Benchmarking Deep RL for Off-Grid Hybrid Microgrids in Sub-Saharan Africa

Jordan Leis
University of Waterloo
jordan.leis@uwaterloo.ca

Devon Kisob
University of Waterloo
dkisob@uwaterloo.ca

Jennifer Yu
University of Waterloo
j545yu@uwaterloo.ca

Behzad Waseem
University of Waterloo
behzad.waseem@uwaterloo.ca

Julian Bauer-Kong
University of Waterloo
jbauerko@uwaterloo.ca

Yalda Nikookar
University of Waterloo
ynikooka@uwaterloo.ca

Ruitong Zhang
University of Waterloo
r544zhan@uwaterloo.ca

Cilo Zhou
University of Waterloo
m69zhou@uwaterloo.ca

Abstract—Reliable electricity access for approximately 570 million unelectrified people in sub-Saharan Africa (SSA) depends on off-grid solar-battery-diesel hybrid microgrids. The energy management system (EMS) governing battery dispatch and diesel throttle critically impacts reliability and fuel cost, yet reinforcement learning (RL) research in this domain is fragmented—most studies evaluate one or two algorithms at a single site. We present a 150-run benchmark of six deep RL algorithms—SAC, DDPG, TQC, PPO, A2C, and Recurrent PPO—across five climatically distinct SSA locations and five independent seeds. Agents are trained on five years of real NASA POWER irradiance data (2019–2023) and evaluated on a held-out 2024 year in a high-fidelity Gymnasium simulation. Off-policy algorithms achieve near-zero unmet energy; DDPG attains the best aggregate performance (7.5 kWh/yr unserved, 20,007 L/yr diesel—a 23% fuel reduction over SAC/TQC). On-policy methods exhibit systematic failure modes. Code: https://github.com/Jordan-Leis/Microgrid-RL.

I. Introduction

As of 2022, approximately 570 million people in SSA—roughly 43% of the global unelectrified population—lacked electricity access [1]. Over 70% of new connections required for universal access by 2030 must come from off-grid or mini-grid solutions, as per-household grid extension costs exceed USD $3,000 in dispersed rural communities [1], [2]. Solar-battery-diesel hybrid microgrids are the primary technology: PV generation covers daytime loads, battery storage bridges overnight demand, and diesel backup provides firm capacity [2].

The EMS governing battery dispatch and diesel throttle must resolve a stochastic, long-horizon trade-off at each 30-minute timestep: minimize fuel cost, unmet energy, and battery wear under uncertain solar irradiance and community load. Rule-based controllers apply fixed priority logic that degrades under seasonal shifts [3]; model predictive control (MPC) requires renewable forecasts that are difficult to maintain in data-scarce environments. Reinforcement learning (RL) offers a model-free alternative capable of learning adaptive policies from simulation without an explicit system model [4], [5].

Prior RL microgrid studies show competitive performance relative to rule-based baselines [3], [6], [7] but predominantly target grid-connected cost optimization. Islanded SSA systems require a reliability-first objective: any blackout is a direct welfare loss with no grid fallback. Most prior studies also evaluate one or two algorithms at a single site, precluding structural conclusions about algorithmic differences or cross-climate generalization [8], [9].

Contributions: (1) An open-source, high-fidelity Gymnasium-compatible simulation environment calibrated to SSA hardware and NASA POWER climate data. (2) The largest RL microgrid algorithm benchmark to date: six algorithms × five SSA locations × five seeds = 150 runs, with zero data leakage via a held-out 2024 evaluation year. (3) A characterization of structural failure modes—gradient collapse (PPO), battery over-cycling (A2C), fixed-throttle satisficing (SAC/TQC), bimodal convergence (RPPO)—invisible in single-seed evaluations, with direct implications for algorithm selection in field deployment.

II. Background

A. Microgrid Architecture

A hybrid microgrid integrates a PV array, battery bank, and diesel generator behind an inverter bus to serve islanded loads. In SSA, seasonal irradiance variability ranges from ∼10% in the Sahel to ∼100% in equatorial zones [2], requiring storage and diesel backup for reliable supply. Fig. 1 shows the system topology. The dispatch problem is a Markov Decision Process (MDP) [4]: at each step the agent observes system state and issues continuous battery and generator commands, receiving rewards that encode reliability and operating cost over a 17,520-step annual episode (365 days at 30-minute resolution).

