 https://doi.org/10.1051/e3sconf/202668903004

Optimization Control of Building Facilities Using Deep Reinforcement Learning Comparison with Dynamic Programming

Yuta Minohara[1*], *Ryozo Ooka*[1], *Shintaro Ikeda*[2*], *Chao Lin*[1] and *Ken Takahashi*[1]

[1]I.I.S., The University of Tokyo, 4-6-1 Komaba, Meguro-Ku, Tokyo, Japan
[2]Mikoto Strategy Co, Kudan Minami, Chiyoda-Ku, Tokyo, Japan

Abstract. In recent years, Artificial Intelligence (AI) has been rapidly advancing worldwide and is being applied in many fields. While its utilization is also expected in the building energy, we need to quantitatively evaluate its effectiveness for full-scale practical implementation. Therefore, in this study, we examined the effect of controlling a photovoltaic power generation system including storage batteries using AI on its power cost reduction by comparing it with a theoretically optimal solution obtained by mathematical programming. More specifically, a comparison was conducted between charge/discharge control using Dynamic Programming (DP) and Soft Actor-Critic (SAC). SAC is a reinforcement learning algorithm known for its high performance in problems with continuous action spaces. In contrast, DP is a method that can obtain a mathematically optimal solution under discretized variable settings based on the principle of optimality. The results showed that SAC-based control achieved a 7.0% to 11.2% reduction in electricity cost compared to the case without utilizing the energy storage system. However, when compared with the theoretical optimum obtained by DP, a maximum cost reduction gap of up to 18% was observed. This may be attributed to the fact that DP considers future demand, whereas SAC focuses solely on the present.

1 Introduction

In recent years, the development of artificial intelligence (AI) has advanced rapidly around the world, and its applications are expanding into many fields. One of the applications is operational optimization of building facilities, which is expected to reduce operating costs and energy consumption. However, for more widespread practical implementation, it is essential to quantitatively evaluate these effects [1].
In Japan, the 6th Strategic Energy Plan outlines a goal to make renewable energy the primary power source, aiming for a 36-38% share of the power generation mix [2]. As part of this initiative, energy storage systems such as battery storage and thermal storage tanks are being promoted. However, due to the variability of electricity demand and renewable power generation, as well as the limited capacity of storage devices, it remains challenging to fully utilize these systems.
Minohara et al. have previously compared reinforcement learning (RL)-based control with the theoretical optimal control derived from dynamic programming in energy system that includes only battery storage, demonstrating the potential for electricity cost reduction [3]. This study investigated the electricity cost reduction effect of a reinforcement learning model for a system including battery storage and thermal storage tanks.

2 Analysis Overview

2.1 Simulation Environment

In previous studies, the optimization control of energy storage systems using artificial intelligence has been examined by Miyano et al. [4] on the CityLearn platform [5]. CityLearn is a platform that enables optimal control of urban energy using reinforcement learning under various constraints and allows simulation using simplified building models. This study aims to quantitatively evaluate the effectiveness of control using reinforcement learning; therefore, the simulation was conducted on the CityLearn platform.

2.2 Analysis Target

In this study, we targeted one standard floor of the building used for heat load simulation, as described in Chapter 5.3 of SHASEG 1008-2016 [6]. We assumed that the building is equipped with a commercial and industrial storage battery with a capacity of 100 kWh, a nominal power of 50 kW, and a round-trip efficiency of 100%. It is also equipped with an air-cooled heat pump packaged air conditioner with a capacity of 980 MJ and a nominal maximum output power of 136 kW.

* Corresponding author: mino0915@iis.u-tokyo.ac.jp

The floor plan of the reference floor and detailed building specifications are shown in Figure 1 and Table 1, respectively.

