UDC 004.8
IRSTI 28.23

https://doi.org/10.55452/1998-6688-2026-23-2-218-232

[1]*Kapparova A.,
PhD student, ORCID ID: 0009-0004-4639-3548,
*e-mail: kapparova_ainur3@kaznu.edu.kz
[1]Zholamanov B.,
PhD student, ORCID ID: 0000-0001-8206-7425,
e-mail: zholamanov.batyrbek@kaznu.kz
[1]Bolatbek A.,
PhD student, ORCID ID: 0009-0004-7613-5507,
e-mail: bolatbek.askhat@kaznu.kz
[1]Kuttybay N.,
PhD, ORCID ID: 0000-0002-5723-6642,
e-mail: kuttybyy.nurzhigit@kaznu.kz
[1]Dosymbetova G.,
PhD, ORCID ID: 0000-0002-3935-7213,
e-mail: gulbakhar.dossymbetova@kaznu.edu.kz
[1]Zhumagaliyev Ye.,
Master student, ORCID ID: 0009-0002-9213-2555,
e-mail: zhumagaliyev_y0@live.kaznu.kz

[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan

IMPLEMENTATION OF MULTI-AGENT FRAMEWORK OF IMPALA FOR A SINGLE ZONE TEMPERATURE CONTROL OF A SIMULATED THERMAL ZONE

Abstract

Buildings account for a significant portion of global energy consumption, with HVAC and humidity-control systems representing the majority of their operational demand. Traditional rule-based strategies often fail to adapt to dynamic indoor-outdoor conditions, motivating the use of data-driven control methods. This study presents a multi-agent reinforcement learning (MARL) framework for simultaneous temperature and humidity control in a single-zone building modeled in EnergyPlus. The proposed approach employs the distributed Importance Weighted Actor-Learner Architecture (IMPALA) algorithm with centralized training and decentralized execution (CTDE), enabling two agents: temperature and humidity to learn coordinated policies directly from high-fidelity simulation feedback. The results demonstrate strong learning performance: both agents improved their per-step rewards substantially (temperature +18.9%, humidity +33.7%), indicating effective convergence and cooperative behavior. The learned controller maintained thermal comfort comparable to the rule-based baseline (mean occupied temperature difference ≈ 0.04 °C; occupied PMV ≈ 0.45) while achieving notable energy savings. Total annual HVAC energy consumption decreased by 8.9%, with the most significant improvement observed in humidification energy, which was reduced by 34.4%. Heating and cooling loads remained nearly unchanged, confirming that energy reductions were achieved without compromising comfort.

Keywords: building model, multi-agent framework, multi-agent reinforcement learning, intelligent control, Importance Weighted Actor-Learner Architecture

Received November 17, 2025; revised February 19, 21, 26, 2026; accepted March 31, 2026

1 Introduction

Globally, buildings account for approximately 30% of total energy consumption, making them one of the largest contributors to global energy use [1]. Among building subsystems, the heating, ventilation, and air-conditioning (HVAC) system is the primary consumer, typically responsible for nearly 50% of a building's total energy demand [2]. In densely populated tropical and subtropical cities, air-conditioning and mechanical ventilation (ACMV) systems can account for up to 70% of building energy use [3]. As global energy resources continue to deplete, improving the efficiency of HVAC and ACMV systems and reducing their energy consumption are key priorities for achieving sustainable and low-carbon buildings.

While reducing energy use is essential, it must be achieved without compromising occupant comfort. Thermal comfort depends on temperature, humidity, air quality, and other environmental factors. Traditionally, buildings have used rule-based control (RBC) systems, which operate HVAC equipment according to predefined rules and schedules derived from expert knowledge [4, 5]. Although such methods are easy to implement, they lack adaptability to nonlinear and dynamic environmental conditions, often leading to inefficient energy performance and comfort degradation.

In contrast, model-based control methods such as Model Predictive Control (MPC) have been widely adopted in recent years. MPC forecasts future thermal dynamics using a mathematical model and computes optimal control actions over a prediction horizon. It can significantly reduce energy use while maintaining comfort by considering forecasted loads and system responses. However, building accurate thermal and moisture models is complex, time-consuming, and requires specialized expertise. When these models fail to represent real-world dynamics accurately, system performance deteriorates. Moreover, MPC methods are computationally expensive, especially in multi-zone buildings where each thermal zone introduces additional state variables [6].

To support such advancements, a variety of high-fidelity building simulation tools have been developed, including EnergyPlus [7–10]. These platforms accurately simulate thermodynamic phenomena such as heat transfer, airflow, humidity regulation, and system dynamics, allowing for in-depth analysis of building performance. Despite their precision, the embedded mathematical models are typically nonlinear and computationally intensive, requiring extensive calibration and user input. Consequently, they are often impractical for real-time control applications or iterative reinforcement learning loops [11].

Effective building operation requires managing the interactions between ventilation, heating, and lighting systems while maintaining overall energy balance. For instance, increased ventilation raises the heating load due to the need to warm incoming air, while continuous heating during unoccupied hours results in energy waste. Adaptive and dynamic control strategies that consider occupancy schedules, solar gains, and internal loads can minimize energy use while maintaining comfort [12]. However, even the most advanced model-based control methods cannot perfectly capture real-time variations in weather, occupancy, or user behavior. Therefore, model-free, intelligent control techniques capable of learning optimal actions directly from environmental feedback have become an emerging research focus.

Recent breakthroughs in artificial intelligence have introduced Reinforcement Learning (RL) and, more specifically, Deep Reinforcement Learning (DRL) as promising approaches for adaptive building energy control. Unlike traditional control strategies, RL does not rely on explicit physical modeling; instead, it learns optimal control policies through trial-and-error interactions with the environment [13]. Data-driven DRL frameworks have demonstrated strong potential for managing complex multi-zone HVAC systems [14]. For example, Barrett and Linder [15] reported that reinforcement learning-based HVAC control achieved up to 10% energy savings compared with programmable thermostats.

Despite these advantages, online DRL methods where agents learn directly from physical building interactions – require long training durations and may risk comfort violations during early

learning stages. For instance, Mocanu et al. [16] demonstrated that DRL agents needed an entire year of continuous operation to converge to near-optimal control policies. As a result, simulation-based pre-training in virtual environments such as EnergyPlus is now a preferred approach, providing a safe and efficient means for policy development and evaluation before deployment.

Within the actor–critic family of RL algorithms, the Deep Deterministic Policy Gradient (DDPG) method has been successfully applied to continuous control tasks such as cooling optimization in data centers [17] and transactive HVAC systems [18]. However, single-agent DRL becomes inefficient in multi-zone environments due to exponentially growing state–action spaces. To overcome this, researchers have proposed Multi-Agent Reinforcement Learning (MARL) frameworks, where multiple agents collaborate or compete to optimize system-level objectives. For example, [19] formulated HVAC control as a Markov Game, enabling decentralized agents to coordinate local policies. Similarly, [20] proposed strategies to reduce action–state complexity and improve convergence while preserving energy efficiency and occupant comfort.

