Reinforcement Learning-Driven Adaptive Control of Complex Systems: A Policy Learning Framework for Long-term Sustainability

1st Lee Zong Han
Universiti Putra Malaysia
Serdang, Selangor, Malaysia
huangziyin0925@gmail.com

2nd Chia Shi Han
Universiti Utara Malaysia
Kedah State, Malaysia
warma8457@gmail.com

3rd Lau Yen Ling
Universiti Malaysia Sarawak
Kuching, Sarawak, Malaysia
zhaosuqiu1828@gmail.com

***Abstract*—Smart microgrids are important complex systems in the shift towards sustainable energy due to the double challenge of global climate change and resource depletion but they struggle to work sustainably over time due to the uncertainties of renewable energy variability, load uncertainty and degradation of storage. Conventional control methodologies including Model Predictive Control (MPC), conventional heuristic rule-based control methods, usually demand accurate system models or manual rules that might restrict their flexibility during changing operating conditions. The paper presents a low-cost and repeatable adaptive control policy learning scheme based on Deep Reinforcement Learning (DRL) to manage smart microgrids in the long-term sustainability-based approach. As a formulation of the energy management problem as a Markov Decision Process (MDP) with safety constraints and a lightweight proxy of storage degradation, the framework will use the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm to optimize continuous control actions. This research does not depend on physical deployment of microgrids or expensive laboratory experiments; instead, it makes use of publicly available hourly time-series data of photovoltaic (PV) generation, wind power, storage states, dynamic load and time-varying cost signals in a purely software simulation setting. Comparative simulation outcomes reveal that the proposed Sustainable-TD3 framework is able to minimize the simulated full operational cost without compromising the supply-demand balance and reducing the proxy cost of storage degradation of energy storage devices. The findings indicate that a storage-health-conscious reward design may enhance the trade-off between economic efficiency and long-term sustainability in normal computing operations. This paper offers a repeatable and minimal-resource-reference model of adaptive control research in smart energy systems and encourages the continued development of sustainable scheduling plans in zero carbon communities and sustainable infrastructure.**

Keywords—Deep Reinforcement Learning, Adaptive Control, Long-term Sustainability, Smart Microgrid, Energy Management

I. INTRODUCTION

In the context of the world-wide quest towards carbon neutrality goals, smart microgrids have become more and more important as the backbone infrastructure to integrate distributed renewable power generation, energy storage devices, and adaptable loads [1]. Smart microgrid is not just a physical energy network but a typical complex system that combines technical design, commercial operation, and user behavior. To realize effective resource scheduling in these very dynamic and unpredictable conditions are closely linked to the effectiveness in practice and the long term socio-economic advantages of energy transition. The International Energy Agency (IEA) states that the building and industrial sectors represent more than 65% of global final energy consumption, with about 30% of the energy wasted due to ineffective scheduling and control practices [2]. It is thus of great practical importance to create intelligent and adaptive energy management systems to enhance sustainable development.

In spite of their enormous potential, smart microgrid control systems have some fundamental pain points caused by non-stationary conditions: the unstable energy output of renewable sources (such as wind and solar power), the random nature of the user load, and the non-linear deterioration of energy storage units (such as lithium-ion batteries) due to the cyclic process of charging and discharging. Real-time supply and demand equilibrium, together with minimization of long-term operational cost and maximization of equipment life span is an extremely important scientific challenge that requires immediate attention.

In the past, the control of microgrid energy has been very much dependent on Rule-Based Control (RBC) or Model Predictive Control (MPC) [3]. Over the last few years, Reinforcement Learning (RL), which is a potent model-free adaptive learning approach, has been extensively used in controlling microgrids and managing building energy, and this is because of the development of artificial intelligence [4]. Nonetheless, many of the current research works on RL-based energy management only aim at maximizing short-term economic gains or prompt response capacity, but do not consider the long-term costs of degrading the components of the system, especially energy storage systems [5]. Moreover, most of the RL models have no rigorous safety constraint mechanisms at the exploration stage, and thus there is a possibility of exceeding the physical operating limits when implemented into real-life complicated systems [6].

In order to overcome these shortcomings, this paper suggests an adaptive control policy learning model based on reinforcement learning which specifically aims at the long term sustainability of smart microgrids. This framework links short-term control optimization with long-term asset-aware scheduling by combining a lightweight equipment-degradation proxy and multi-objective safety constraints into the RL reward mechanism. This research will be confined to a software-only islanded microgrid benchmark and not involve any deployment of physical microgrid, hybrid storage devices scaled down to laboratory size, or the topological

Corresponding Author: Lau Yen Ling, Universiti Malaysia Sarawak
Jalan Datuk Mohammad Musa, Sarawak, MALAYSIA, 94300,
zhaosuqiu1828@gmail.com

restructuring of large scale cross-regional transmission and distribution networks.