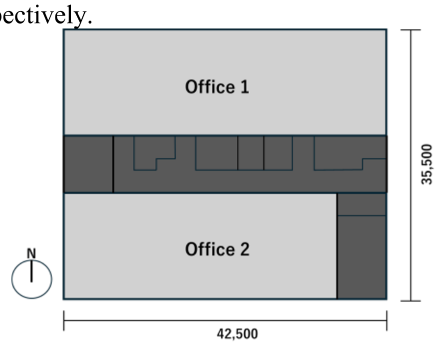


Figure 1. Floor plan (Unit: mm)

Table 1. Target Building Specifications

Weather Conditions	Expanded AMeDAS Typical Meteorological Year (1991–2000,Tokyo)
Region	Tokyo
Usage	Office
Standard Floor Area	1560m²

The building shape, wall composition, and internal heat gains were modeled using OpenStudio, and heat load calculations were simulated using EnergyPlus's Zone-HVAC:IdealLoadsAirSystem. For electricity pricing, we adopted the time of use rates from Tokyo Electric Power Company's Peak Shift Plan (Seasonal Time-of-Use Electricity Pricing with Peak Demand Control), excluding the basic charge (Figure2). In this study, two evaluation periods were used: August 1–3 and August 22–24. Figure 3,4 shows the electricity demand and Cooling load. The Electricity demand was assumed to consist of lighting and office equipment loads, while the Cooling load was assumed to include heat gains from lighting, equipment, occupants, and ventilation. Figure 5,6 show the solar power generation. PV panels were assumed to be installed on the rooftop with a south-facing orientation and a tilt angle of 20°, covering an area of 750 m², which is approximately half of the typical floor area.

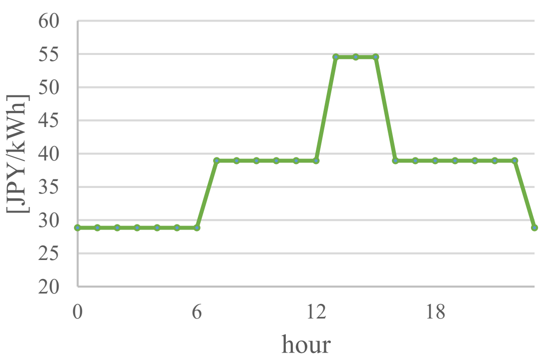


Figure2. Time Series of Electricity price in a day

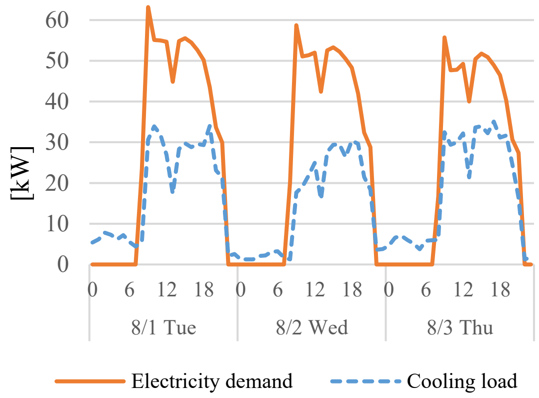


Figure3. Time Series of Electricity Demand, Cooling load (8/1~8/3)

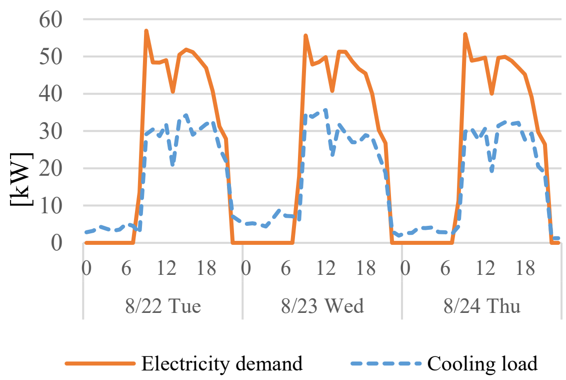


Figure4. Time Series of Electricity Demand, Cooling Load (8/22~8/24)

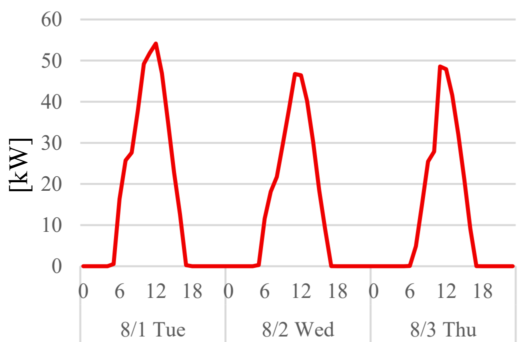


Figure5. Time Series of PV generation (8/1~8/3)

