explicitly and to leverage forecasts for proactive decision-making. However, practical applications of MPC in buildings face several challenges[2]. First, developing accurate physics-based building models is time-consuming and requires domain expertise. Second, solving optimization problems in real time can be computationally intensive, particularly for large-scale systems with nonlinear dynamics. Finally, while MPC can in theory handle multiple objectives by formulating a weighted-sum optimization problem, tuning such weights is non-trivial and often results in oversimplified trade-offs between comfort, cost, and energy efficiency [3]. These limitations hinder the scalability and flexibility of MPC for real-world deployment.

2.3 Reinforcement Learning (RL) in Building Control

Reinforcement learning (RL) has recently gained attention as a promising alternative for building energy management. Unlike MPC, which requires explicit system models, RL agents learn control policies through trial-and-error interaction with the environment [4]. This makes RL particularly well-suited for environments with nonlinear dynamics, uncertainty, and stochastic variations in demand and weather. RL methods have been applied to tasks such as HVAC control, energy storage scheduling, and demand response. They offer several advantages, including the ability to capture long-term dependencies and adapt to changing conditions. Nonetheless, RL faces its own challenges: training is data-intensive, ensuring safety and comfort constraints is difficult during exploration, and interpretability of learned policies remains limited. Recent advances, such as model-based RL and the integration of deep neural networks, have improved sample efficiency and stability, but the field remains under active research.

2.4 Multi-Objective Optimization in BEMS

A critical challenge in building energy management lies in balancing multiple, often conflicting, objec-

tives. For example, minimizing energy cost may come at the expense of occupant comfort, while strategies that prioritize comfort can increase consumption and carbon emissions. Traditional RBC and MPC approaches typically reduce these objectives to a single weighted-sum formulation, which oversimplifies the inherent trade-offs. Recent studies have explored multi-objective reinforcement learning (MORL) to address this issue, enabling agents to learn policies that explicitly account for competing objectives and to approximate Pareto-optimal trade-offs [5]. However, applications of MORL in building energy management are still limited, and more research is needed to translate theoretical advances into practical control frameworks.

2.5 Regional Studies and Saudi Context

Most of the existing literature on BEMS originates from Europe, North America, and East Asia, where energy consumption patterns and climatic conditions differ significantly from those in Saudi Arabia. In the Saudi context, residential cooling demand is the dominant driver of electricity consumption, accounting for a substantial share of peak load during summer [1]. Despite this, relatively few studies have investigated advanced building control methods tailored to the Kingdom's unique climatic, cultural, and infrastructural conditions. The scarcity of region-specific research creates a gap in understanding how advanced frameworks such as RL-based BEMS can be adapted to Saudi buildings. Addressing this gap is critical for supporting the Kingdom's sustainability targets under Vision 2030.

While RBC and MPC remain the most widely adopted approaches for BEMS, they suffer from rigidity, high modeling effort, and limited adaptability. Data-driven and RL-based approaches offer promising alternatives, with RL in particular showing potential for adaptive, long-term, multi-objective decision-making. However, their application in the Saudi context remains underexplored. This motivates the development of frameworks such as CORE-EMS, which leverage reinforcement learning to incorporate comfort, operational cost, resilience, and energy intelligence into a unified building energy management strategy.

3 Methodology

This section describes the problem formulation, environment and dataset, reward design, agent architecture, and the training/evaluation protocol used in CORE-EMS. The implementation follows the public code in our project repository (agents, training scripts, and evaluation utilities).

3.1 Problem Formulation

We model building energy management as a Markov Decision Process (MDP). At each hourly time step t the agent observes a state $\mathbf{s}_t$, applies a continuous action $\mathbf{a}_t$, receives a scalar reward r_t, and the CityLearn environment advances to the next state $\mathbf{s}_{t+1}$.

- **State $\mathbf{s}_t$**: concatenation of per-building observations such as month/hour/day-type, indoor and outdoor temperatures, solar irradiance, cooling demand, electricity price, and storage state-of-charge.
- **Action $\mathbf{a}_t$**: continuous control of *cooling device* intensity ($[0, 1]$), *cooling storage* charge/discharge, and *DHW storage* charge/discharge (bounded by device limits). In our main setup, we use a *centralized* policy that outputs a single action vector covering all buildings.
- **Transition**: $\mathbf{s}_{t+1} = f_{\text{env}}(\mathbf{s}_t, \mathbf{a}_t, \boldsymbol{\omega}_t)$ where f_{env} is the CityLearn simulator and $\boldsymbol{\omega}_t$ captures exogenous inputs (weather, non-shiftable loads, occupancy proxies).

