Comparing model predictive control and reinforcement learning for the optimal operation of building-PV-battery systems

Sicheng Zhan[1], *Yue* Lei[1], and *Adrian* Chong[1*]

[1]Department of the Built Environment, National University of Singapore, Singapore

Abstract. The integration of renewable energy, such as solar photovoltaics (PV), is critical to reducing carbon emissions but has exerted pressure on power grid operations. Microgrids with buildings, distributed energy resources, and energy storage systems are introduced to alleviate these issues, where optimal operation is necessary to coordinate different components on the grid. Model predictive control (MPC) and reinforcement learning (RL) have been proven capable of solving such operation problems in proof-of-concept studies. However, their applications in real-world buildings are limited by the low reproducibility and the high implementation costs. There is a lack of systematic and quantitative understanding of their strength and weakness in actual applications. Hence, this study aims to improve the scalability of optimal control solutions for smart grid operations by comparing MPC and RL regarding their requirements and control performance. We leveraged the CityLearn simulation framework to implement and compare alternative control solutions based on MPC and RL for the energy management of microgrids. In addition to the control performance of cost saving and carbon reduction, other factors such as robustness and transferability were also examined. While both methods achieved promising results, MPC had slightly better performance and could be transferred more smoothly. Given the standardized framework, MPC is more suitable in most cases for the purpose of microgrid operations. However, RL could be preferable for its quickness in making decisions if a large number of energy systems are involved.

1 Introduction

Buildings contribute to over a third of global energy consumption and the corresponding greenhouse gas emissions. In the meantime, building operations possess great potential for energy saving and carbon mitigation [1]. Therefore, the building sector plays an essential role in the campaign for decarbonization, where two primary approaches are replacing the fossil fuel end-use with electricity and integrating more renewable energy resources for power generation [2]. When reducing carbon emissions, the progress in these two aspects presents increasingly notable challenges to the operations of the current power grid.

While the electricity load of buildings is altered and amplified by electrification, the penetration of renewable energy results in higher variability and more irregular patterns of power generation. Consequently, problems such as power outages or curtailment are caused by the mismatch between the supply and the demand side [3]. One pathway to addressing these issues is to form microgrids with buildings and other energy systems. The pressure on the main grid can be relieved by strategically operating these microgrids [4]. The realization of such operations remains a challenge and is the focus of this study.

1.1 Microgrid and energy management

Microgrids connect energy consumers, i.e., buildings, with distributed energy generation and storage systems. PV and batteries are often involved in decarbonization and electrification [5]. Although most microgrids cannot be completely stand-alone, managing the energy systems provides an opportunity to have less dependence and disturbance on the main grid when necessary. For example, over 20% of peak demand can be shaved by adjusting the thermostats [6]; a considerable reduction in operating cost and carbon emission can be achieved by coordinating energy resources and buildings [7].

Demand response is a practical approach to driving the energy management of microgrids, where the grid sends requests or varies utility prices to encourage consumers to adjust their electricity demand. However, this strategy cannot fully realize the potential of microgrids due to the lack of consumers' willingness to fulfill the requests for incentives [8]. The complication of the operations also makes it difficult to exploit the interaction capabilities of the energy systems [9]. Hence, it is necessary to apply automated and optimized control for the energy management of microgrids.

* Corresponding author: adrian.chong@nus.edu.sg

1.2 Optimal control in practice

Predictive control is critical to making optimal decisions for the smart operations of microgrids. For example, the battery charging and discharging should be based on the buildings' electricity load and the PV generation in the coming future, and thermostat adjustment needs to account for the thermal response of the buildings.

Model predictive control (MPC) is a well-established optimal control framework that has been successfully applied in many fields. The core of an MPC framework is a numerical model that represents the energy systems. Based on the predictive model and the forecasted boundary conditions, the trajectory of control actions is optimized at every time step with a receding control horizon. It is able to adapt to various control scenarios for microgrid operations by adjusting the predictive models and the objective functions [10]. Yet, since the control performance highly relies on the model quality, the difficulty in obtaining an adequate model has been hindering the actual application of MPC. Moreover, the modeling effort has to be repeated for every new building because of the heterogeneity in terms of building dynamics and data availability. Considering the time-varying system characteristics, the long-term reliability of predictive models is also of concern [11].