The rest of the paper is structured in the following way: Section 2 is the literature review and research gap identification. The proposed DRL framework and the microgrid model are discussed in detail in Section 3. The data sources and preprocessing methods are introduced in Section 4. Experimental results and performance comparisons are presented in Section 5. The detailed discussion is given in Section 6. Finally, Section 7 summarizes the paper and discusses the further direction of research.

II. Literature Review

A. Limitations of Traditional Control Methods for Complex Systems

The traditional control strategies in the areas of microgrid and smart building energy management are mainly classified into heuristic methods and optimization-based methods. The heuristic methods (e.g., state machine control, fuzzy logic control) have minimal computational complexity and can be used easily in engineering but their rules are pre-set by people and thus cannot respond to changing environmental conditions and cannot be used to schedule globally [7]. Multi-constrained problems are excellently addressed by optimization-based methods especially MPC which solves a sequence of control actions at a future rolling horizon [8]. However, MPC is strongly based on correct mathematical models of the system and exact predictions of future disturbances. High-precision forecasting models are not feasible in typical research environments when non-stationary conditions are present, including extreme variations in wind and solar power. Besides, the solution of non-linear optimization problems of significant scale frequently suffers the curse of dimensionality restricting its use in low-resource real-time adaptive control problems [9]. Also, the computation time of conventional optimization algorithms such as mixed-integer programming is exponential to the size of the system, so it is hard to satisfy the demands of real-time decision-making.

B. Application of Reinforcement Learning in Energy System Control

Deep Reinforcement Learning (DRL) has become an optimization tool of complex system control based on data and without a model due to the limitations of traditional methods. DRL is capable of identifying the best control policies by interacting with the environment continuously, without the requirement to know the exact physical models [10]. The use of DRL in smart building energy management was reviewed systematically by Yu et al. [11], and it was emphasized that Deep Q-Network-based approaches (DQN) could be highly effective in optimizing the energy usage of Heating, Ventilation, and Air Conditioning (HVAC) systems. In order to manage energy resources in large scale smart buildings safely, Sun et al. [12] suggested a multi-agent DRL model to schedule the distributed energy assets collaboratively. Adaptive control and reinforcement learning were thoroughly examined by Annaswamy [13], who observed that the two had supplementary benefits when it came to managing parameter uncertainty. These researches show that DRL can be used successfully and is better at solving non-linear, high-dimensional problems related to energy management.

C. Research Gaps in Long-term Sustainability and Safety Constraints

However, despite the fact that DRL is already far advanced in the process of enhancing the efficiency of the short term system, there are clear drawbacks in implementation situations that focus on long term sustainability. To begin with, many of the existing models merely state that the reward should be minimum electricity bill or instantaneous energy consumption, and do not consider long-term degradation of energy storage devices (e.g., battery capacity fade, cycle life loss) as part of optimization goals [14]. As energy storage systems represent more than 40 percent of the overall lifecycle cost of a microgrid, neglecting degradation expenses causes control policies to promote high frequency deep charging and discharging, which considerably decreases the life of equipment and contradicts the initial purpose of sustainable development [15]. Thirdly, DRL algorithms also display randomness at the exploration stage, and it is hard to guarantee that control actions will comply with physical safety requirements (e.g., the upper and lower limit of battery State of Charge (SOC), power ramp rate limits) [16]. An adaptive safety-certified reinforcement learning approach was introduced by Zhang and Yang [17], which satisfies the constraints through control barrier functions; nonetheless, their study subject was robotics and had not been applied to the field of energy management.

D. Interdisciplinary Methodological Reference and Research Positioning

The originality of this work can be seen in its interdisciplinary modification of Safe RL concepts [18] and Asset Lifecycle Management concepts [19] to design a policy learning model that explicitly measures long-term degradation costs plus hard constraint penalties. Using the viewpoint of interdisciplinary design innovation, the present paper strongly combines the engineering control theory, data science and sustainable development economics to address the most significant weakness of the general DRL ignoring the lifespan of equipment. It offers an innovative way to reach the best possible equilibrium between economic efficiency and sustainability in complex systems. In comparison to what has already been researched, this paper makes three main contributions: (1) it presents the first application of battery degradation cost model, which is based on the rainflow-counting algorithm to microgrid DRL control; (2) it creates a multi-objective reward function that considers economic efficiency, safety and long-term sustainability; (3) it validates the effectiveness and strength of the framework using systematic comparative experiments.