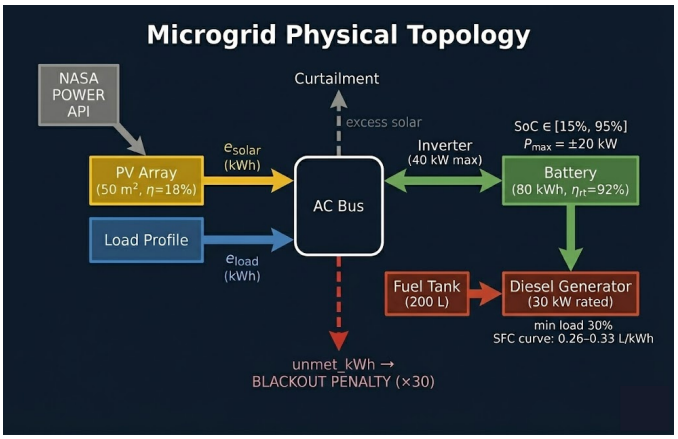

Fig. 1. Physical topology of the simulated solar-battery-diesel hybrid microgrid.

B. RL Algorithms

All six algorithms are implemented in Stable-Baselines3 [10] with $[256, 256]$ policy networks. The three off-policy methods maintain a replay buffer enabling experience reuse across gradient updates: SAC [11] adds entropy regularization for exploration; DDPG [12] uses a deterministic actor with target networks; TQC [13] adds distributional critic ensembles with quantile truncation to reduce overestimation bias. The three on-policy methods discard trajectories after each update: PPO [14] constrains gradient steps via a clipped surrogate objective; A2C [15] uses synchronous parallel workers with advantage estimation; RPPO [14] extends PPO with LSTM layers. The replay buffer is the decisive structural advantage for long-horizon (17,520-step) episodes where on-policy methods must credit decisions made hundreds of steps earlier.

III. Methodology

A. Simulation Environment

The environment is a Gymnasium-compatible [16] Python class parameterized via YAML. The default system models a small rural community: 50 m^2 PV array (18% efficiency), 80 kWh lithium-ion battery (SOC bounds: 15–95%), and 30 kW diesel generator.

PV model. Output at each 30-minute step:

$$P_{\mathrm{pv},t} = A\,\eta_{\mathrm{ref}}\,G_t\big[1-\gamma(T_t-25)\big](1-f_s), \qquad (1)$$

where $A = 50$ m^2, $\eta_{\mathrm{ref}} = 0.18$, G_t is surface irradiance (kW/m^2), $\gamma = 0.0035$ °C^{-1} is the temperature derating coefficient, and $f_s = 0.05$ is a soiling derate. Derating is critical at tropical and Sahelian sites where ambient temperatures routinely exceed 35 °C.

Battery model. SOC evolves with symmetric charge/discharge efficiency $\eta_c = \eta_d = \sqrt{0.92} \approx 0.959$ [17]. The rainflow cycle counting algorithm [18] is applied to the SOC history at episode end to compute equivalent full cycles (EFC) for degradation estimation.

Diesel model. A specific fuel consumption (SFC) curve—0.33, 0.30, 0.28, 0.26 L/kWh at load fractions 0.3, 0.5, 0.7, 1.0—penalizes part-load operation. A 0.5 L startup penalty and minimum on/off time constraints (2 h and 1 h) prevent excessive cycling.

B. Observation and Action Spaces

Fig. 2 summarizes the interface. The observation is a six-dimensional vector: battery SOC $\in [0, 1]$, fuel level $\in [0, 1]$, irradiance (kW/m^2), normalized temperature ($T/50$), hour of day $\in [0, 1]$, and load demand (kWh/step). The action is two-dimensional: $a_{\mathrm{batt}} \in [-1, 1]$ (negative=charge, positive=discharge up to 20 kW) and $a_{\mathrm{diesel}} \in [0, 1]$ (throttle fraction of rated 30 kW). Continuous actions match real inverter and generator controller interfaces.