Reinforcement learning (RL) has attracted a lot of research interest in recent years as an alternative optimal control approach. The key advantage of RL over MPC is that it is capable of making control decisions without a predictive model. The control problem is treated as a Markov Decision Process, and the RL agent gradually learns the optimal control policy by interacting with the environment. Another merit of RL that is particularly suitable for microgrid operations is that the multi-agent RL can naturally handle the coordination of different energy systems [12]. However, RL agents typically converge after a long training period. The performance cannot be guaranteed when directly learning from the actual systems, so a large amount of training data is required. There is an option of learning from a simulation model instead of the actual environment, which brings back the challenge of getting a reliable model. Besides, the agent-based control policy also suffers a paucity of interpretability [13]. Thus, the application of RL is also at its infant stage.

In summary, MPC and RL are theoretically capable of providing optimal control for microgrid operations, but both of them are difficult to be implemented in practice. While many studies have tried to address the aforementioned limitations, it is equally important to comprehensively compare the two families to understand their pros and cons. A few comparative studies were conducted recently [14, 15], which only focused on the ultimate performance metrics. Their properties regarding the challenges in actual applications were overlooked and remain unclear. Therefore, further investigation is needed to support the decision of which method to apply in a certain situation.

1.3 Research outline

The electrification of building energy usage and the integration of renewable energy resources led to the emergence of microgrids, which calls for automated and optimized energy management. MPC and RL are two major categories of optimal control solutions with demonstrated potential. MPC is known to require significant modeling effort, and RL typically relies on an arduous tuning and training procedure. Thus far, there is no decisive evidence to select one over the other. Well-defined comparative studies are needed to understand their strength and weakness for actual applications quantitatively. Hence, the objective of this study is to guide the application of optimal control for smart grid operations by comparing MPC and RL regarding their requirements and control performance.

The comparative experiments were conducted using the CityLearn Gym environment, which is introduced in the next section, together with the details about control configurations and experimental design. The results of control experiments are presented from multiple aspects in section 3, followed by the key takeaways discussed in section 4. Lastly, section 5 concludes the study.

2 Design of experiments

Figures 1 illustrates the workflow of the simulation framework, which consists of four main parts: CityLearn environment, training dataset, MPC, and RL. The environment served as a virtual testbed to generate the training dataset and test the control performance. Two test cases were implemented in parallel, where MPC and RL were respectively configured and examined. This section describes the details of these four components.

2.1 CityLearn environment and challenge

CityLearn is an open-source OpenAI Gym environment to standardize the evaluation of optimal control for microgrid operations [16]. The platform and interface were designed to facilitate the implementation of RL algorithms with the state and action spaces specified but can be used for any other control methods including MPC. The environment includes energy models for heat pumps, electrical heaters, heating and cooling storage systems, batteries, and PV panels. The electricity, heating and cooling loads, as well as the thermal response of buildings, were pre-computed with EnergyPlus or measured from actual buildings. More details about the environment can be found in the GitHub repository[†].

The CityLearn challenge is an annual public control strategy competition based on the CityLearn environment. Participants can submit their RL agents with customized reward functions or other control algorithms for online evaluation. For the challenge 2022 [‡], different end-use of building loads are