III. Methodology

A. Research Strategy

The technical strategy used in this research is: lightweight modeling and reproducible simulation validation. Initially, an abstract mathematical model of smart microgrid energy management is developed using a Markov Decision Process (MDP) that converts long term sustainability objectives into particular reward values and constraining conditions. Next, it presents the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm as a model training method in a lower computation environment. Lastly, the policy is tested and compared via a simulation system that runs in software only with constant random seeds and explicit parameters.

Figure 1 shows the overall architecture of the framework. The smart microgrid system model is described as follows.

In this paper, the microgrid system is deployed as a software-only benchmark model consisting of four key modules: photovoltaic (PV) generation, wind turbine (WT) generation, a lithium-ion battery storage unit with an optional abstract reserve-storage channel denoted as H_2, and local dynamic load. Only the H_2 term is maintained as a generic secondary storage variable in the simulator; therefore, no physical hydrogen equipment or laboratory-scale hybrid storage platform is required.

The power balance equation constraint is as follows:

$$P_PV(t) + P_WT(t) + P_batt(t) + P_{H^2}(t) = P_load(t) \quad (1)$$

Here, $P_PV(t)$ and $P_WT(t)$ are the power generated by PV and wind at time t respectively, and $P_load(t)$ is the local load demand. The charge/discharge powers of the battery and the abstract reserve-storage channel, $P_batt(t)$ and $P_{H^2}(t)$, respectively, have positive values indicating discharging and negative values indicating charging.

Dynamics of State of Charge Battery: The discrete recursive equation that describes the evolution of the battery SOC over time is:

$$SOC(t+1) = SOC(t) - [P_batt(t) \times \Delta t]/[E_cap \times \eta] \quad (2)$$

Where E_cap is the rated capacity of the battery (kWh), η is the charging/discharging efficiency, and Δt is the time step.

Battery Degradation Model: This paper provides a simple empirical proxy to quantify long-term sustainability with regular reproducible conditions instead of using a high fidelity electrochemical or rainflow counting model. In accordance with the overall battery degradation concerns raised in the literature [14], [15], we define the degradation cost $C_{deg}(t)$ of the battery at time t as a non-linear function of charge/discharge intensity and the state of charge:

$$C_deg(t) = \alpha \times |P_batt(t)|^{\beta} \times \exp[-\gamma \times SOC(t)] \quad (3)$$

where $\alpha = 0.02$, $\beta = 1.5$, and $\gamma = 0.8$ are fixed proxy coefficients used for reproducible simulation comparison. The model is not intended to substitute complex electrochemical aging models, but rather, provides two general trends of degradation in a lightweight form: high power charging/discharging causes more degradation pressure and low SOC states increase the chances of degradation.

B. Formalization of the Reinforcement Learning Framework

The energy management problem is formulated as an MDP described by the tuple $\langle S, A, R, \gamma \rangle$:

State Space $\mathcal{S}$: It includes both the environmental data and the state of the system at time t represented by a continuous vector of six dimensions:

$$s_t = [P_PV(t), P_WT(t), P_load(t), SOC_batt(t), SOC_{H^2}(t), \lambda(t)] \quad (4)$$

where $\lambda(t)$ is a time-dependent cost signal; in cases where local electricity price data are unavailable, it may be substituted by a normalized public tariff profile or a user-defined benchmark cost sequence.

Action Space $\mathcal{A}$: The agent's output is a two-dimensional continuous control variable, defined as:

$$a_t = [P_batt(t), P_{H^2}(t)] \quad (5)$$

Representing the charge/discharge power commands for the energy storage devices, bounded within P_{max}.

The reward function design is the main novelty that supports long-term sustainability. The overall reward r_t consists of three components: economic operating cost $C_eco(t)$, equipment degradation cost $C_deg(t)$, and a safety-constraint violation penalty $P_vio(t)$:

$$r_t = -[w_1 \times C_eco(t) + w_2 \times C_deg(t) + w_3 \times P_vio(t)] \quad (6)$$

Where $w_1 = 0.4, w_2 = 0.5, w_3 = 0.1$ are constant values of weights obtained using normalization-based scaling and a minor ablation test without requiring a significant amount of grid search. Economic cost is given by:

$$C_eco(t) = \max[0, P_load(t) - P_PV(t) - P_WT(t) - P_batt(t)] \times \lambda(t) \quad (7)$$

The safety constraint penalty term is defined as:

$$P_vio(t) = max[0, SOC_min - SOC(t)] + max[0, SOC(t) - SOC_max] \quad (8)$$

Where $SOC_{min} = 0.1$ and $SOC_{max} = 0.9$.