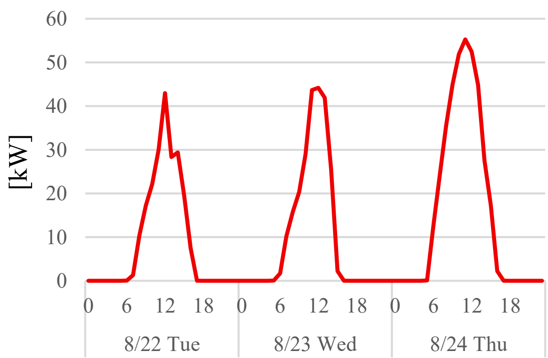


Figure6. Time Series of PV generation (8/22~8/24)

2.3 Energy Storage System and Electricity Cost Calculation Method

In this simulation, the energy storage system is configured as shown in Figure 7. The electricity costs were calculated based on Equations (1) to (3).

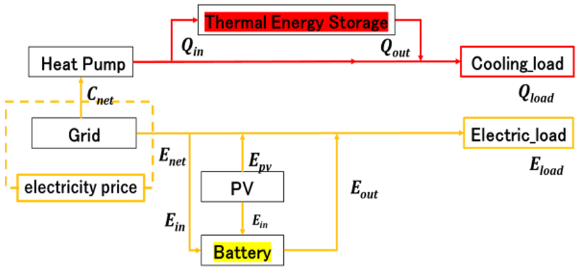


Figure 7. Energy Storage System

$$P = \sum_{t=1}^{T} \max\left(0, \{E_{net}(t) + C_{net}(t)\} \cdot p(t)\right) \tag{1}$$

$$E_{net} = E_{load} - E_{pv} + E_{in/out} \tag{2}$$

$$C_{net} = (Q_{load} - Q_{in/out})/cop \tag{3}$$

$Q_{in/out}$	Thermal charging/discharging power	[kWh]
p(t)	Electricity cost per hour	[JPY/kWh]
E_{load}	Electricity load	[kWh]
E_{pv}	Solar power generation	[kWh]
$E_{in/out}$	Battery charging/discharging power (negative for discharging)	[kWh]
Q_{load}	Cooling load	[kWh]
$Q_{in/out}$	Thermal charging/discharging power (negative for discharging)	[kWh]
COP	Coefficient of performance for cooling	[-]

3 Control Methods

In this study, the states of the battery storage and the thermal storage tank are controlled using two methods: Dynamic Programming (DP) [7], a type of mathematical optimization technique, and Soft Actor-Critic (SAC), a type of machine learning . For comparison, the baseline case is where neither the battery nor the thermal storage tank system is utilized, i.e., no energy storage control is applied [1].

3.1 Dynamic Programming (DP)

DP is a method that can precisely find the optimal solution under discrete variable settings with an optimality principle. Figure 8 shows the optimization algorithm based on Dynamic Programming. The effectiveness of applying DP to the control of energy storage systems for reducing electricity costs has been demonstrated by Ikeda et al. [8]. In DP, the optimal solution can be obtained through $d^{2n}\times T$ calculations, where d is the discrete value setting for each variable, n is the number of variables at each time step, and T is the number of analysis steps(in this study, $n = 2$, as the state of the battery and thermal storage are the variables). In this study, with a discrete value of 2.5%, and the number of analysis steps was set to 24×3=72, as the simulation was conducted hourly over a three-day period, a total of $(100\div2.5+1)^{2\times2}\times72$ = 203,454,792 calculations are required. In general, DP determines the optimal solution based on two phases: backward calculation and forward calculation—based on complete information on electricity and thermal loads, photovoltaic generation, and costs over the entire evaluation period. In the backward calculation, the process moves backward in time from the final time (t=T). and proceeds backward to the initial time step (t = 0), recursively deriving the optimal control policy. In this study, the battery and thermal storage state-of-charge rates were set to zero at t = T. At each time step, the transition cost from the current battery and thermal storage state to all possible states at the previous time is calculated, and only the path with the minimum cumulative cost is retained. This backward approach is based on the principle of optimality, which states that the optimal decision at each time step depends on the optimal policy for subsequent stages. By starting from the terminal cost defined by the energy levels of the battery and thermal storage at the final time, the optimal control for the entire period can be obtained recursively.