[†] https://github.com/intelligent-environments-lab/CityLearn

[‡] https://www.aicrowd.com/challenges/neurips-2022-citylearn-challenge

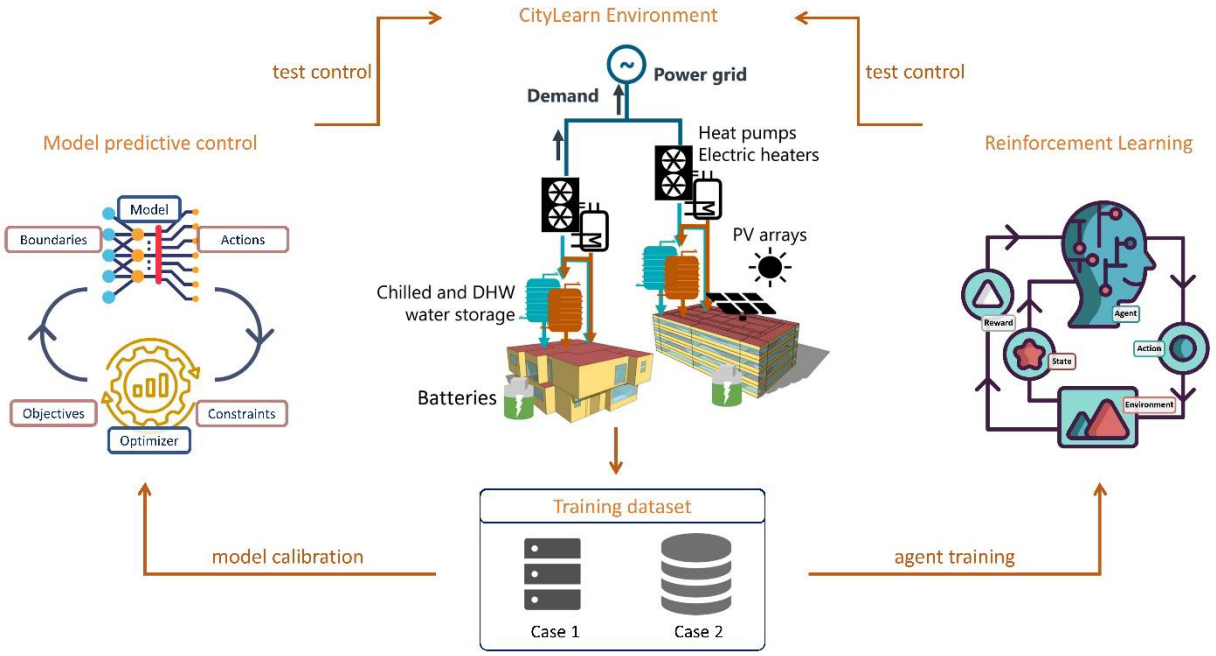

Fig. 1. Components and workflow of the simulation framework.

aggregated into the electricity consumption, and the control actions are the charging and discharging of batteries that come with each building. The goal is to minimize the electricity cost and the CO_2 emissions of the microgrid. The control performance was benchmarked against the operation cost without the energy storage systems, quantified by the ratio of electricity cost and CO_2 emissions over the baseline. Besides the online evaluation results, the simulation was also conducted locally for investigation.

2.2 Two test cases

The first test case was exactly as defined in phase I of the CityLearn challenge 2022. The simulation was based on one year of electricity demand and PV generation data of five single-family houses in Fontana, California. Other boundary conditions included the weather data, electricity price, and carbon intensity during the year. Each house came with a battery and a PV panel, the capacity and nominal power of which were also specified. An ideal situation was assumed in this case, where the controllers were tested on the same dataset that they learned from. While not accounting for the uncertainty of forecasting the future, this test case explored the controllers' capability of achieving optimal performance.

The second test case emulated a more realistic situation by training the controllers with one year of data and testing them with three years of unseen data. The microgrid consisted of nine buildings, including one medium office, one restaurant, one standalone store, one strip mall, and five multi-family buildings. The electricity loads, including heat pump and electrical heater consumption, were simulated using the typical meteorological year weather data of climate zone 5. Each building had a battery, but only four of them were equipped with PV panels. The boundary conditions and metadata were provided in the same way as the first test case.

2.3 Model predictive control

There are many modeling and optimization techniques available for MPC, and a traditional basic MPC framework was used here for comparison. In this case, the coordination between different buildings did not show significant benefit. Therefore, the control action of each building could be optimized independently from each other. The entire MPC framework was implemented in Python, consisting of three main components: an energy model, a forecasting model, and an optimization solver.

As the buildings were incorporated as pre-computed aggregated loads, the only numerical model involved in MPC was for the batteries. The model inputs were the battery's state of charge (SOC), the charging (or discharging) action, and the properties, including the capacity and the nominal power. The maximum allowed power and the energy efficiency were dynamically calculated according to the performance curve, and the battery power was accordingly derived. The power value can be negative, which means the battery is releasing energy. Given the metadata, this model yielded identical predictions as in the CityLearn environment.

Long Short-Term Memory (LSTM), a typical kind of recurrent neural network, was applied to forecast the boundary conditions. The historical data and timestamps of the past eight hours were used to forecast the value for the next hour. Such one-step-ahead forecasting was applied recursively (i.e., outputs appended to inputs for the next time step) twelve times to obtain the forecast throughout the control horizon. For pre-processing, the historical data was standardized, and the timestamps were converted into a cosine signal with a period of 24 hours. The variables to forecast include the electricity

load of each building and the global solar irradiance. The energy price and carbon intensity had regular patterns and therefore were assumed to be perfectly known.