The integration of Deep Reinforcement Learning (DRL) and Multi-Agent Reinforcement Learning (MARL) in HVAC systems thus represents a significant step toward intelligent, adaptive, and energy-efficient buildings. By overcoming the limitations of traditional rule-based and model-based methods, these data-driven approaches enable real-time optimization of temperature, humidity, and ventilation in response to dynamic environmental and occupancy conditions. As research in this field advances, MARL-based frameworks are expected to play a key role in the development of smart, self-learning building systems that jointly optimize comfort, energy, and sustainability.

In this study, we implement a multi-agent reinforcement learning framework for the automatic control of temperature and relative humidity in a single zone building model simulated in EnergyPlus. The framework is built upon the Importance Weighted Actor-Learner Architecture (IMPALA) algorithm which is a distributed actor–critic DRL method capable of parallelizing data collection and learning across multiple agents. Two agents are designed: a temperature-control agent that manages heating and cooling setpoints, and a humidity-control agent that regulates humidifier operation.

The novelty of this research lies in the implementation of IMPALA with a multi-agent framework for building energy management system. The control algorithm is integrated the EnergyPlus simulation platform via a custom Python–EnergyPlus interface. This connection allows the agents to interact with the high-fidelity building simulation in real time and learn optimal control policies that balance:

- Occupant comfort, measured through Predicted Mean Vote (PMV) and relative humidity;
- Energy efficiency, quantified via heating, cooling, and humidifier energy consumption; and
- Occupancy-dependent operation, distinguishing between occupied and unoccupied periods.
- Compared with conventional rule-based control, the proposed IMPALA-based MARL system achieved:
- Up to 9% total energy savings, primarily due to a 34% reduction in humidifier energy consumption,
- Minimal comfort deviation, maintaining nearly the same indoor temperature and PMV levels as the baseline, and
- Improved policy stability, with agents showing consistent reward improvement over training iterations.

This work contributes to the development of simulation-integrated multi-agent reinforcement learning architectures for intelligent building energy management. It establishes a scalable experimental foundation for extending toward multi-zone, multi-agent EnergyPlus simulations, facilitating the design of next-generation sustainable, comfort-preserving, and autonomous HVAC control systems for low-carbon smart buildings. The paper's structure is outlined as follows: In Methods section, the methodology of the proposed architecture and the simulation environment framework are delineated, Results section presents the training and simulation results. Finally, Conclusions concludes the study.

2 Materials and methods

2.1 Formulation of the HVAC Control Problem

In this study, the control of single zone building HVAC systems is formulated as a Markov Decision Process (MDP) framework. Under the simplified assumption of building thermal dynamics, the indoor temperature at a given time step depends primarily on the system's previous state such as the indoor temperature and control inputs from the preceding interval and is independent of earlier time steps. Consequently, the HVAC control process can be modeled as a finite Markov process, making it suitable for solution using Deep Reinforcement Learning (DRL) techniques.

The DRL-based control framework for building HVAC systems comprises five core components: the environment, the agent, the state space (s), the action space (a), and the reward function (r). In this context, the zone and its HVAC system represent the environment, while the DRL controller acts as the agent. The agent observes the current conditions of the environment through the state space, which encodes information such as zone temperature, humidity, occupancy, and energy usage. Based on these observations, the agent selects an action from its discrete or continuous action space such as adjusting temperature setpoints or modifying humidity states to influence the environment.

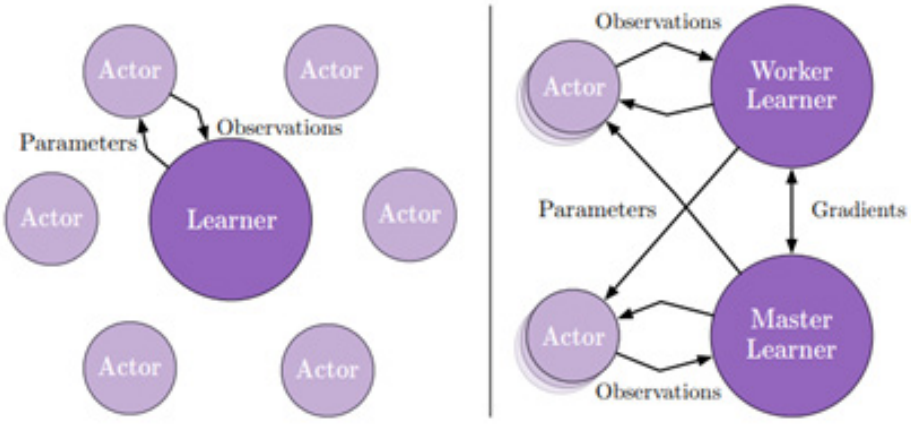
After the environment transitions to a new state in response to the agent's action, a reward signal is generated, reflecting the effectiveness of the chosen action in achieving the control objectives: energy efficiency and occupant comfort.

This reward function serves as the learning feedback that guides the agent toward developing an optimal control policy, enabling it to make progressively better decisions through iterative interaction with the environment. As a DRL technique Importance Weighted Actor-Learner Architecture (IMPALA) is used.

2.2 IMPALA algorithm

In this study, the Importance Weighted Actor-Learner Architecture (IMPALA) algorithm is adopted as the distributed reinforcement learning framework. IMPALA efficiently utilizes computational resources on a single machine and can scale to large distributed systems while maintaining high data efficiency. It employs an actor–critic architecture where multiple actors generate experience trajectories in parallel, and centralized learners update the global policy π and value function V_π based on received trajectories.

The learning and acting processes (Figure 1) are decoupled, enabling high-throughput training. To address off-policy discrepancies caused by asynchronous updates between actors and the learner, IMPALA integrates a correction mechanism known as V-trace, which ensures stable and efficient policy optimization under distributed conditions.



a) b)

Figure 1 – The learning and acting processes: a – Single Learner; b – Multiple Synchronous Learners [21]

IMPALA efficiently utilizes computational resources on a single machine and scales seamlessly across multiple systems while maintaining data efficiency. It follows an actor–critic architecture, where multiple actors generate experience trajectories in parallel, and a centralized learner updates the global policy (π) and value function (Vπ). The decoupling of experience collection from learning allows high-throughput training [21].

To address the off-policy nature of this setup caused by delays between actors and the learner IMPALA integrates a correction mechanism known as V-trace, which stabilizes learning. In the V-trace actor–critic algorithm, the value parameters θ are updated through gradient descent to minimize the difference between the estimated value $V_\theta(x_s)$ and the V-trace target v_s. The policy parameters ω are optimized using the policy gradient weighted by the importance-sampling ratio ρ_s, which corrects for deviations between the behavior policy μ and the target policy π_ω:

$$\rho_s \nabla_\omega \log \pi_\omega(a_s \mid x_s)(r_s + \gamma v_{s+1} - V_\theta(x_s)) \quad (1)$$

To enhance policy exploration and avoid premature convergence, an entropy regularization term is added, encouraging the agent to explore diverse actions:

$$-\nabla_\omega \sum_a \pi_\omega(a \mid x_s) \log \pi_\omega(a \mid x_s) \quad (2)$$

The overall parameter update is obtained by combining these gradients with appropriate weighting coefficients (hyperparameters) [21]. This configuration enables stable and efficient learning under distributed asynchronous conditions, making IMPALA with V-trace particularly well-suited for continuous control tasks such as building HVAC and humidity regulation.