C. Policy Optimization Algorithm: TD3

Since the control actions of the microgrid (charging/discharging power) are continuous, and the environment is highly uncertain, this paper takes the TD3 algorithm as a policy optimizer. TD3 is a more stable form of Deep Deterministic Policy Gradient (DDPG) algorithm, based on three main mechanisms to increase stability in learning:

Clipped Double Q-learning: It is based on two independent Critic networks Q_{θ_1} and Q_{θ_2}, which use the minimum value as the target Q-value and it is an effective way of reducing overestimation bias.

Target Policy Smoothing: Clipped Gaussian noise is added onto the target action during target Q-value calculation so that the policy does not overfit to narrow peaks of Q-function.

Delayed Policy Update: Actor network is updated less often than the Critic networks (1 Actor update every 2 Critic updates) to make sure that the value estimates will have had enough time to converge before updating the policy.

In terms of network architecture, the actor and critic networks follow similarly lightweight fully connected structures with two hidden layers of 64 neurons each. Hidden layers have ReLU activation functions, and the Actor output layer has a Tanh activation function in order to fit the normalized action space limits [-1, 1]. The experience replay buffer capacity is limited to 5×10^4 to ensure that the experiment can be repeated when using standard computers, the size of the mini-batch is restricted to 64, and the learning rate is restricted to 1×10^{-4}. The discount factor is taken as $\gamma = 0.99$ and the soft update coefficient of the target network

is $\tau = 0.005$, and all runs are performed with constant random seeds.

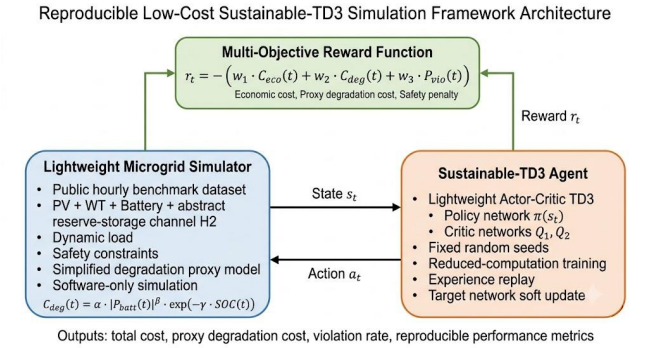

Fig. 1. Reproducible Low-Cost Sustainable-TD3 Simulation Framework Architecture

IV. Data

A. Basic Data Information

This study data is structured into a publicly available hourly benchmark that has PV generation, wind generation, load demand, and time-varying cost data. The benchmark has 8760 hourly records which could be replicated with no need to deploy sensors in the field, proprietary sensor networks or laboratory microgrid setupsTable I presents the descriptive statistics of the core variables. The final version of the manuscript should indicate the data access link, preprocessing script, random seeds, and train-test split to ensure that the research is reproducible.

B. Data Preprocessing Methods

In order to conform with the input requirements of deep neural networks, the raw data was subjected to the following preprocessing operations:

Dealing with Missing Data: Reproducible public time-series processing involves the use of linear interpolation to deal with missing entries when they exist. Proportion and location of missing data must be reported along with the preprocessing program to guarantee that the same clean dataset could be re-created.

The 3 sigma rule was used to identify outliers in the public benchmark time series. Outliers identified were substituted with a weighted mean between neighboring time steps, and the amount and index of smoothed points should be registered in order to have the preprocessing process fully repeatable.

The Min-Max scaling technique was used to scale every state variable to the interval of [-1, 1]:

$$x_norm = 2 \times [(x - x_min)/(x_max - x_min)] - 1 \quad (9)$$

TABLE I. Descriptive Statistics of Microgrid Operational Data

Variable	Mean	Std Dev	Median	Min	Max	Unit
PV Power	2.85	1.92	2.6	0	6.5	kW
Wind Power	3.1	1.75	3.05	0.2	7.2	kW
Load Demand	4.75	2.1	4.6	1.1	9.8	kW
Battery SOC	0.52	0.18	0.5	0.1	0.95	-
Cost Signal	0.68	0.25	0.65	0.3	1.2	CNY/kWh

V. Results

A. Experimental Setup and Baselines

In order to confirm the effectiveness of the proposed TD3 framework with regard to degradation costs (further called Sustainable-TD3), three baseline strategies were developed to compare them with:

The Rule-Based Control (RBC) approach uses elementary threshold logic where it is based on the use of renewable energy, discharge the battery if there is low power and charge when excessive power is available, without regard to price signals.

The Standard Deep Q-Network (Standard DQN) discretizes the action space in 11 levels. The reward function only takes into account the economic cost incurred, and does not consider the degradation cost.

Standard TD3 (Standard TD3): It uses a continuous action space, however, the reward function is the same as Standard DQN, except that it does not have the long-term degradation penalty term.