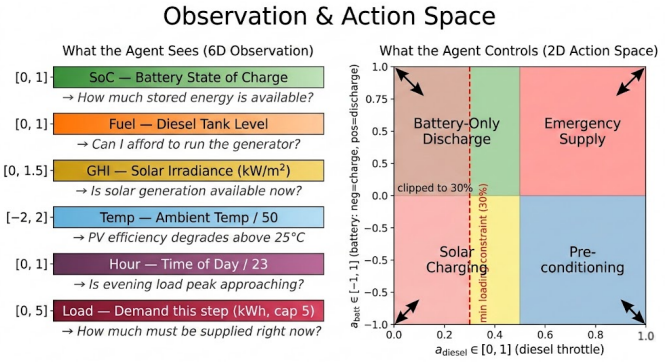

Fig. 2. Six-dimensional observation space and two-dimensional continuous action space.

C. Reward Formulation

$$R_t = -\big(p_u\,u_t + c_f\,f_t + \lambda_c\,\mathrm{EFC}_t + p_s\,s_t + p_g\,g_t\big), \qquad (2)$$

where u_t (kWh) is unmet energy, f_t (L) is diesel consumed at cost $c_f = 1.5$ \$/L, EFC_t is incremental battery cycling, $s_t \in \{0, 1\}$ is the generator start indicator, and g_t is generator operating hours. Coefficients: $p_u = 30$ \$/kWh, $\lambda_c = 0.005$, $p_s = 0.5$ \$, $p_g = 0.05$ \$/h. Setting $p_u = 30$ ensures the blackout penalty dominates the marginal cost of diesel dispatch, enforcing reliability as the primary objective. Fig. 3 visualizes this decomposition.

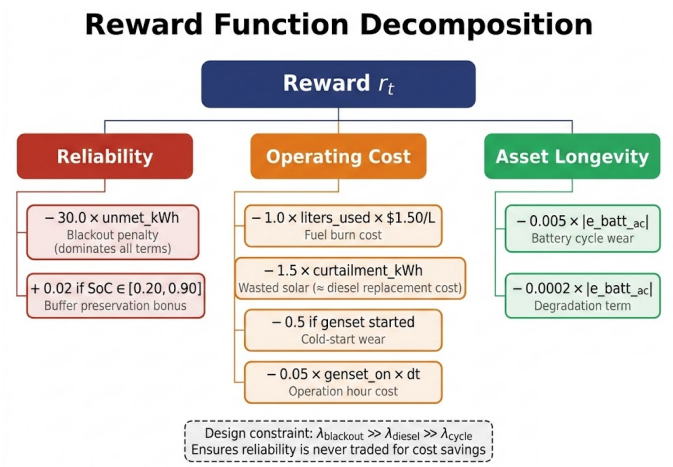

Fig. 3. Reward function decomposition showing reliability, operating cost, and asset longevity terms with their respective coefficients.

D. Study Locations and Data

Five SSA sites spanning three Köppen climate zones [19] were selected: Niamey, Niger (Sahel semi-arid); Dakar, Senegal (Sahel/Atlantic coast); Kumasi, Ghana (tropical rainforest); Libreville, Gabon (equatorial humid); Addis Ababa, Ethiopia (highland tropical). Niamey and Dakar offer high-irradiance, low-variability profiles; Kumasi and Libreville stress-test agents under persistent cloud cover; Addis Ababa presents a sharp monsoon transition at 2,355 m. Irradiance (ALLSKY_SFC_SW_DWN) and temperature (T2M) were obtained from NASA POWER [20] at hourly resolution and resampled to 30-minute intervals via linear interpolation. Years 2019–2023 form the training set; 2024 is fully held out for evaluation.

E. Training Configuration

Each run trains for 750,000 timesteps; 150 runs total. Off-policy hyperparameters: train_freq=256, gradient_steps=16, batch_size=4096 (SAC/TQC); gradient_steps=8 (DDPG). On-policy algorithms use SubprocVecEnv parallel rollouts. Evaluation uses 50 held-out 2024 episodes per run.

IV. Results

A. Aggregate Performance

Table I reports mean performance across all five locations and five seeds. A structural performance gap between off-policy and on-policy families is the dominant finding.