Subsequently, a forward calculation is performed starting from t=0. By utilizing the recorded paths, the optimal route from t=0 to t=T is determined, enabling optimal control using DP.

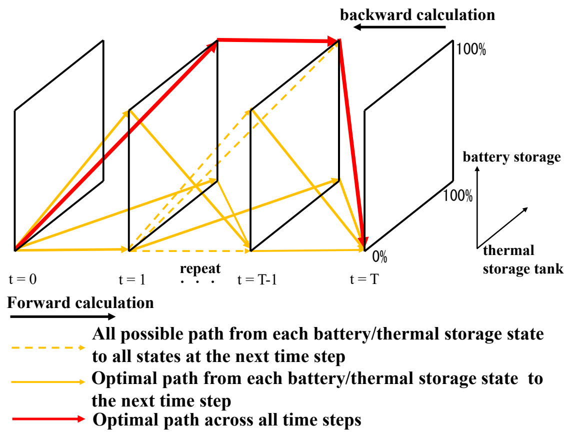


Figure 8. Optimization Algorithm of Dynamic Programming

3.2 Soft Actor-Critic (SAC)

Soft Actor-Critic (SAC) is a deep reinforcement learning algorithm that combines entropy regularization with off-policy learning. By encouraging action diversity while efficiently optimizing the policy, SAC can achieve high performance in complex environments with continuous action spaces. In this study, we set the hyperparameters as shown in Table 2 and used CityLearn's observations (Table 3) as observation variables to control the agent. The reward function, with electricity cost defined as P, is given by Equation (4). To enable the agent to learn a control policy that minimizes electricity costs, the reward function was defined as the negative of the electricity cost. In this way, minimizing the cost is equivalent to maximizing the cumulative reward.

$$Reward = -P \tag{4}$$

The training period was from July 1 to July 30. To evaluate the model's generalization performance, the validation period consisted of two weekday periods: August 1 to 3, and August 22 to 24. Both validation periods were weekdays from Tuesday to Thursday, and there were no significant differences in the temporal patterns of electricity demand, cooling demand, and PV generation. However, there were slight differences in the peak values of demand and generation. These two periods were selected to examine how such differences might affect the control performance of the SAC algorithm.

Table 2. SAC Hyperparameter

Parameter	data
Optimization Method	Adam [9]
Learning Rate	0.0003
Discount Factor	0.99
Replay Buffer Size	1000000
Number of Hidden Layers	2
Number of Neurons	256
Batch Size	256
Activation Function	ReLu
Target Update Coefficient	0.005
Target Network Update Interval	1
Number of Gradient Steps	1
Number of Training Episodes	100,500,1000,2000

Table 3. Observations of CityLearn

Observations	Time of measurement
Month, day, hour	current
Outside temperature [°C]	current 6h_predicted, 12h_predicted
Outside relative humidity [%]	current 6h_predicted, 12h_predicted
Diffuse solar irradiance [W/m²] Direct solar irradiance [W/m²]	current 6h_predicted, 12h_predicted
Inside temperature [°C] Indoor set temperature [°C]	current
Indoor relative humidity [%]	current
Cooling demand[kWh]	current
Non shiftable load[kWh]	current
Solar generation[kWh]	current
Electrical storage rate[kWh/kWh]	current
Cooling storage rate [kWh/kWh]	current
Electricity pricing [JPY/kWh]	current 6h_predicted, 12h_predicted

4 Results and Discussion

Figure 9 shows the reduction rate from the baseline for each control method from August 1–3, and Figure 10 shows the same for August 22–24. From August 1 to 3, electricity costs were reduced by 28.2% under DP control and by 10.8% under SAC control (with n = 1000), resulting in a difference of 17.4 percentage points between the two methods. Similarly, from August 22 to 24, electricity costs were reduced by 22.9% under DP control and by 7.0% under SAC control, with a reduction difference of 15.9 percentage points. Furthermore, the reduction rate achieved by SAC changed by only 1.1% at most when the number of training iterations was varied, suggesting that, under the simulation conditions in this study, the contribution of the number of training episodes to control performance was relatively small.