Each of the DRL models was trained on 800 episodes under the low-budget protocol using early stopping when the moving-average reward deviated by less than 1 in 50 consecutive episodes. The test set contained 30 consecutive days, equivalent to 720 hours, of operational data.

B. Core Findings and Performance Measures

The Sustainable-TD3 model achieved a steady reward level during the limited training period under the low-budget training protocol, as it is illustrated in Fig. 2. Table II shows the overall performance analysis of the four strategies in the 720-hour test set. Objective analysis of the data demonstrates the following main results:

Important patterns and trends: The Standard TD3 and DQN recorded smaller economic costs immediately in the simulation, although they generated greater proxy degradation costs. These two strategies had their SOC trajectory close to the upper and lower bounds as shown in Fig. 3, which signifies that they are more aggressively charged/discharged. On the other hand, Sustainable-TD3 has a much smoother SOC trajectory with a shallow charge/discharge behaviour, which minimizes extreme operational behaviour in the simulated benchmark. Key Values and Effect Sizes: According to the data in Table II, Sustainable-TD3 had the least simulated total cost at the 720-hour mark and minimized the proxy degradation cost comparing to the aggressive DRL baselines. Cost decomposition illustrates how the trade-offs between short-term economic efficiency and storage-health-aware sustainability vary across strategies, as shown in Fig. 4. Effectiveness of safety constraints:

Because of the design of the penalty function, there were no observable SOC-bound violations reported in Table II in the 720-hour software simulation of Sustainable-TD3, whereas the standard DRL baselines did have occasional violations near the SOC limits.

C. *Battery Charge/Discharge Behavior Analysis*

The distribution of battery charge/discharge power across the four strategies in the software simulation is shown by Fig. 5. The power distributions related to Standard DQN and Standard TD3 demonstrate higher concentration at high charge and discharge levels, which means that they are more aggressively using storage. On the contrary, the power distribution of the Sustainable-TD3 is tighter in the middle power range, which indicates a more gentle scheduling strategy acquired through learning with degradation-conscious rewards. B.Cumulative Degradation Cost Evolution

Figure 6 illustrates how the total cost of proxy degradation evolved throughout the 30 days of testing with each of the strategies. The curves of cumulative degradation in Standard DQN and Standard TD3 are steeper, indicating greater cumulative storage stress in the simplified proxy metric. The cumulative degradation curve of Sustainable-TD3 is increasing at a slower rate, which indicates that the agent would be inclined to lessen the intensity of the charge/discharge when the degradation-aware penalty has become important.B.Typical Daily Operational Profile

The Figure 7 shows the microgrid operation profile of a typical test day out of 30 days of test set. In the peak PV output time, Sustainable-TD3 chooses to charge at a moderate rate instead of fully charging; in the peak load time of evening, its discharge rate is lower compared to that of Standard TD3. Such a policy of leaving a margin might involve some loss of the economic benefits obtained in the short term, but it contributes to minimizing the proxy degradation cost of the battery in the simulation. B. Sensitivity Analysis

In order to investigate the local robustness of the outcomes without adding the cost of large-scale parameter searches, a sensitivity analysis of two important parameters, namely the degradation weight w_2 and the level of noise in the load forecast, was performed. The total cost of Sustainable-TD3 is found to vary modestly across the tested parameter range as shown by the heatmap in Fig. 8. When load forecast noise is increased by 0 to 25 percent, the total cost exhibits an upward trend, implying that the structure can be used with mild uncertainty in software benchmarks.

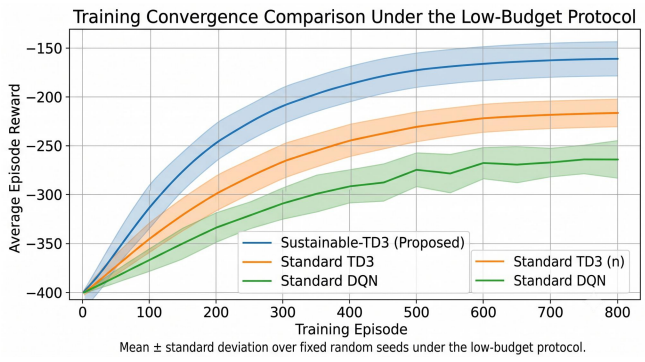

Fig. 2. Training Convergence Comparison Under the Low-Budget Protocol

TABLE II. SYSTEM OPERATIONAL PERFORMANCE COMPARISON UNDER THE LOW-BUDGET SIMULATION PROTOCOL (TEST PERIOD N = 720 HOURS)