TABLE I
Mean benchmark performance across 150 runs (6 algos × 5 locations × 5 seeds), evaluated on held-out 2024 data. EFC = equivalent full battery cycles per year.

Algo	Type	Unmet (kWh)	Diesel (L)	EFC
DDPG	Off	7.5	20,007	∼2,300
SAC	Off	<1	26,018	∼2,300
TQC	Off	<1	26,018	∼3,000
A2C	On	∼5,800	∼19,000	∼4,000
RPPO	On	∼3,000	∼13,000	∼300
PPO	On	∼11,500	∼0	∼0

Fig. 4 shows all four primary metrics side by side with seed variability. Fig. 5 shows training reward trajectories.

B. Off-Policy Algorithms

DDPG achieves the best aggregate performance: 7.5 kWh/yr unmet energy and 20,007 L/yr diesel—a 23% fuel reduction relative to SAC and TQC at equivalent reliability. DDPG's deterministic actor learns an adaptive dispatch strategy, modulating diesel throttle in response to battery SOC and irradiance, making targeted fuel use only when stored energy is insufficient. SAC and TQC converge instead to a fixed-throttle satisficing strategy: 26,018 ± 1 L/yr, invariant across all five climate zones and all five seeds. This striking result—identical diesel consumption regardless of whether the site is the dry Sahel or the wet equatorial zone—indicates the entropy regularization term drives SAC (and by inheritance TQC) to a single robust but fuel-wasteful equilibrium, preventing the fuel-optimal adaptation learned by DDPG.

C. On-Policy Failure Modes

PPO collapses to a do-nothing attractor: ∼0 L/yr diesel and ∼0 EFC, with ∼11,500 kWh/yr unserved. With no resources dispatched, all load goes unmet. The long-horizon credit assignment challenge—17,520-step episodes where early decisions yield rewards only hours later—overwhelms the on-policy update, which discards all experience after each rollout.

A2C avoids complete inaction but exhibits battery over-cycling: ∼4,000 EFC/yr (compared to ∼2,300 for DDPG/SAC), while still leaving ∼5,800 kWh/yr unserved. A2C thrashes the battery without learning when to dispatch diesel, resulting in the most negative mean episodic return of any algorithm.

RPPO shows bimodal convergence: ∼50% of seeds train to a partially functional policy (visible in the large error bars in Fig. 4), while the remainder collapse. The LSTM hidden state aids temporal credit assignment in successful seeds, but initialization sensitivity causes frequent collapse. This bimodal distribution is invisible in single-seed evaluations.