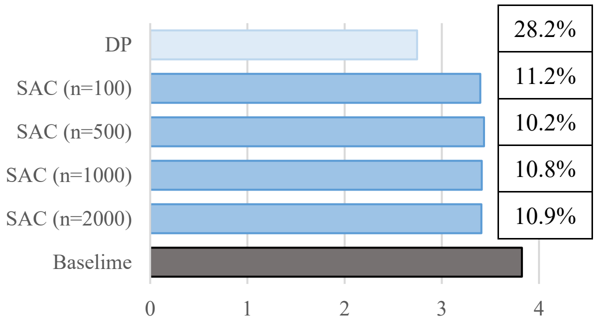


Figure 9. Reduction Rate from Baseline for Each Control Method

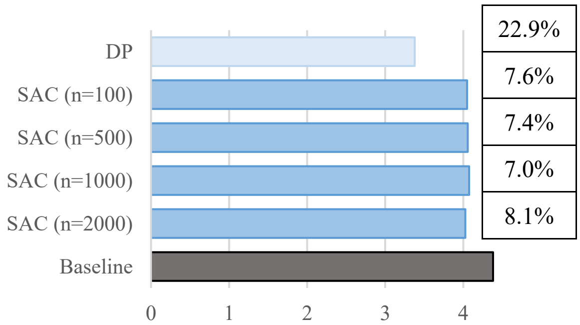


Figure 10. Reduction Rate from Baseline for Each Control Method

Figure 11,12 shows the transition of battery and thermal storage states from August 1–3 for DP and SAC (n = 1000), and Figure 13,14 shows the transition of battery and thermal storage states from August –24 for DP and SAC (n = 1000). Regarding the battery state of charge, the control trend of SAC was generally similar to that of DP on August 1,2, and 24. In contrast, on August 3, 22, and 23, the SAC control deviated significantly from the DP control. On the other hand, with respect to thermal storage state of charge, substantial discrepancies between SAC and DP were observed regardless of the date.

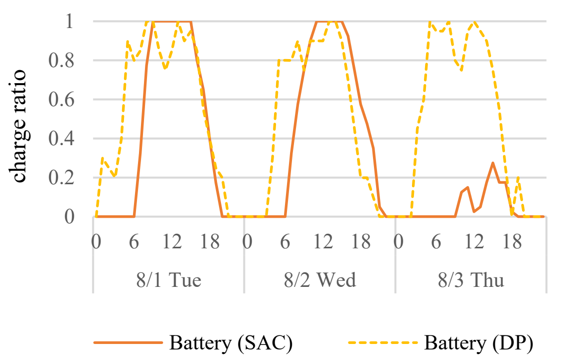


Figure 11. Transitions of Battery Storage by DP and SAC (8/1 – 8/3)

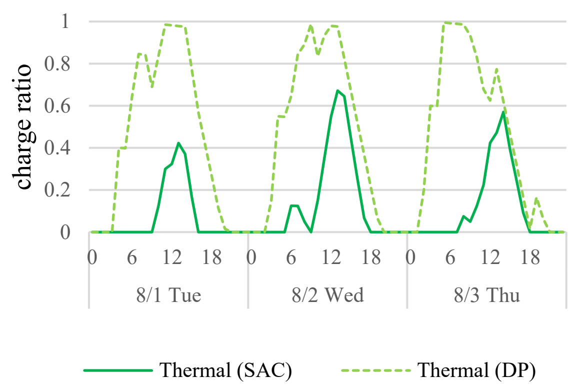


Figure 12. Transitions of Thermal Storage States by DP and SAC (8/1 – 8/3)