The optimization problem was formulated with an objective function as the weighted sum of electricity cost and carbon emission. The actions were constrained between -1 and 1. The weighting factors q_{cost} and q_{emis} were decided by trial and error with respect to the scale of the two items, which will be further illustrated in the next section. The electricity load of each building was calculated by summing up the predicted battery power, the forecasted non-shiftable load, and the forecasted PV generation. The control horizon was determined as twelve hours based on the preliminary tests. The optimization was solved for each building using Powell's method, which efficiently finds a local minimum for non-differentiable functions. To approach the global optimal solution, the optimization was repeated three times with different simple control rules as the starting point.

2.4 Multi-agent reinforcement learning

Many RL algorithms have been proposed during the past ten years that can be used for microgrid energy management, such as the actor-critic method [17] and the CommNet [18]. Noting the continuous emergence of many new algorithms every year, the one that was designed specifically for the CityLearn Challenge 2020 was selected as the representative: Multi-Agent Reinforcement Learning with Iterative Sequential Action Selection (MARLISA) [19].

MARLISA consists of a group of agents for each building that can exchange information about the reward and the states. The soft actor-critic (SAC) method was adopted as the agents, which learns the optimal actions with an actor network, a critic network, and a value function. Huber loss, layer normalization, and dimension reduction were applied to save computational resources and speed up training. At each time step, the agents took turns selecting actions based on the estimated overall reward and iteratively updated their actions based on other agents' decisions. When conducting deterministic control evaluation, the optimal actions can be identified as the mean of the policy distribution. MARLISA was implemented in PyTorch, and more information can be found in the original paper.

There are many hyperparameters when defining and training MARLISA. The reward function was standardized using the mean and variance of rewards during the random exploration to avoid majorly re-tuning the hyperparameters for different tasks. Thus, most hyperparameters were inherited from the original paper, and a few of them were adjusted according to the new control task, which will be explained later. The reward function was set as the negative weighted sum of electricity cost and carbon emission, different from the original MARLISA reward function but similar to the objective function of MPC.

3 Control experimental results

3.1 Test case I

3.1.1 MPC implementation and behavior

As the first test case assumed perfect information about the future, the forecasting model was not used. The upper image in figure 2 displays the electricity price and the carbon intensity on two consecutive days, which is regular and similar on each day. The carbon emission in the objective function was weighted by three (q_{emis}=3) to counter the larger scale of the price. The optimization was solved immediately at each time step, and the year-long simulation took about an hour to finish.

The lower image in figure 2 shows the resulting net electricity consumption of the microgrid with MPC on these two days, compared with the consumption without the optimal battery control. Essentially, the batteries were charged during the daytime when energy price was low and there was PV generation, and then discharged in the evening to reduce the peak load. Consequently, the net consumption profile was flatter than the baseline, which benefitted the operations of the main grid. The electricity cost and carbon emissions were respectively lowered by 34.8% and 15.8% throughout the year.

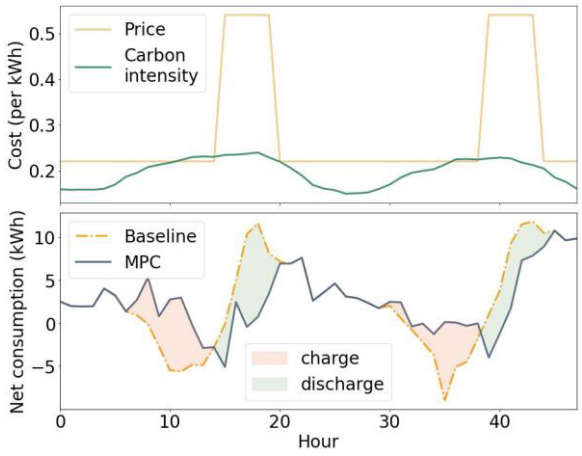

Fig. 2. Energy price, carbon emissions, and net electricity consumption with MPC of the microgrid in test case I.

3.1.2 RL and performance comparison

Similar to MPC, the carbon emission was weighted by three in the reward function. The hyperparameters were optimized to have three hidden layers with 256 neurons and a learning rate of $3e^{-3}$. As shown in figure 3, the RL agent converged after 20 epochs (around 20 hours), and the final CityLearn score was 0.756.