2.3 State space, Action space, Reward function

Observation vector (11 variables): [OAT, IAT, RH, CO2, CLG_SPT, HTG_SPT, CLG_J, HTG_J, OCC, HUM_J, PMV] covering environmental and energy states. Two mechanisms ensure numerical stability and smooth learning: RunningStat class: Implements Welford's algorithm to maintain running mean and variance for observation normalization. EMAStd class keeps an exponential moving estimate of reward standard deviation to normalize and stabilize reward signals. Normalization is essential because EnergyPlus variables vary across wide ranges.

Action Spaces and Mapping: 1) Temperature agent: MultiDiscrete action [cool_index, heat_index] mapped to physical setpoints (e.g., 25–27 °C for cooling, 21–23 °C for heating). A deadband constraint ensures logical order: heating setpoint < cooling setpoint. 2) Humidity agent (z1_hum): Discrete levels mapped linearly to the humidifier command [0.0, 1.0]. This discretization makes control tractable for RL while preserving physical interpretability.

Reward function used in this study:

$$R = -(\alpha_{\text{temp/hum}})(E_{\text{cost}} + \text{comfort cost} + \text{penalties}) \quad (3)$$

with normalization and clipping to stabilize training.

The reward balances comfort and energy efficiency, using both linear and nonlinear components (Table 1).

Table 1 – Reward function components

Component	Description	Applies to
PMV comfort cost	Penalizes deviation of PMV from tolerance band	Temperature agent
Humidity discomfort	Penalizes deviation of RH from comfort range (35–55%)	Humidity agent
Energy cost	Penalizes cooling, heating, or humidifier energy (kWh)	Both
Occupancy weighting	Different weights for occupied vs unoccupied times	Both
Work-hour boost	Comfort is more important during working hours	Temp agent
Setpoint churn penalty	Discourages rapid setpoint changes	Temp agent
Extreme setpoint penalty	Penalizes setting at physical limits	Temp agent
No-fight penalty	Penalizes simultaneous heating and cooling	Temp agent

2.4 Algorithm structure and training parameters

The proposed control framework employs the IMPALA integrated with Centralized Training and Decentralized Execution (CTDE) over the EnergyPlus simulation environment. The framework enables distributed multi-agent learning for indoor temperature and humidity control.

1. Actors (Rollout Workers): Multiple rollout workers sample trajectories from the EnergyPlus environment, each containing two agents: temperature and humidity.

2. V-trace Correction: Off-policy correction converts actor trajectories collected under a behavior policy into learner updates consistent with the target policy.

3. Learning Phase: The central learner updates a shared or separate policy for each agent using the IMPALA actor-critic setup. An LSTM module can be applied to handle partial observability in building dynamics.

4. Callback Metrics: Episode-level data such as energy consumption (kWh), thermal comfort (PMV), and humidity discomfort are aggregated. Reward normalization may be applied online to stabilize training.

5. CTDE Scheme: During centralized training, the environment provides a global critic observation that incorporates all agents' information. During decentralized execution, each agent acts independently using its own local policy.

IMPALA employs V-trace to correct for policy lag between actors and the learner. The hyperparameters of the current architecture are given in Table 2.

Table 2 – Optimization hyperparameters

Hyperparameter (flag)	Meaning	Values
--gamma	Discount factor	0.95
--lr	Adam learning rate	5e-4
--entropy-coeff	Policy entropy bonus	1e-3
--grad-clip	Global grad clip (L2)	40.0
--train-batch-size	SGD batch size	4000
--rollout-fragment-length	Steps per sample batch	50
--vtrace-clip-rho / --vtrace-clip-pg-rho	Importance ratio clips	1.0 / 1.0
--use-lstm + --lstm-cell + --lstm-seq-len	Recurrent policy	off / 256 / 20

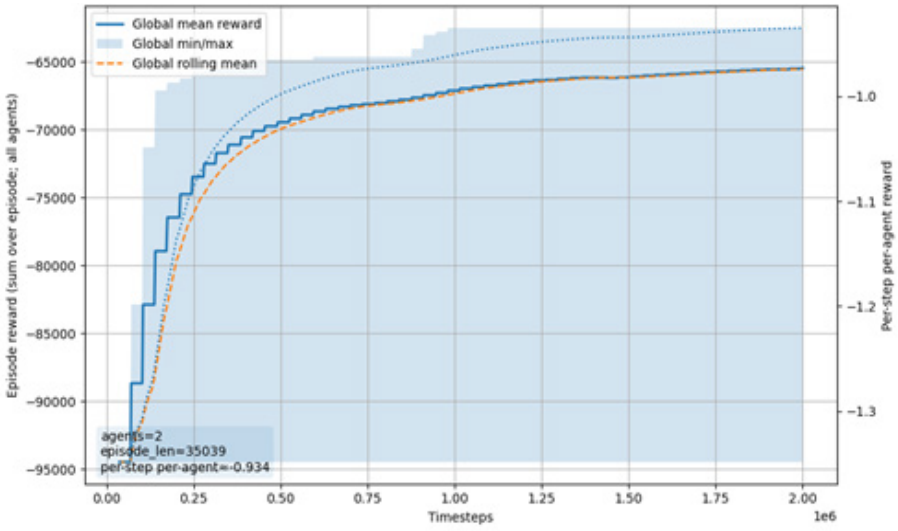
Recurrent policies LSTM are configured with --lstm-cell 256 and --lstm-seq-len 20, while the burn-in argument is retained for compatibility with previous configurations.

3 Results and discussion

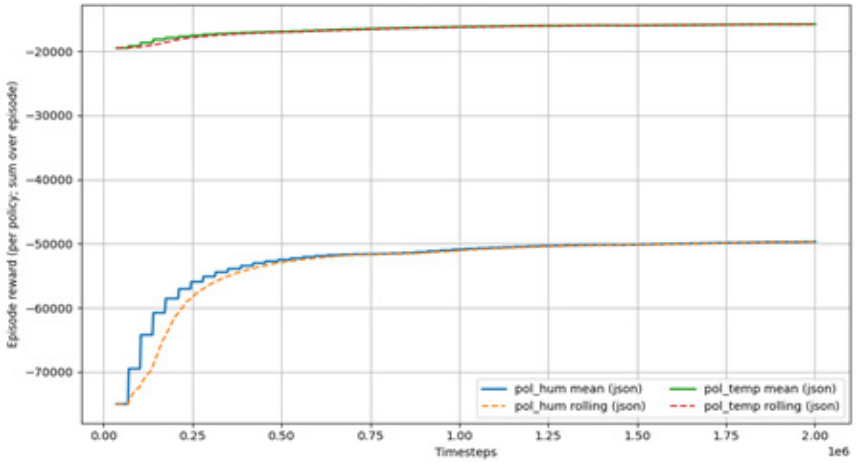
As studied environment a room with an area 41.54 m² and volume 147.88 m³ is modeled in EnergyPlus. The simulated environment uses weather of Almaty for one year: KAZ_ALA_Almaty.368700_TMYxEPW.epw which is from the database of EnergyPlus [22]. Two actuators were used in the simulation: zone temperature control (cooling/heating setpoints) and humidifier power value. These mappings allow EnergyPlus to execute actions from RL agents in real-time simulation via the PyEnergyPlus API. Simulations can be daily episodes, in this case, 24 hours × 4 steps per hour.