The control objective is to maximize the expected discounted return

$$J(\pi) = E_\pi \left[\sum_{t=0}^{T} \gamma^t \, r_t \right], \quad \gamma \in (0, 1]. \tag{1}$$

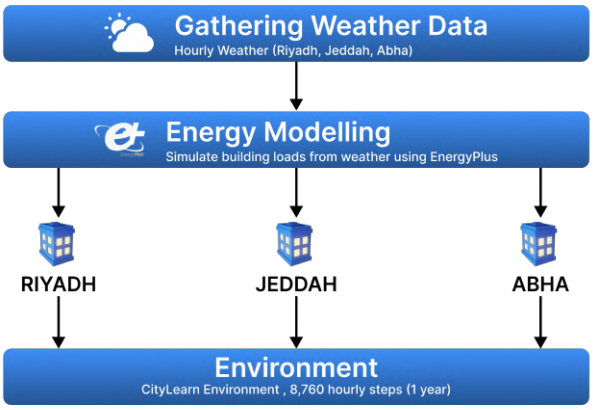


Figure 1: Dataset pipeline: hourly weather (Riyadh, Jeddah, Abha) feeds EnergyPlus to simulate building loads; the resulting traces are used by CityLearn for 8,760 hourly steps (1 year).

3.2 Environment and Dataset

We use the CityLearn environment[6] configured with three residential-style buildings and Saudi weather files (Riyadh, Jeddah, Abha). Episodes span one simulation year (8,760 hourly steps). Building load traces are generated from EnergyPlus[7] and provided to CityLearn as CSV files; CityLearn handles the state transition, constraints, and KPI computation. The environment is run in *central-agent* mode so that one policy controls all buildings jointly.

3.3 Reward Function

The reward mirrors the operational objectives implemented in the codebase: thermal comfort around a setpoint, smooth operation (low ramping of net power), and reliability during outages (serve non-shiftable load when possible). For building i at time t:

$$\Delta T_t^{(i)} = T_{\text{in},t}^{(i)} - T_{\text{set},t}^{(i)}, \tag{2}$$

$$\text{err}_t^{(i)} = \max\Big(0,\ \big|\Delta T_t^{(i)}\big| - \delta\Big) \cdot \begin{cases} m, & \Delta T_t^{(i)} < 0 \\ 1, & \text{otherwise} \end{cases}, \tag{3}$$

$$\text{ramp}_t^{(i)} = \Big|P_t^{(i)} - P_{t-1}^{(i)}\Big|, \tag{4}$$

$$\text{out}_t^{(i)} = \mathbf{1}\{o_t^{(i)} = 1\}\ \max\Big(0,\ \text{NSL}_t^{(i)} - P_t^{(i)}\Big). \tag{5}$$

The per-building reward and the centralized aggregation are

$$r_t^{(i)} = -0.3\,\text{err}_t^{(i)} - 0.15\,\text{ramp}_t^{(i)} - 0.1\,\text{out}_t^{(i)}, \tag{6}$$

$$R_t = \frac{1}{N}\sum_{i=1}^{N} r_t^{(i)}, \tag{7}$$

where T_{in} is indoor temperature, T_{set} a cooling setpoint, δ is a small deadband, $m > 1$ increases penalty when too cold, P is net electricity, NSL is non-shiftable load, and o_t flags outages.

3.4 Agent Architecture (SAC)

We use Soft Actor–Critic (SAC) implemented with Stable Baselines3[8]. Observations and actions are flattened via a wrapper that ensures strict bounds, handles NaNs/infs, and maps actions back to CityLearn per-building vectors. The policy and critics are multi-layer perceptrons with ReLU activations. Representative hyperparameters are: learning rate 1×10^{-4}, buffer size 10^5, batch size 128, discount $\gamma = 0.95$, target update coefficient $\tau = 0.01$, entropy coefficient 0.1, training frequency every step, and network widths $[128, 128]$.