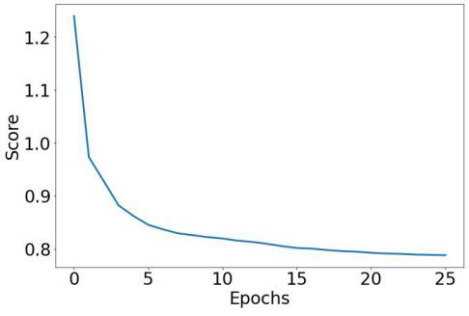

Fig. 3. Learning procedure of the RL agent

Figure 4 plots the control action of one building and the microgrid operating costs on the same two days as in figure 2. MPC and RL are compared with each other, as well as a simple schedule-driven rule-based control (RBC) provided by the CityLearn challenge organizer. It can be seen that the control behavior of MPC and RL are similar and very different from RBC. Instead of continuously charging or discharging the batteries with low power, the optimal control finished the charging cycle in a few hours and stayed idle for the rest of the day. This strategy made better use of the variation of the price and carbon intensity, and resulted in the higher energy efficiency of the batteries.

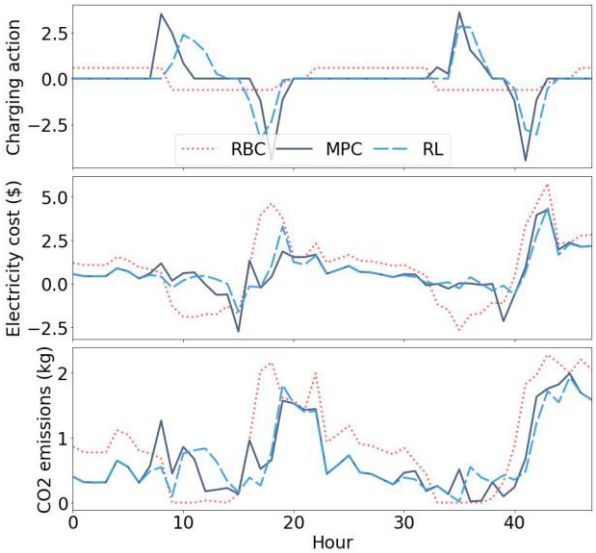

Fig. 4. Comparison of control actions and the resulting performance of MPC, RL, and RBC.

Corresponding to the control actions, MPC and RL had similar profiles of the resulting electricity cost and CO2 emissions, lower and more stable than RBC. Table 1 summarizes the evaluation results of alternative control methods, where the annual control performance is divided by the value without the automated control of batteries. RBC yields higher operating costs than the baseline, especially for carbon emissions. MPC and RL significantly improved the performance with a marginal in-between difference.

Table 1. Control performance comparison in test case I

Metrics	RBC	MPC	RL
Electricity cost	1.033	0.652	0.678
CO_2 emissions	1.156	0.842	0.834

3.2 Test case II

3.2.1 MPC training and testing

The control was first tested using one year of training data, where perfect information about the future was assumed again to exploit the limit of MPC control. The MPC configurations were almost the same as in test case I except for the objective function. The optimal weighting strategy was found to be 2 for the electricity cost and 1 for the carbon emissions. This is because the scale of carbon intensity in the new location was about three times higher than in California. With the larger number of buildings, one year of simulation took over two hours to finish. The evaluation scores were 0.855 for electricity cost and 1.017 for carbon emissions. There was no improvement in carbon emissions due to the lack of regular variation in the carbon intensity. As displayed in figure 5, the decision of charging and discharging was dominated by the electricity price.

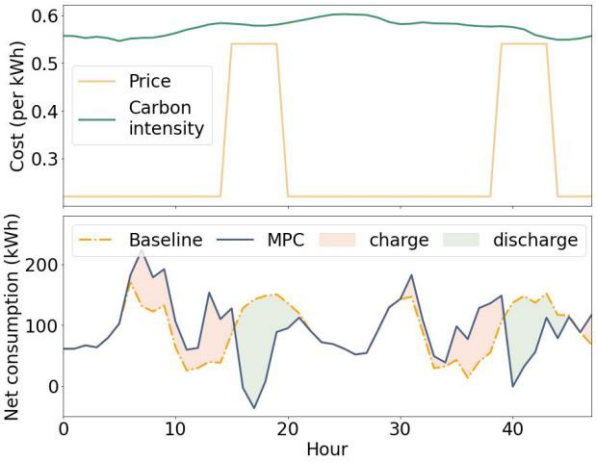

Fig. 5. Energy price, carbon emissions, and net electricity consumption with MPC of the microgrid in test case II.

The first major difference when testing MPC for unseen data is the involvement of forecasting models. Ten LSTM models were trained with one year of data and tested for the following three years. Figure 6 shows the root mean square error (RMSE) over the forecasting horizon of the solar generation and the electricity load of three representative buildings. It was expected to observe that the RMSE grew with the forecasting steps, and even the largest values by the end were at a satisfactory level compared with the literature. The variation of RMSE across buildings was caused by the difference in the scale of the raw data.