3.1 Training Performance

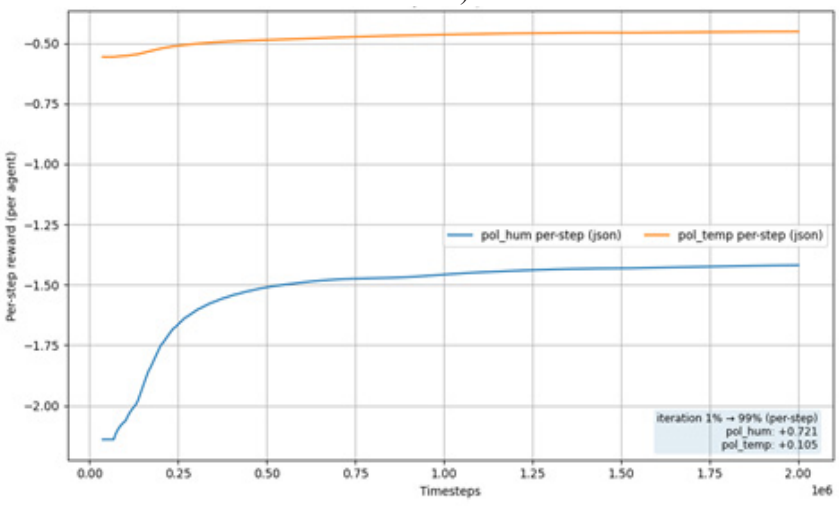
The training performance in reinforcement learning is evaluated by reward progress. Figure 2 shows reward progress of multi-agent IMPALA algorithms which have 2 agents. The training process used 2,000,000 timesteps and consisted of 56 episodes, each spanning one year. During training, both agents showed stable learning progress: 1) Temperature policy: mean step reward improved from −0.557 to −0.451, representing a +18.9% increase in average per-step reward; 2) Humidity policy: mean step reward improved from −2.140 to −1.418, corresponding to a +33.7% increase in control efficiency. These results indicate that both controllers successfully learned more balanced strategies, reducing penalties associated with comfort violations and energy use.



a)



b)



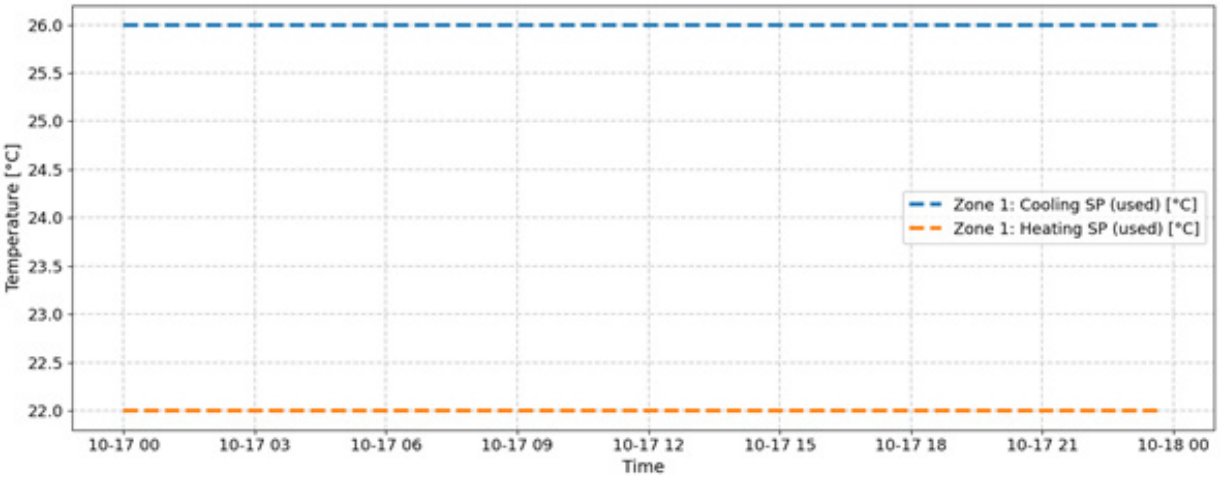
c)

Figure 2 – Reward results: a – episode reward (sum over episode for all agents); b – per-policy episode rewards; c – per-policy per-step rewards

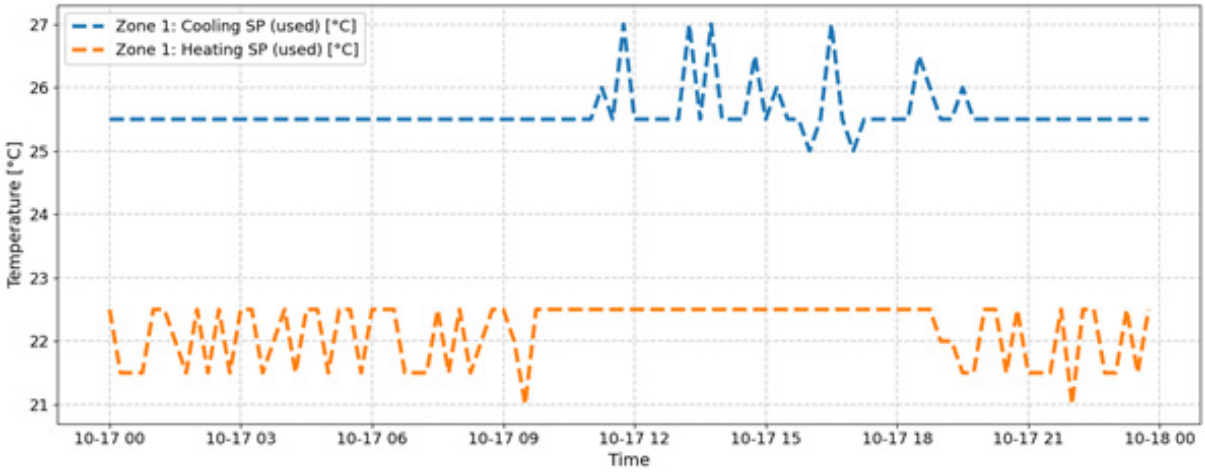
As Figure 2 shows, in the first step of training, the reward value of both agents was low, and then the training began to stabilize from 500,000 timesteps. This may indicate that the RL control system is unstable in the first steps and that the energy saving or the comfort level of people may decrease. Therefore, using the weights of trained agents in practice can give the desired result.

3.2 Thermal Comfort Analysis

In the case of rule-based control method, the zone thermostat gets constant values which are 22°C for heating and 26°C for cooling. In the case of scheduled RBC control, the heating setpoint is set to 22°C between 09:00 and 18:00 and 21°C at other times, and the cooling setpoint was set to 26°C between 09:00 and 18:00 and 27°C at other times. Figure 3a and Figure 3b show the thermostat setpoint values in two cases for one day, which is the 17th of October: rule-based control and multi-agent IMPALA control methods. In the case of MA IMPALA, the thermostat could choose heating setpoint in the range of 21-23°C, cooling setpoint in the range of 25-27°C.



a)



b)

Figure 3 – Heating and cooling setpoints for the 17th of October:
a – rule-based control; b – IMPALA control

PMV (Predicted Mean Vote) from EnergyPlus' Fanger model is used as the primary comfort indicator. The PMV and temperature profiles demonstrate that the multi-agent controller could maintain occupant comfort within acceptable limits throughout the year. Figure 4 demonstrates

annual PMV values for each month when using IMPALA. Annual mean PMV, when the zone was occupied, reached +0.453, corresponding to a slightly warm but comfortable range (−0.5 < PMV < +0.5). The monthly PMV trend peaked during summer months (June–August), aligning with higher indoor air temperatures (≈ 25 °C). During winter (January–March), PMV values dropped slightly below zero (cool sensation), indicating seasonal adaptability of the learned policy.