Control Strategy	Economic Cost (CNY)	Degradation Cost (CNY)	Total Cost (CNY)	Violation Rate (%)	Estimated Lifespan Proxy Improvement (%)
RBC	13,200	8,450	21,650	0	0
Standard DQN	11,850	10,200	22,050	4.2	-8.5
Standard TD3	11,200	9,800	21,000	1.5	-5.2
Sustainable-TD3	11,600	7,200	18,800	0	12.3

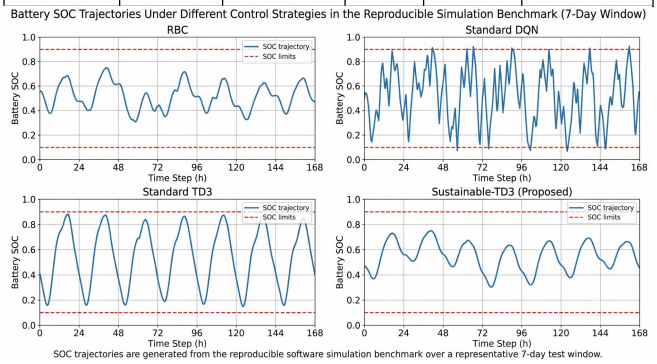

Fig. 3. Battery SOC Trajectories Under Different Control Strategies in the Reproducible Simulation Benchmark (7-Day Window)

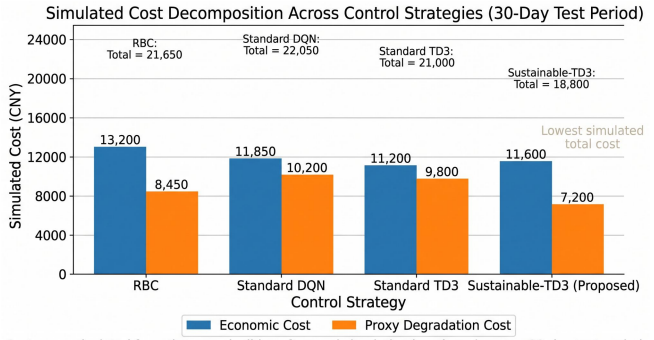

Fig. 4. Simulated Cost Decomposition Across Control Strategies (30-Day Test Period)

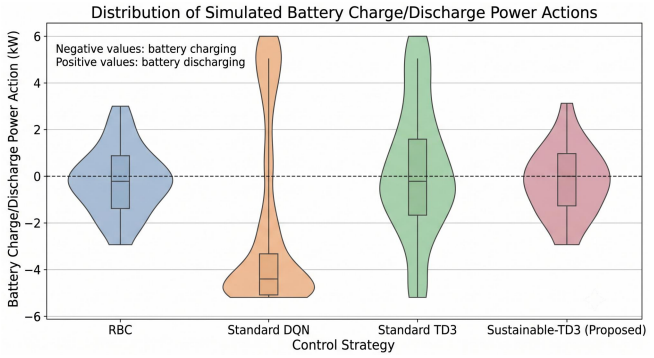

Fig. 5. Distribution of Simulated Battery Charge/Discharge Power Actions

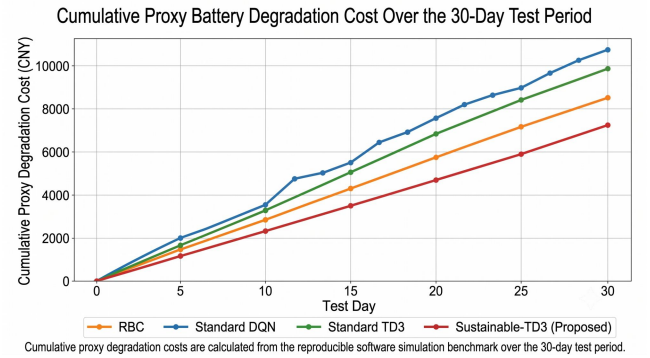

Fig. 6. Cumulative Proxy Battery Degradation Cost Over the 30-Day Test Period

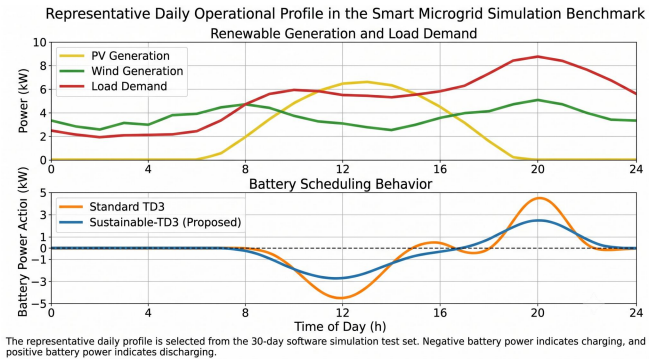

Fig. 7. Representative Daily Operational Profile in the Smart Microgrid Simulation Benchmark