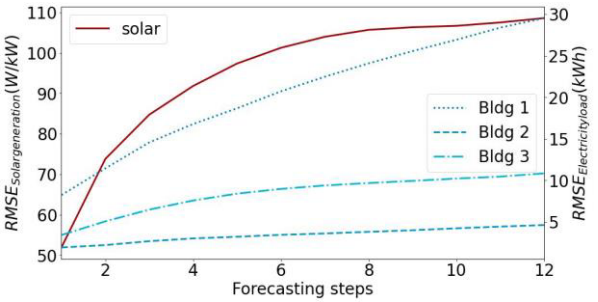

Fig. 6. Multi-step Root mean square error of representative forecasting models.

The forecasted boundary conditions were used when testing MPC for the following three years. This extra step of forecasting and the longer simulation period increased the runtime to more than seven hours. The evaluation results are 0.887 for electricity price and 1.009 for CO_2 emissions, which is slightly deteriorated by the uncertainty brought by the disturbance forecast. Figure 7 takes a closer look at the factors that affect the control performance by visualizing the results of evaluating the monthly electricity costs. It can be seen

that the performance was consistent over the three years and had an apparent seasonality. It was found that the performance was highly correlated to the trend of total electricity load, which was increased in summer by the cooling load.

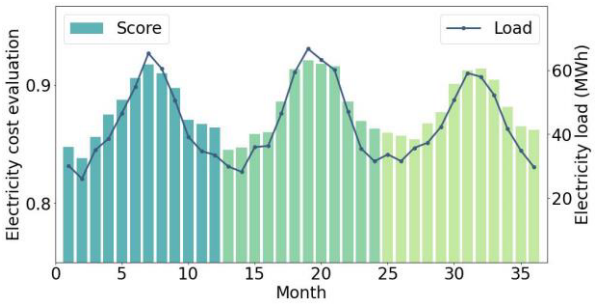

Fig. 7. Monthly MPC performance and total electricity load during the three testing years.

3.2.2 RL results and comparison

The weighting strategy of the reward function in case II again followed the MPC objective function. Another important adjustment of the hyperparameter is the discount factor, decreased to 0.95 to focus more on the coming twelve hours. The training converged to 0.933 for electricity cost and 1.002 for carbon emissions after ten epochs of year-long learning. With higher dimensionality of the state and action spaces, each epoch took longer than in case I, and the training procedure spent over 15 hours.

After training, the RL agents were applied for the operations in the following years. Skipping the online learning by backpropagation, three years of simulation finished within an hour. The performance evaluation yielded 0.949 for electricity cost and 1.006 for carbon emissions. Table 2 summarizes the runtime and electricity cost score during the training (Y1) and testing (Y2-4) stages in test case II. It shows that both MPC and RL extrapolated reasonably well during the testing period. Carbon emissions were omitted as the values were all close to 1.

Table 2. Runtime and performance comparison in test case II

	Runtime (Hr)		Performance	
	Y1	Y2-4	Y1	Y2-4
MPC	2	7	0.855	0.887
RL	15	1	0.933	0.949

The electricity cost reduction of RL was around 50% less than MPC. This was because of the suboptimal control actions spotted in two typical situations. The first situation is presented in figure 8, where the charging actions of one building on a Friday and a Saturday are plotted with the electricity price on these two days. While the control actions were similar on weekdays, the amplitude of the RL charging cycle was significantly smaller than MPC on weekends in response to the smaller peak-valley difference in the electricity price.

The second situation is regarding the heterogeneity across buildings. Figure 9 provides an example by comparing the control actions for two of the nine buildings on two consecutive weekdays. Although charged at different times, the discharging time and scales of RL were similar to MPC in building 7. Meanwhile, the battery of building 8 was only partially utilized by RL in each charging cycle, and the cost-saving potential was not fully exploited. This could probably be related to the difference in PV availability and battery capacity.

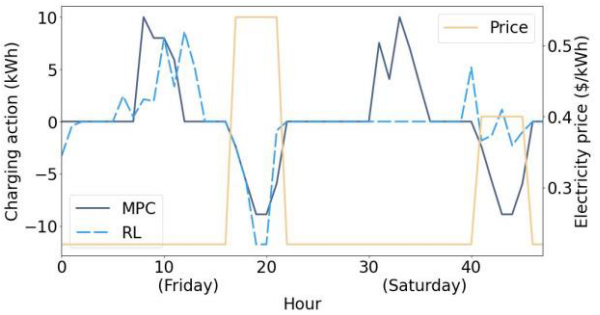

Fig. 8. Charging action comparison of one building by and the electricity price on a Friday and a Saturday.