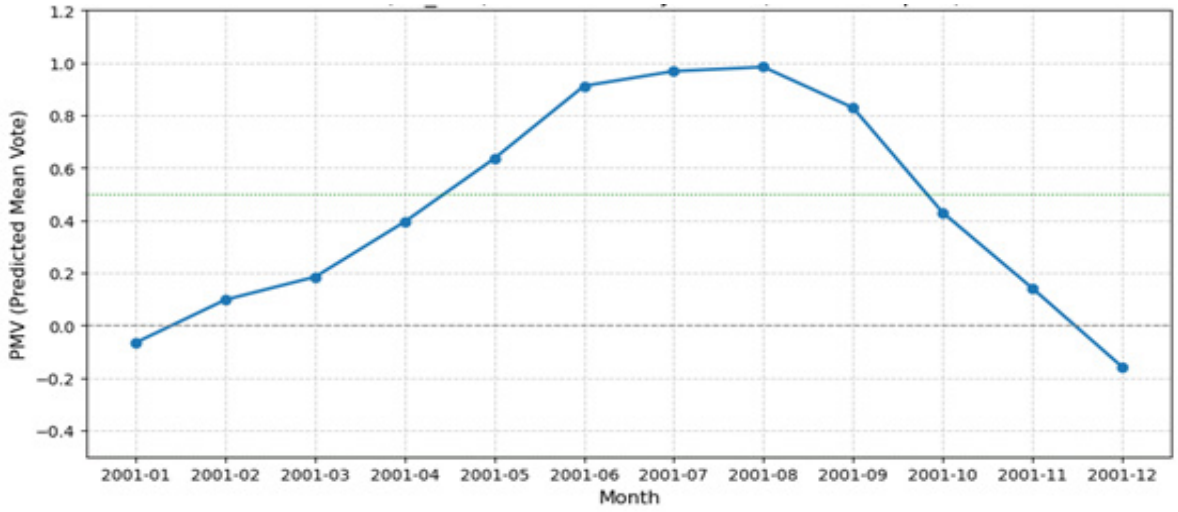


Figure 4 – Annual PMV values for each month when the zone is occupied

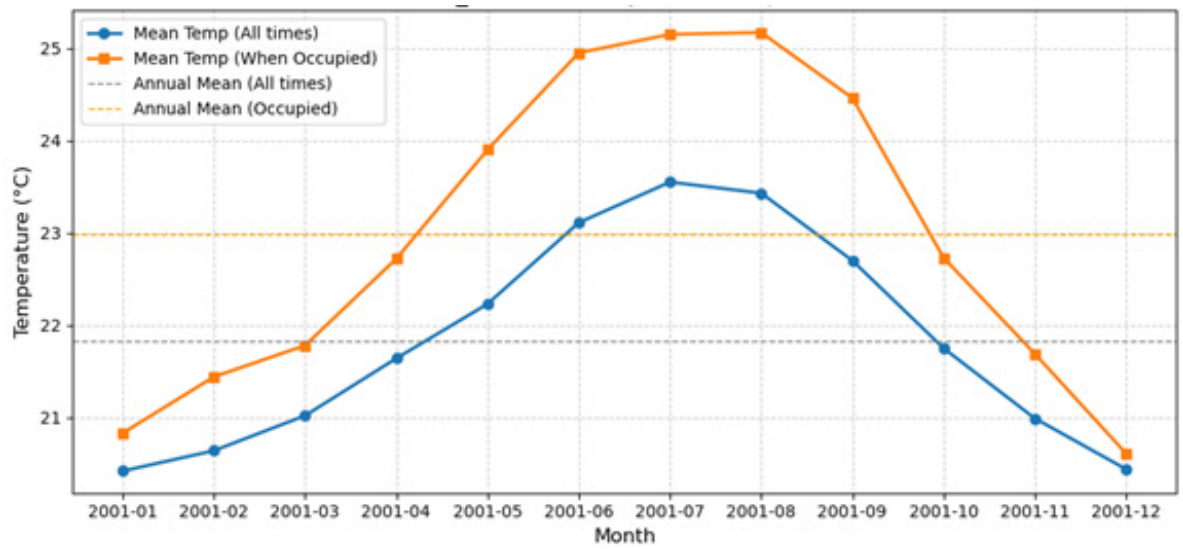


Figure 5 – Monthly mean temperature

Figure 5 demonstrates monthly mean temperature when using IMPALA in both cases: occupied and unoccupied. The mean indoor temperature when occupied was maintained close to the comfort reference. The value of mean indoor temperature of one year was 23.019 in rule-based control case and 22.978 in MARL control case which is similar in two cases. This small temperature difference shows that comfort was preserved while improving energy efficiency.

Mean temperature deviation was compared between episodes of MARL, rule-based control and scheduled control algorithms. The results are shown in Figure 6, according to the graph, while in the first episode the mean temperature deviation from the comfort range was 17%, in the last 56 episodes it decreased to 14%. And this indicator is about 15% better than the scheduled RBC.