3.5 Training and Evaluation Protocol

Training is conducted with a centralized policy and per-step updates using experience replay. We periodically save checkpoints and log episode returns. For evaluation, we run deterministic rollouts on the full-year episode and compute CityLearn challenge KPIs. Our evaluation utilities normalize action shapes for all agents to ensure comparability and export KPIs/plots (CSV/PNG) for analysis.

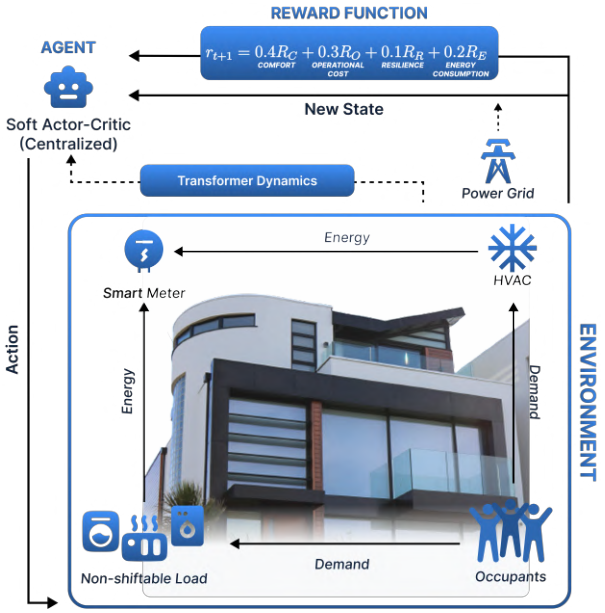


Figure 2: Methodology overview: a centralized SAC agent interacts with the CityLearn environment; the reward balances comfort, smooth operation (ramping), and outage unserved energy.

3.6 Baselines

We compare against: (i) an improved RBC that uses comfort bands for cooling and simple storage heuristics; and (ii) a practical MPC baseline that solves a short-horizon optimization and is mapped to CityLearn actuators via a reverse-engineering step for device intensities. All methods share the same environment schema and KPI pipeline.

4 Experimental Results

4.1 Baselines in Detail

Rule-Based Controller (RBC). The RBC turns the cooling device on when indoor temperature exceeds a comfort band above the setpoint and off otherwise. Thermal storages follow simple heuristics: charge when solar/cheap hours are available and discharge during peak windows; actions are clipped to device limits. This yields predictable behavior with low compute but limited adaptability across seasons and buildings.

Model Predictive Control (MPC). The MPC baseline uses a receding horizon (short prediction and control horizons) to plan setpoints and storage charge/discharge subject to comfort and actuator constraints. Because CityLearn expects device intensities, we map the planned energy/setpoints to actuator commands via a calibrated reverse-engineering step; when the solver fails or is infeasible, a safe fallback policy is used. This baseline captures some look-ahead but is sensitive to model/forecast errors and solver settings.

4.2 Evaluation Metrics

We use CityLearn's official KPIs. All values are normalized and reported so that *lower is better*. Below we give a plain-language definition for each metric used in this paper:

- **Discomfort proportion**: Fraction of hours when indoor temperature is outside the comfort band around the setpoint. Captures occupant comfort.
- **Ramping average**: Average hour-to-hour change in net electricity consumption. Lower values mean smoother, less jerky operation.
- **Daily 1−load factor average**: Measures how flat the daily load profile is. A value near 0 indicates a flatter (more efficient) daily profile.
- **Daily peak average**: Typical daily maximum demand, normalized. Lower values reduce daily peak stress on the grid.
- **All-time peak average**: Maximum demand across the entire period, normalized. Lower values mean smaller absolute peaks.
- **Annual normalized unserved energy**: Portion of required energy not served (e.g., during outages). Lower is safer and more reliable.

Metric	RBC	MPC	SAC
Discomfort proportion	0.4242	0.4107	**0.2287**
Ramping average	1.1270	1.0069	**1.0019**
Daily peak average	0.9949	**0.9611**	0.9861
All-time peak average	1.2719	**0.9559**	0.9922
Daily 1−load factor avg	1.0249	1.0215	**1.0091**
Annual unserved energy	**0.4360**	0.5003	0.5270
Thermal resilience (1−)	0.4244	0.4099	**0.2239**

Table 1: District-level KPIs (lower is better). Winners are bolded. SAC strongly reduces discomfort and ramping; MPC slightly lowers peaks and consumption; aggregate differences are modest due to static operational pricing.