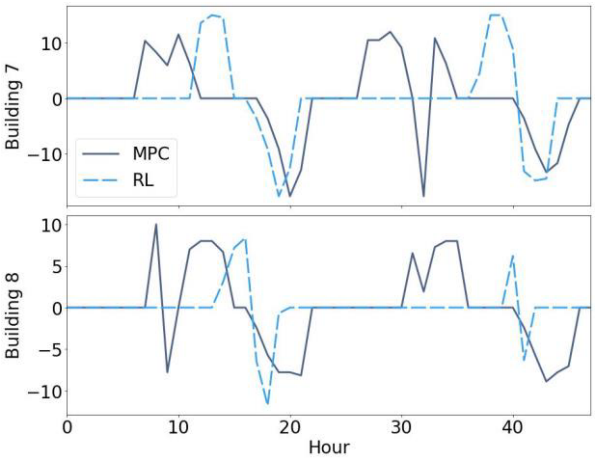

Fig. 9. Charging actions comparison for two of the buildings.

4 Discussion

4.1 Factors affecting control performance

The experimental results revealed several factors that could affect the control performance, which can be categorized into physical systems, boundary conditions, and control algorithms.

First, the characteristics of buildings and other energy systems on the microgrids determined their capability of shifting loads and, thereby, the upper limit of improving control performance. The dominant factor in this CityLearn challenge is the batteries' capacity and nominal power, as well as their relationship with the buildings' loads. In test case I, the ratios of battery nominal power over building average electricity load for the five buildings were 4.14, 4.68, 6.11, 4.06, and 4.97. In contrast, these numbers were 1.49, 2.21, 1.07, 1.95, 1.18, 0.44, 0.59, 0.61, and 0.91 in test case II. This ratio serves as an indicator of the relative capability of shifting loads. As a result, the saving in electricity costs in case II, at its best, was less than half of the savings in case I. In other control scenarios, the primary factor

could be the profile of non-shiftable loads [10] or the building thermal characteristics [20].

The boundary conditions refer to the weather conditions and the electricity costs, including carbon intensity. For example, the carbon intensity in case II had little to optimize. Therefore, the short-term load shifting was driven by the price, and the carbon emissions stayed the same. Also, the seasonality of monthly evaluation results in figure 7 corresponds to the heating and cooling seasons. As the amount of cost savings was relatively stable throughout the years, the total cost drastically increased in the summer. The larger denominator reduced the saving percentage, yielding larger evaluation scores.

Ultimately, the realized performance was decided by the control algorithm. Noting that both MPC and RL can be further fine-tuned to achieve better performance, the experiments represented the typical procedures of implementing optimal control with limited time and resources. In general, MPC has a more solid theoretical basis and a more intuitive workflow. With more details to be discussed in 4.2, MPC is the method of choice in most cases. Besides, it is worth noting that the selection and tuning of algorithms are also affected by the physical systems and boundary conditions. For example, it is always necessary to customize the objective or reward functions based on the control scenarios.

4.2 MPC versus RL from different aspects

When it comes to microgrid operations, the traditional impressions regarding the limitations of MPC and RL do not always apply. For the modeling difficulty of MPC, it is usually the complicated building dynamics that are of concern. In the CityLearn environment and most grid-scale applications, buildings are simplified as time-series load data. Compared with the thermal response of buildings, the dynamics of other energy systems are easier to account for, given the better-defined specifications and boundary conditions. Hence, the model discrepancy was not a problem in the test cases. As pointed out in [21], the modeling requirements can be relaxed when optimizing load management. The ease of obtaining adequate models also helps alleviate the high data requirements of RL. A reliable simulation environment can be established with a short period of data, where the agents can be trained for as many episodes as needed.

Transferring the MPC solution to a new control scenario is straightforward. Apart from adjusting models based on the metadata, the objective function needs to be updated according to the control aim and the characteristics of data (relative scale in this case). It is not difficult to at least obtain a near-optimal solution. On the other hand, RL-based solutions can hardly be directly applied to a new test case. The hyperparameters need to be intensively fine-tuned, which is typically guided by experience. In addition to those mentioned in 3.2.2, the selections of learning rate and exploration duration are also critical to control.