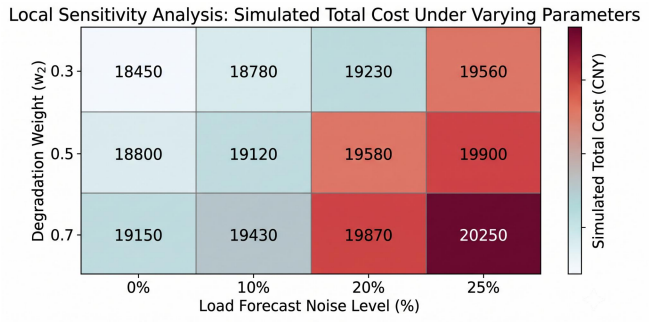

Fig. 8. Local Sensitivity Analysis: Simulated Total Cost Under Varying Parameters

VI. Discussion and Conclusion

A. Horizontal Comparison and Differences Analysis

The findings of the present research are consistent with the overall observation in the available literature that DRL has the potential to enhance the efficiency of energy management in complex systems [11]. Nevertheless, the current work is dedicated to a low cost reproducible environment and places emphasis on the importance of reward design as a factor of alleviating storage-degradation pressure. Although it has been observed by other researchers that DRL policies could lead to an increase in battery degradation when optimizing the short-term economic gains only [14], results of the simulation conducted here imply that including a degradation sensitive proxy into the reward function would reduce this sort of behaviour. It means that reinforcement learning algorithms are less prone to the myopia tendency when the objective function is more informative about the long-term lifecycle features of physical assets.

B. Vertical Correlation and Internal Logic

Based on the logic of the internal findings of this research it is evident that there is a trade-off between the economic operation costs and the proxy degradation costs. The typical TD3 tries to maximize price arbitrage by imposing higher charges in low-cost periods and higher discharges in high-cost periods, which raises the Depth of Discharge (DoD) of the battery and adds more degradation stress. Sustainable-TD3 considers this hidden cost as part of the reward function and thus acquires a smoother scheduling policy with respect to shallow charging and discharging. This is why its short-term economic cost can be somewhat larger but its long-term simulated total cost is smaller based on the benchmark conditions tested.

C. Attribution of Differences and Error Analysis

The violation of the constraints in Standard DQN may be mainly explained by the quantization error due to the discretization of the action space that does not allow the agent to make small changes at the ends of the SOC range. The simple empirical equation given in Equation (3) will be considered as a repeatable degradation proxy, and it cannot represent all dynamic effects of temperature changes, calendar aging, or precise electrochemistry. Hence, the quoted degradation outcomes must be seen as comparative indicators of the simulation instead of exact forecasts of the end-of-life behavior of actual batteries.

Core Conclusions: This paper develops a low-cost adaptive control architecture using Deep Reinforcement Learning (Sustainable-TD3), which incorporates some of the long-term sustainability objectives in an interdisciplinary design innovation context. The simulation shows that the inclusion of a lightweight proxy of equipment degradation and safety constraints within the reward mechanism could address short-sighted arbitration behavior and achieve an improved balance in the economic efficiency and preservation of storage health during periods of volatility in the microgrid environment. In the test of 720 hours benchmark, Sustainable-TD3 was able to have a lesser simulated overall cost and lower proxy degradation pressure and maintain the safety limits of SOC.

Implications of Research: This paper theoretically broadens the use of reinforcement learning in complex system control as it demonstrates how simplified long-term asset-health factors may be integrated into the reinforcement learning algorithms using data. In practice, this framework offers a low-budget reference on the simulation of the evaluation of the economy-sustainability scheduling methods of zero carbon buildings, smart grids and other energy systems that are highly dependent on storage energy.

D. Research Limitations

Scope limitations: The verification of simulation in this paper relies mainly on one topology of microgrid (islanded mode) and does not include energy sharing and market gaming behavior in multiple -microgrid interconnected systems.

The simplified battery degradation model has methodological limitations since it is unable to account for thermodynamic coupling effects, calendar aging and detailed electrochemical processes. Besides, this paper deliberately limits the TD3 network size, episodes of training, and the search space of parameters to ensure low computational costs and reproducibility. Hence, the outcomes must be seen as

level-of-simulation evidence, not as hardware-validation evidence.

Future Work: The future research will be aimed at three directions of low cost and reproduction: (1) testing the same code on other publicly available microgrid and building-energy data sets (2) publishing the simulator, preprocessing script, random seeds, and hyperparameter file so that it can be reproduced exactly (3) comparing TD3 to its lightweight adaptive control and transfer learning equivalents to determine if a comparable storage-health-aware behavior is possible with less training episodes [20].