- **Thermal resilience (1−)**: Proxy for resilience under stress; lower values indicate fewer/shorter comfort violations during difficult periods.

4.3 Results Summary

Table 1 compares district-level KPIs across RBC, MPC, and SAC. Values are normalized, and lower is better.

4.4 Findings

Comfort and smoothness. SAC achieves the lowest discomfort and near-minimal ramping, indicating smoother operation without frequent on/off swings.

Peaks and consumption. MPC edges out on daily and all-time peaks and attains the lowest electricity consumption, consistent with short-horizon planning aligned to device/constraint limits.

Resilience proxy. The (1−thermal resilience) metric is lowest for SAC, reflecting fewer and smaller comfort violations during stress periods.

Static pricing effect. With largely static operational pricing in this Saudi setup, temporal incentives are weak; consequently, RBC, MPC, and SAC converge to similar aggregate scores and grid-facing KPIs. Differences emerge primarily in comfort/ramping (favoring SAC) versus peaks/consumption (slightly favoring MPC).

Parity with baselines. While SAC does not win every KPI, it matches or closely tracks the baselines on most grid metrics, *while simultaneously delivering the best comfort and smoothness.* This demonstrates that a single learned policy can attain baseline-level efficiency without sacrificing comfort.

Practical takeaway. In environments with flat tariffs and limited price signals, policy choice is driven more by comfort targets and smoothness preferences than by cost arbitrage; adding dynamic prices or DR signals would likely widen performance gaps.

5 Conclusion

CORE-EMS demonstrates that a single, centralized SAC policy can manage building cooling and thermal storage effectively across diverse Saudi climates. While SAC does not dominate every KPI, it consistently attains grid-facing performance on par with rule-based and MPC baselines and simultaneously delivers superior comfort and smoother operation. Under largely static operational pricing, these methods converge to similar aggregate scores; nevertheless, SAC's adaptability and comfort advantages make it a compelling choice for real deployments. Future work will incorporate dynamic tariffs and demand-response signals, which we expect will amplify the benefits of learning-based control.

References

[1] Saudi Energy Efficiency Center, "Buildings sector," https://www.seec.gov.sa/en/about/energy-efficiency-sectors/buildings-sector, 2025, accessed: 2025-08-20.

[2] T. Zangato, A. Osmani, and P. Alizadeh, "Data-driven policy mapping for safe rl-based energy management systems," 2025, arXiv preprint arXiv:2506.16352. [Online]. Available: https://arxiv.org/abs/2506.16352

[3] H. Zhang, S. Seal, D. Wu, B. Bouffard, and G. Joos, "Data-driven model predictive and reinforcement learning-based control for building energy management: a survey," *IEEE Access*, vol. 10, pp. 1–1, 2022, licensed

under CC BY 4.0. [Online]. Available: https://arxiv.org/abs/2106.14450

[4] L. Yu, S. Qin, M. Zhang, C. Shen, T. Jiang, and X. Guan, "A review of deep reinforcement learning for smart building energy management," *arXiv preprint arXiv:2008.05074*, 2020, revised September 2021. [Online]. Available: https://arxiv.org/abs/2008.05074

[5] T. N. Y. Huynh, A. T. Nguyen, Y. Ahn, B. L. Oo, and B. T. Lim, "Multi objectives reinforcement learning for smart buildings: A systematic review of algorithms, applications and future perspectives," *Energy and Buildings*, vol. 345, p. 116045, 2025.

[6] J. R. Vázquez-Canteli, L. García *et al.*, "Citylearn: A multi-agent reinforcement learning environment for demand response," 2023, https://github.com/intelligent-environments-lab/CityLearn.

[7] D. B. Crawley, L. K. Lawrie, F. C. Winkelmann, W. F. Buhl, Y. J. Huang, C. O. Pedersen, R. K. Strand, R. J. Liesen, D. E. Fisher, M. J. Witte *et al.*, "Energyplus: creating a new-generation building energy simulation program," *Energy and buildings*, vol. 33, no. 4, pp. 319–331, 2001.

[8] A. Raffin, A. Hill, A. K. Ernestus, Adam Gleave *et al.*, "Stable-baselines3: Reliable reinforcement learning implementations," in *Journal of Machine Learning Research*, 2021, https://github.com/DLR-RM/stable-baselines3.