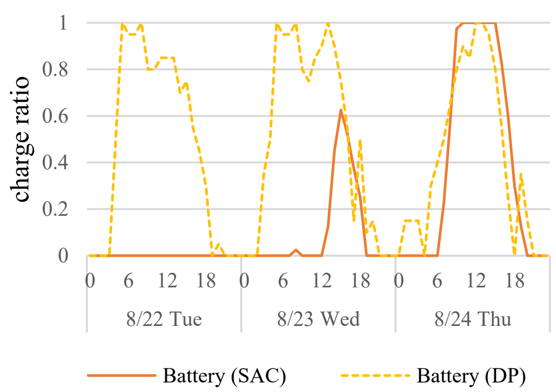


Figure 13. Transitions of Battery Storage States by DP and SAC (8/22 – 8/24)

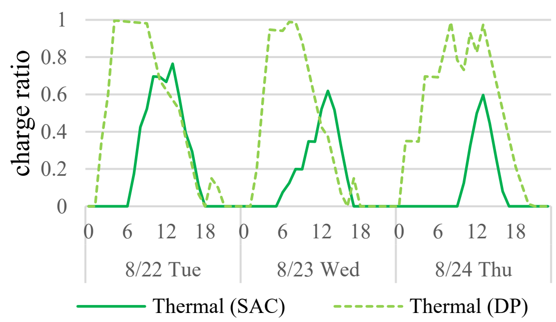
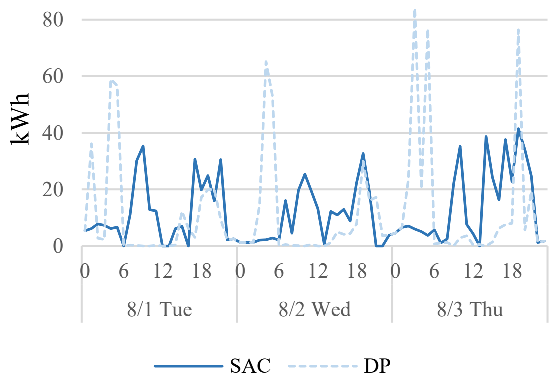
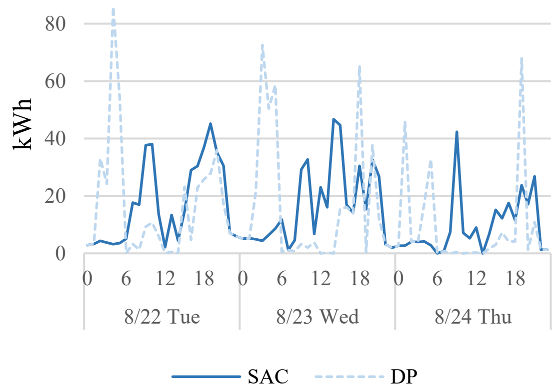


Figure 14. Transitions of Thermal Storage States by DP and SAC (8/22 – 8/24)

Figures 15 and 16 show the amount of electricity purchased from the grid from August 1–3, and from August 22–24, respectively. From both graphs, it can be observed that for DP control, the amount of purchased electricity tends to significantly increase between 1:00 AM and 5:00 AM, which is when the electricity unit price is lowest. Conversely, for SAC control, the amount of purchased electricity tends to increase after 7:00 AM, when the electricity unit price rises. Unlike DP, SAC controls the battery and thermal storage states based only on the current time step. Therefore, it is presumed that the increase in purchased electricity after 7:00 AM—when electricity and cooling demand rise—occurred as a response to those immediate demands.

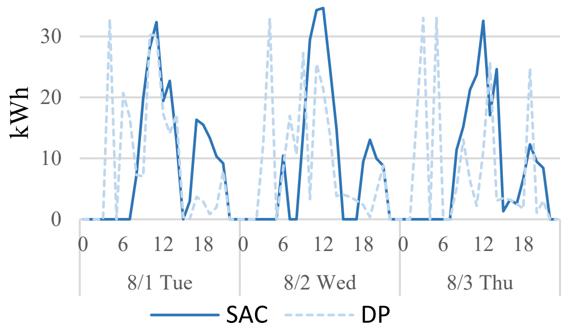


Figures 15. Purchased Electricity from the Grid (8/1 – 8/3)