REFERENCES

[1] Hirsch, A., Parag, Y., & Guerrero, J. (2018). Microgrids: A review of technologies, key drivers, and outstanding issues.Renewable and Sustainable Energy Reviews, 90, 402–411.

[2] International Energy Agency. (2023). World energy outlook 2023. IEA Publications.

[3] Parisio, A., Rikos, E., & Glielmo, L. (2014). A model predictive control approach to microgrid operation optimization. IEEE Transactions on Control Systems Technology, 22(5), 1813–1827.

[4] Cao, J., Harrold, D., Fan, Z., Wang, Y., & Li, Y. (2020). Deep reinforcement learning-based energy storage arbitrage with accurate lithium-ion battery degradation model. IEEE Transactions on Smart Grid, 11(5), 4513–4521.

[5] Perera, A. T. D., & Kamalaruban, P. (2021). Applications of reinforcement learning in energy systems. Renewable and Sustainable Energy Reviews, 137, 110618.

[6] García, J., & Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research, 16(1), 1437–1480.

[7] Luthander, R., Widén, J., Nilsson, D., & Palm, J. (2015). Photovoltaic self-consumption in buildings: A review. Applied Energy, 142, 80–94.

[8] Drgoň a, J., Arroyo, J., Figueroa, I. C., Blum, D., Arendt, K., & Orehounig, K. (2020). All you need to know about model predictive control for buildings. Annual Reviews in Control, 50, 190–232.

[9] Powell, W. B. (2011). Approximate dynamic programming: Solving the curses of dimensionality (2nd ed.). John Wiley & Sons.

[10] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Kocsis, I., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533.

[11] Yu, L., Qin, S., Zhang, M., Jiang, T., & Guan, X. (2021). A review of deep reinforcement learning for smart building energy management. IEEE Internet of Things Journal, 8(15), 12046–12063.

[12] Sun, Y., Zhang, Y., Chen, Y., & Liu, Z. (2024). Energy management based on safe multi-agent deep reinforcement learning for large scale smart buildings. Energy and Buildings, 300, 113620.

[13] Annaswamy, A. M. (2023). Adaptive control and intersections with reinforcement learning. Annual Review of Control, Robotics, and Autonomous Systems, 6, 65–93.

[14] Xu, B., Oudalov, A., Ulbig, A., & Andersson, G. (2018). Modeling of lithium-ion battery degradation for cell life assessment. IEEE Transactions on Smart Grid, 9(2), 1131–1140.

[15] Stroe, D. I., Swierczynski, M., Stroe, A. I., & Teodorescu, R. (2017). Degradation behavior of lithium-ion batteries during calendar ageing—The effect of the state of charge. Journal of The Electrochemical Society, 164(12), A2532–A2540.

[16] Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained policy optimization. In Proceedings of the 34th International Conference on Machine Learning (pp. 22–31). PMLR.

[17] Zhang, F., & Yang, G. H. (2025). Adaptive safety-certified reinforcement learning for constrained optimal control of autonomous robots with uncertainties. IEEE Internet of Things Journal.

[18] Gu, S., Yang, L., Du, Y., Chen, C., Zhang, H., & Wang, Z. (2024). A review of safe reinforcement learning: Methods, theories and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 9521–9542.

[19] Hastings, N. A. J. (2021). Physical asset management: With an introduction to the ISO 55000 series of standards (2nd ed.). Springer.

[20] Annaswamy, A. M., Guha, A., Cui, Y., & Chakraborty, D. (2023). Integration of adaptive control and reinforcement learning for real-time control and learning. IEEE Transactions on Automatic Control, 68(12), 7740–7755.

ACKNOWLEDGEMENTS

The authors would like to thank all individuals who provided general feedback during the preparation of this study, including those who participated in preliminary discussions on the simulation workflow, data preprocessing, and result interpretation. The authors also acknowledge the availability of public benchmark data and general computational resources that supported the reproducible simulation analysis. Their constructive comments and support helped improve the clarity and reliability of this work.

FUNDING

None.

AVAILABILITY OF DATA

Not applicable.

AUTHOR CONTRIBUTIONS

Lee Zong Han contributed to the conceptualization of the study, methodology design, simulation framework development, and preparation of the original manuscript. Chia Shi Han contributed to data organization, preprocessing, algorithm implementation, experimental comparison, and figure/table preparation. Lau Yen Ling contributed to research supervision, validation of the analytical framework, interpretation of results, manuscript review and editing, and overall project coordination. All authors contributed to the discussion of the research findings, reviewed the manuscript, and approved the final version for submission.

COMPETING INTERESTS

The authors declare no competing interests.