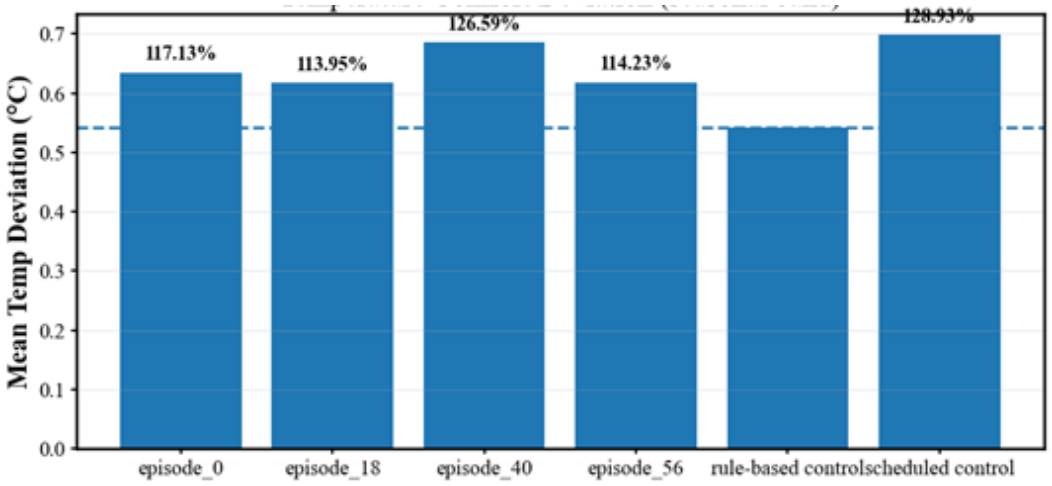


Figure 6 – Temperature comfort deviation

3.3 Humidity Control

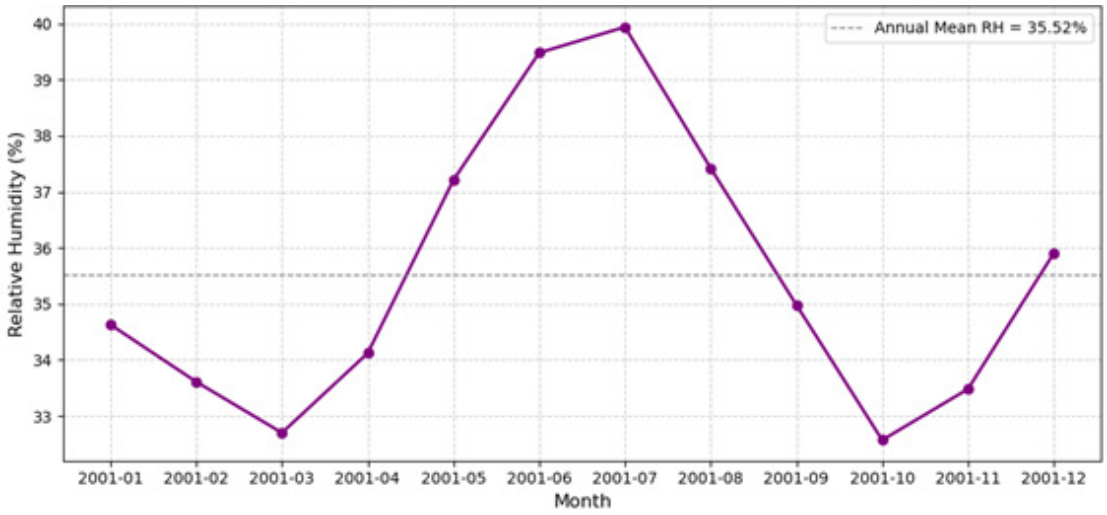


Figure 7 – The average annual relative humidity

When using IMPALA, the average annual relative humidity (RH) was approximately 35.5%, with seasonal variation between 33–40% (Figure 7). The multi-agent controller effectively maintained RH within the desired range (35–55%) most of the time, avoiding excessive humidifier usage.

The mean RH deviation, when the zone is occupied, of MARL episodes was compared with RBC and scheduled RBC control methods. The results in Figure 8 show that the deviation was very large in the first episode, while the relative humidity deviation decreased in the last episode. Scheduled control was less effective in maintaining relative humidity in the comfort range than RBC and MARL.

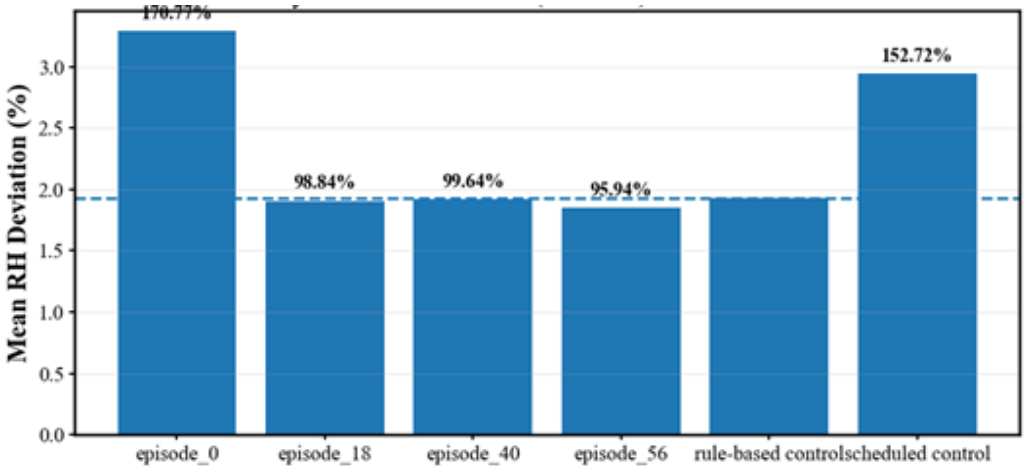


Figure 8 – Humidity comfort deviation when the zone is occupied

3.4 Energy Consumption

Table 3 shows a detailed comparison between rule-based and multi-agent control for the zone that shows significant energy savings.

Table 3 – Comparison between rule-based and multi-agent control

Energy End-Use	Rule-based (kWh)	Multi-agent (kWh)	Δ (kWh)	Δ (%)
Total	28,220	25,698	−2,521.7	−8.9%
Humidifier	7,257.5	4,759.6	−2,497.9	−34.4%
Heating	17,624.8	17,571.9	−53.0	−0.3%
Cooling	3,337.8	3,367.0	+29.2	+0.9%

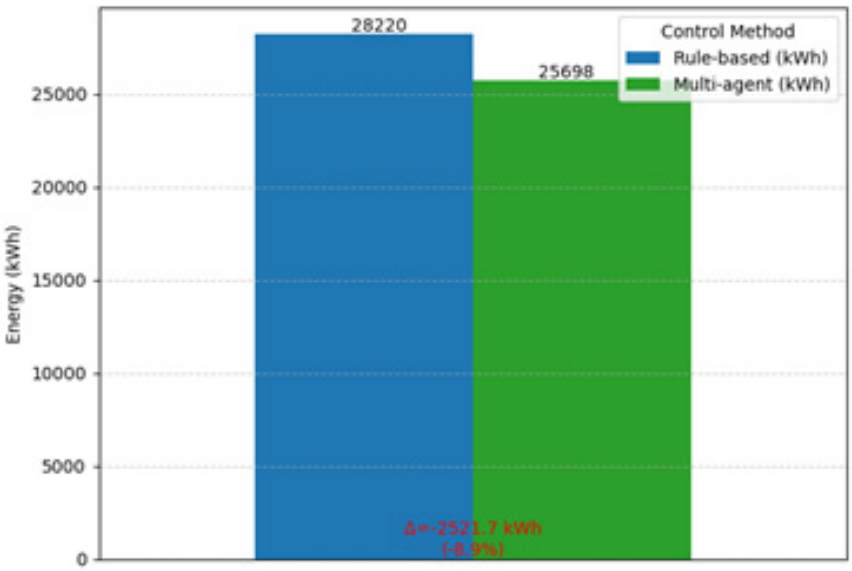


Figure 9 – Comparison of total energy consumption

The multi-agent controller notably reduced humidifier electricity usage by 34.4%, while maintaining comparable heating and cooling loads. This resulted in an overall 8.9% reduction in total annual HVAC energy consumption (Figure 9). This shows that treating the problem of energy efficient use of a building as a Markov game, evaluating the system's decisions in the certain states through rewards is an advantageous method for minimizing energy consumption while maintaining target temperatures for people in the zone.