Another potential issue of RL that is likely to happen in the context of microgrid operations is heterogeneous agents [22]. Should the buildings be very different in terms of load scales or patterns, sharing information between agents could lead to suboptimal actions, such as illustrated in figure 9. The agent architecture and the training procedure should be carefully designed to resolve this issue.

One scenario where RL could be preferable is when there are a large number of systems to coordinate, which could make the optimization of MPC intractable. Although the RL training is prone to take long, trained agents could derive the control actions instantly. Yet, implementing MPC in practice may suffer from the increased runtime when many systems are involved in the microgrid, which could be more severe if the decisions are expected to be made in seconds.

Disregarding the implementation cost, both methods could be further improved to pursue the globally optimal control. For example, MPC with a time-varying horizon may reduce the carbon emissions in case II, and the hyperparameters of RL could be optimized through systematic approaches such as Bayesian optimization. However, MPC can be implemented more smoothly to obtain near-optimal control decisions for microgrid operations in a practical setting.

5 Conclusion

In this study, MPC and RL were comprehensively investigated in a standardized simulation framework for microgrid operations. To sum up, both methods achieved promising performance and extrapolated well in the out-of-sample tests. While the difference in performance was marginal, MPC comes with a more standardized framework and can quickly adapt to a new control task. Therefore, unless optimal control is needed instantly for large scale systems, MPC is the more suitable solution.

Acknowledgement

This work is supported by SinBerBEST (Singapore-Berkeley Building Efficiency and Sustainability in the Tropics) research program 2.0 (Project No. A-0005078-09-00).

References

1. C. Carraro, A. Lanza, M. Tavoni, Rev. of Environ. Energy Econ. (2014)
2. B. D. Leibowicz, C. M. Lanham, M. T. Brozynski, J. R. Vázquez-Canteli, N. C. Castejón, Z. Nagy, Appl. Energy, **230**, 1311 (2018)
3. Q. Cui, L. He, G. Han, H. Chen, J. Cao, Appl. Energy, **267**, 115114 (2020)
4. M. W. Khan, J. Wang, Renewable Sustainable Energy Rev., **80**, 1399 (2017)
5. M. A. Jirdehi, V. S. Tabar, S. Ghassemzadeh, S. Tohidi, J. Energy Storage, **30**, 101457 (2020)
6. K. Zhang, M. Kummert, Build. Simul., **14**, 1439 (2021)

7. G. R. Aghajani, H. A. Shayanfar, H. Shayeghi, Energy Convers. Manage., **106**, 308 (2015)
8. B. Zeng, G. Wu, J. Wang, J. Zhang, M. Zeng, Appl. Energy, **202**, 125 (2017)
9. J. Wang, H. Zhong, Z. Ma, Q. Xia, C. Kang, Appl. Energy, **202**, 772 (2017)
10. H. Fontenot, B. Dong, Appl. Energy, **254**, 113689 (2019)
11. E. Atam, L. Helsen, Renewable Sustainable Energy Rev., **54**, 1653 (2016)
12. J. R. Vázquez-Canteli, Z. Nagy, Appl. Energy, **235**, 1072 (2019)
13. K. Nweye, B. Liu, P. Stone, Z. Nagy, Energy AI, **10**, 100202 (2022)
14. J. Arroyo, F. Spiessens, L. Helsen, Front. Built Environ., **8** (2022)
15. A. Mugnini, F. Ferracuti, M. Lorenzetti, G. Comodi, A. Arteconi, Energy Convers. Manage.: X, **15**, 100264 (2022)
16. J. R. Vázquez-Canteli, S. Dey, G. Henze, Z. Nagy, arXiv preprint arXiv:2012.10504 (2020)
17. T.P. Lillicrap, J.J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra. arXiv preprint arXiv:1509.02971. (2015)
18. S. Sukhbaatar, R. Fergus, Adv. NeurIPS, **29** (2016)
19. J. R. Vázquez-Canteli, G. Henze, Z. Nagy, Proceedings BuildSys '20, 170 (2020)
20. S. Zhan, B. Dong, A. Chong, Energy Build., **278**, 112606 (2023)
21. S. Zhan, A. Chong, Renewable Sustainable Energy Rev., **142**, 110835 (2021)
22. T. T. Nguyen, N. D. Nguyen, S. Nahavandi, IEEE Trans. Cybern., **50(9)**, 3826 (2020)