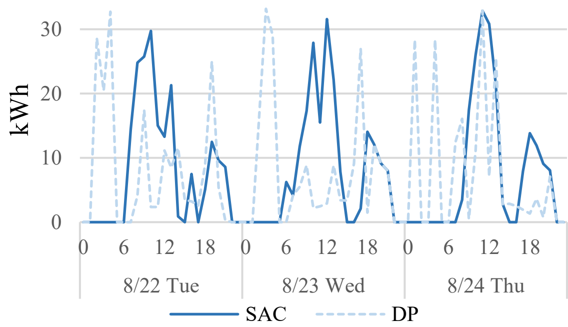


Figures 16. Purchased Electricity from the Grid (8/22 – 8/24)

Figures 17 and 18 show the electricity consumption of cooling equipment during August 1–3 and August 22–24, respectively. Regarding the electricity consumption of cooling equipment, for DP control, there's a tendency for the electricity consumption of cooling equipment to significantly increase between 1:00 AM and 4:00 AM, when the electricity unit price is low. For SAC (n=1000) control, an increase in the electricity consumption of cooling equipment is observed after 6:00 AM. These results suggest that increasing electricity consumption during periods with the lowest electricity prices contributes significantly to reducing total electricity costs.



Figures 17. Electricity Consumption of Cooling Equipment (8/1 – 8/3)



Figures 18. Electricity Consumption of Cooling Equipment (8/22 – 8/24)

5 Conclusion and Future Outlook

This study investigated the change in electricity cost when controlling the state of battery storage and the state of a thermal storage tank using SAC. As a result, we observed a reduction in electricity costs ranging from 7.0% to 11.2% compared to the baseline. However, when compared to the control results from DP, which represents the theoretically optimal solution, the difference between DP and SAC was confirmed to be up to 18%. These results indicate that, under the present simulation conditions, the deep reinforcement learning control could not find the theoretical optimal solution. This may be because, while DP determines control actions based on electricity and cooling demand over the entire evaluation period, SAC makes decisions based only on the current time step. As a result, SAC may have failed to implement a long-term cost reduction strategy. Moving forward, we believe that refining the SAC algorithm and modifying the explanatory variables to enable learning of cost-reduction strategies from a longer-term perspective will lead to improved control accuracy using deep reinforcement learning.

[1] In cases where the electricity generated by photovoltaic (PV) panels exceeded the power consumption, the surplus electricity was assumed to be consumed for purposes outside the scope of the energy system considered in this study.

Acknowledgment
This research was conducted with support from the New Energy and Industrial Technology Development Organization (NEDO) and JSPS KAKENHI Grant-in-Aid for Scientific Research (A), "Development of a General-Purpose Energy System Platform Integrating Digital Information, AI, and IoT." The authors would like to express their sincere gratitude for this support.

References

1. R. Ooka, *Application of AI in building energy management and operation, J. Inst. Electr. Install.* Eng. Jpn. 43, 159-161 (2023)
2. Agency for Natural Resources and Energy, *Future directions for promoting the dissemination of ZEB/ZEH-M* (2023)
3. Y. Minohara, et al., *Optimization control of building equipment using deep reinforcement learning (Part 1): Comparison with dynamic programming for operational cost reduction* Summaries of Technical Papers of Annual Meeting, Architectural Institute of Japan (to be published) (2025)
4. K. Miyano, et al., *Study on explainability of energy optimization control using AI (Part 1): Optimization control for a single building,* Summaries of Technical Papers of Annual Meeting, Architectural Institute of Japan (2024)
5. K. Nweye, et al., *The CityLearn Challenge 2022,* Texas Data Repository (2023)
6. Society of Heating, Air-Conditioning and Sanitary Engineers of Japan, *Guideline for evaluation methods of building energy simulation tools* (2016)
7. T. Suzuki, et al., Summaries of Technical Papers of Annual Meeting, Architectural Institute of Japan, Environmental Engineering II, 1125 (2013)
8. S. Ikeda, et al., *Development of an integrated optimal operation method for batteries, thermal storage tanks, and heat sources using metaheuristics*, J. Soc. Heat. Air-Cond. San. Eng. Jpn. 79, 957 (2014)
9. D. P. Kingma, J. Ba, *Adam : A method for stochastic optimization, in Proc. Int.* Conf. Learn. Represent. (2015)