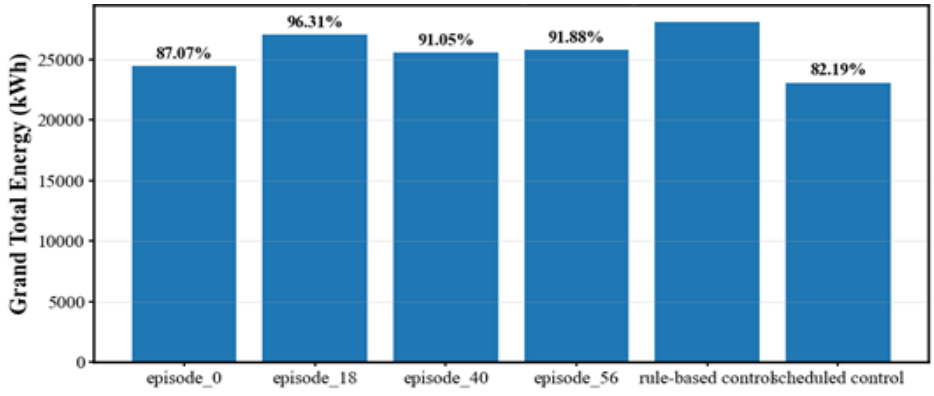


Figure 10 – Total energy consumption of different algorithms

Figure 10 shows the least energy-efficient control system, the scheduled RBC system with 17.81% reduction. However, based on the above analysis, we can see that the scheduled RBC reduces occupancy comfort when controlling temperature and humidity. In the first episode of subsequent MARL control, although the energy consumption was reduced by 12.93%, the deviation of temperature and humidity from the comfort range was greater than in episode 56.

Overall, the advantages of the proposed method are 1) physically consistent control: all actuators and feedbacks are real EnergyPlus quantities; 2) multi-objective optimization: combines energy, comfort, and operational stability; 3) scalable design: easily extendable to multi-zone and multi-agent systems; 4) supports CTDE: ideal for MARL algorithms like DDPG, IMPALA, or MAPPO; 5) reusable: flexible normalization and reward shaping for different climates and buildings.

The limitation of the proposed algorithm is its difficulty in real-world implementation. The implementation of the method requires more computer resources and technical support than RBC and scheduled control. However, this algorithm can be used to control the operation of the other thermostat and humidifier using transfer learning. Thus, we can implement the algorithm in the energy management system of another building, which indicates a high level of scalability of the method.

Conclusions

The proposed multi-agent reinforcement learning approach successfully optimized temperature and humidity control for the single-zone building model. Both temperature and humidity agents demonstrated effective learning convergence and coordination, improving their per-step rewards by 18.9% and 33.7%, respectively. The learned controller maintained indoor comfort close to the rule-based baseline ($\Delta T \approx 0.04$ °C, PMV ≈ 0.45) while reducing energy consumption by nearly 9%. The largest improvement was achieved in humidifier energy savings (−34.4%), showing that the RL agent effectively avoided unnecessary operation under favorable conditions. These findings confirm that data-driven RL-based controllers can achieve energy-efficient comfort management in dynamic building environments, outperforming static rule-based strategies.

Acknowledgment

This study was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (grant No. AP26197509 - Intelligent control system for building energy consumption and occupant comfort using multi-agent reinforcement learning).

REFERENCES

1 Dean, B., et al. Towards zero-emission efficient and resilient buildings. Global Status Report (2016).

2 Razmara, M., et al. Optimal exergy control of building HVAC system. Applied Energy, 156, 555–565 (2015). https://doi.org/10.1016/j.apenergy.2015.07.051

3 Chua, K.J., et al. Achieving better energy-efficient air conditioning – a review of technologies and strategies. Applied Energy, 104, 87–104 (2013). https://doi.org/10.1016/j.apenergy.2012.10.037

4 Maasoumy, M., et al. Handling model uncertainty in model predictive control for energy efficient buildings. Energy and Buildings, 77, 377–392 (2014). https://doi.org/10.1016/j.enbuild.2014.03.057

5 Salakij, S., et al. Model-Based Predictive Control for building energy management. I: Energy modeling and optimal control. Energy and Buildings, 133, 345–358 (2016). https://doi.org/10.1016/j.enbuild.2016.09.044

6 Yang, S., et al. Experiment study of machine-learning-based approximate model predictive control for energy-efficient building control. Applied Energy, 288, 116648 (2021). https://doi.org/10.1016/j.apenergy.2021.116648

7 Crawley, D.B., et al. EnergyPlus: creating a new-generation building energy simulation program. Energy and Buildings, 33(4), 319–331 (2001). https://doi.org/10.1016/S0378-7788(00)00114-6

8 Strachan, P.A., Kokogiannakis, G., Macdonald, I.A. History and development of validation with the ESP-r simulation program. Building and Environment, 43(4), 601–609 (2008). https://doi.org/10.1016/j.buildenv.2006.06.025

9 Salvalai, G. Implementation and validation of simplified heat pump model in IDA-ICE energy simulation environment. Energy and Buildings, 49, 132–141 (2012). https://doi.org/10.1016/j.enbuild.2012.01.038

10 Shrivastava, R.L., Kumar, V., Untawale, S.P. Modeling and simulation of solar water heater: A TRNSYS perspective. Renewable and Sustainable Energy Reviews, 67, 126–143 (2017). https://doi.org/10.1016/j.rser.2016.09.005

11 Koeln, J., et al. Multi-zone temperature modeling and control. In: Intelligent Building Control Systems: A Survey of Modern Building Control and Sensing Strategies. Springer, Cham, 139–166 (2017). https://doi.org/10.1007/978-3-319-68462-8_6

12 Perera, D.W.U., Pfeiffer, C.F., Skeie, N.-O. Control of temperature and energy consumption in buildings – a review. International Journal of Energy & Environment, 5(4) (2014).

13 Sutton, R.S., Barto, A.G. Reinforcement Learning: An Introduction. MIT Press, Cambridge (1998). https://doi.org/10.1017/S0263574799271172

14 Li, F., Du, Y. Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning. In: Deep Learning for Power System Applications. Springer, Cham, 71–96 (2023). https://doi.org/10.1007/978-3-031-45357-1_4

15 Barrett, E., Linder, S. Autonomous HVAC control, a reinforcement learning approach. In: ECML PKDD. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23461-8_1

16 Mocanu, E., et al. On-line building energy optimization using deep reinforcement learning. IEEE Transactions on Smart Grid, 10(4), 3698–3708 (2018). https://doi.org/10.1109/TSG.2018.2834219

17 Li, Y., et al. Transforming cooling optimization for green data center via deep reinforcement learning. IEEE Transactions on Cybernetics, 50(5), 2002–2013 (2019). https://doi.org/10.1109/TCYB.2019.2927410

18 Liu, B., Akcakaya, M., McDermott, T.E. Automated control of transactive HVACs in energy distribution systems. IEEE Transactions on Smart Grid, 12(3), 2462–2471 (2020). https://doi.org/10.1109/TSG.2020.3042498

19 Kazmi, H., et al. Multi-agent reinforcement learning for modeling and control of thermostatically controlled loads. Applied Energy, 238, 1022–1035 (2019). https://doi.org/10.1016/j.apenergy.2019.01.140

20 Blad, C., Bøgh, S., Kallesøe, C. A multi-agent reinforcement learning approach to price and comfort optimization in HVAC-systems. Energies, 14(22), 7491 (2021). https://doi.org/10.3390/en14227491

21 Espeholt, L., et al. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. Proceedings of ICML (2018).
22 EnergyPlus Weather Data. URL: https://energyplus.net/weather (accessed: 2026.06.01).

[1]*Каппарова А.,
докторант, ORCID ID: 0009-0004-4639-3548,
*e-mail: kapparova_ainur3@kaznu.edu.kz
[1]Жоламанов Б.,
докторант, ORCID ID: 0000-0001-8206-7425,
e-mail: zholamanov.batyrbek@kaznu.kz
[1]Болатбек А.,
докторант, ORCID ID: 0009-0004-7613-5507,
e-mail: bolatbek.askhat@kaznu.kz
[1]Құттыбай Н.,
PhD, ORCID ID: 0000-0002-5723-6642,
e-mail: kuttybyy.nurzhigit@kaznu.kz
[1]Досымбетова Г.,
PhD, ORCID ID: 0000-0002-3935-7213,
e-mail: gulbakhar.dossymbetova@kaznu.edu.kz
[1]Жұмағалиев Е.,
магистрант, ORCID ID: 0009-0002-9213-2555,
e-mail: zhumagaliyev_y0@live.kaznu.kz

[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан

СИМУЛЯЦИЯЛАНҒАН ЖЫЛУ АЙМАҒЫНЫҢ БІР БӨЛІГІНДЕГІ ТЕМПЕРАТУРАНЫ БАСҚАРУҒА АРНАЛҒАН IMPALA КӨПАГЕНТТІ ҚҰРЫЛЫМЫН ЖҮЗЕГЕ АСЫРУ

Аңдатпа

Ғимараттар әлемдік энергия тұтынудың айтарлықтай бөлігін құрайды, олардың ішінде жылыту, желдету және ауаны баптау (HVAC), сондай-ақ ылғалдылықты реттеу жүйелері негізгі үлеске ие. Ережеге негізделген дәстүрлі басқару стратегиялары ішкі және сыртқы ортаның динамикалық жағдайларына бейімделе алмайды, сондықтан деректерге негізделген басқару әдістерін қолдану өзекті болып отыр. Зерттеу жұмысында EnergyPlus бағдарламасында модельденген бір зоналы ғимараттағы температура мен ылғалдылықты бір мезгілде басқаруға арналған көпагентті күшейте оқыту (MARL) жүйесін ұсынады. Ұсынылған тәсіл орталықтандырылған оқыту және орталықсыз орындау (CTDE) қағидасына негізделген таратылған Importance Weighted Actor-Learner Architecture (IMPALA) алгоритмін қолданады. Мұнда екі агент – температура және ылғалдылық агенттері – жоғары дәлдіктегі симуляциялық кері байланыс негізінде үйлестірілген саясаттарды үйренеді. Нәтижелер оқытудың жоғары тиімділігін көрсетті: екі агенттің де орташа қадамдық марапаттары айтарлықтай артты (температура бойынша +18.9%, ылғалдылық бойынша +33.7%), бұл олардың тиімді өзара әрекеттесуін және тұрақты үйренуін дәлелдейді. Үйретілген басқару жүйесі ережеге негізделген базалық жүйемен салыстырғанда жылулық жайлылық деңгейін (орташа айырмашылық ≈ 0.04 °C, PMV ≈ 0.45) сақтай отырып, энергияны үнемдеуге қол жеткізді. Жылдық HVAC жүйесінің энергия тұтынуы 8.9%-ға азайды, ал ең үлкен үнем ылғалдандыру энергиясында байқалды – 34.4%-ға төмендеді. Жылыту және салқындату жүктемелері іс жүзінде өзгеріссіз қалды, бұл жайлылық деңгейін төмендетпей-ақ энергия үнемдеуге қол жеткізілгенін растайды.

Түйін сөздер: ғимарат моделі, көпагенттік құрылым, көпагенттік күшейту арқылы оқыту, интеллектуалды басқару, Importance Weighted Actor-Learner Architecture.

[1]*Каппарова А.,
докторант, ORCID ID: 0009-0004-4639-3548,
*e-mail: kapparova_ainur3@kaznu.edu.kz
[1]Жоламанов Б.,
докторант, ORCID ID: 0000-0001-8206-7425,
e-mail: zholamanov.batyrbek@kaznu.kz
[1]Болатбек А.,
докторант, ORCID ID: 0009-0004-7613-5507,
e-mail: bolatbek.askhat@kaznu.kz
[1]Куттыбай Н.,
PhD, ORCID ID: 0000-0002-5723-6642,
e-mail: kuttybyy.nurzhigit@kaznu.kz
[1]Досымбетова Г.,
PhD, ORCID ID: 0000-0002-3935-7213,
e-mail: gulbakhar.dossymbetova@kaznu.edu.kz
[1]Жумагалиев Е.,
магистрант, ORCID ID: 0009-0002-9213-2555,
e-mail: zhumagaliyev_y0@live.kaznu.kz

[1]Казахский национальный университет им. аль-Фараби, г. Алматы, Казахстан

РЕАЛИЗАЦИЯ МНОГОАГЕНТНОЙ АРХИТЕКТУРЫ IMPALA ДЛЯ УПРАВЛЕНИЯ ТЕМПЕРАТУРОЙ В ОДНОЙ ЗОНЕ СИМУЛИРОВАННОГО ЗДАНИЯ

Аннотация

Здания составляют значительную долю мирового энергопотребления, в том числе системы отопления, вентиляции и кондиционирования воздуха (HVAC) и управления влажностью, формируют основную часть их эксплуатационных энергозатрат. Традиционные стратегии управления, основанные на фиксированных правилах, часто не способны адаптироваться к динамически изменяющимся внутренним и внешним условиям, что обуславливает необходимость применения методов управления на основе данных. В данном исследовании представлен фреймворк многоагентного обучения с подкреплением (MARL) для одновременного управления температурой и влажностью в однозонном здании, смоделированном в среде EnergyPlus. Предложенный подход использует распределенную архитектуру Importance Weighted Actor-Learner Architecture (IMPALA) с централизованным обучением и децентрализованным исполнением (CTDE), что позволяет двум агентам – температурному и влажностному – обучаться согласованным стратегиям непосредственно на основе обратной связи высокоточной симуляционной модели. Полученные результаты демонстрируют высокую эффективность обучения: оба агента существенно улучшили среднее вознаграждение на шаге (температура +18,9%, влажность +33,7%), что свидетельствует об успешной сходимости и кооперативном поведении. Обученный контроллер обеспечивает уровень теплового комфорта, сопоставимый с базовой стратегией управления на основе правил (средняя разница температуры в занятый период ≈ 0,04 °C; среднее значение PMV в занятый период ≈ 0,45), при одновременном достижении значительной экономии энергии. Совокупное годовое энергопотребление системы HVAC снижается на 8,9%, при этом наибольшее сокращение было достигнуто по энергии увлажнения на 34,4%. Нагрузки на отопление и охлаждение практически не изменяются, что подтверждает достижение энергосбережения без ухудшения показателей комфорта.

Ключевые слова: модель здания, многоагентная архитектура, многоагентное обучение с подкреплением, интеллектуальное управление, Importance Weighted Actor-Learner Architecture.