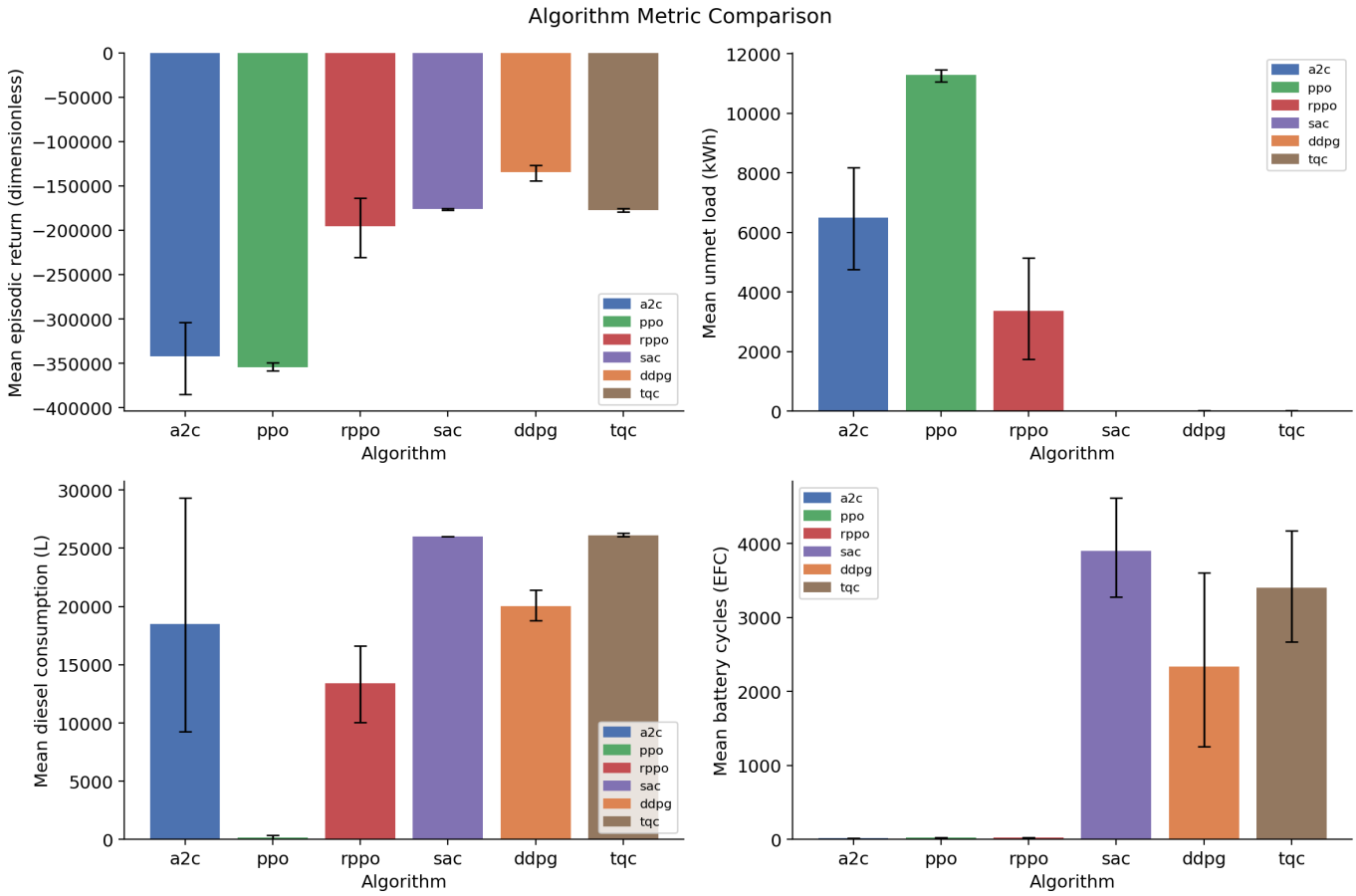

Fig. 4. Benchmark results across 150 runs: mean episodic return, unmet load, diesel consumption, and battery EFC per algorithm. Error bars indicate ±1 standard deviation across seeds and locations.

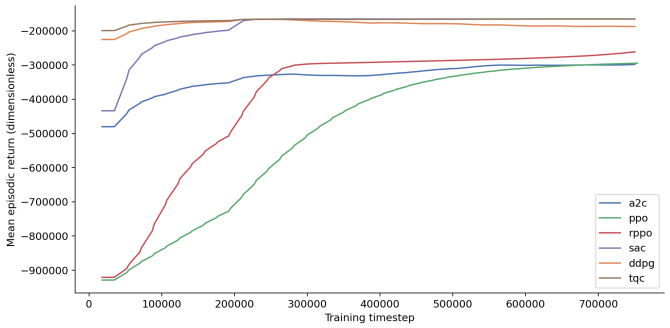

Fig. 5. Training reward curves (mean across 5 seeds, Niamey). TQC/SAC/DDPG converge from the very first episodes; PPO and A2C spend most of training near −900,000 before slowly recovering.

D. Cross-Climate Generalization

The SAC/TQC climate-invariant fuel result (26,018 ± 1 L/yr across five distinct Köppen zones) reveals a policy that ignores the site-specific irradiance signal entirely. DDPG, by contrast, shows meaningful cross-location variation in diesel use, confirming that its adaptive strategy responds to climate-driven differences in solar availability. These results suggest entropy regularization can impede policy specialization in continuous-action dispatch tasks despite improving exploration during training.

E. Statistical Reproducibility

Following [8], all results are reported across five seeds. SAC/TQC seed variance on diesel is essentially zero ($\sigma \approx 1$ L/yr), confirming convergence to a single attractor. DDPG shows low cross-seed variance on both reliability and fuel. On-policy methods show high variance driven by the collapse/no-collapse split, reinforcing that any single-seed evaluation of these algorithms on long-horizon tasks is statistically unreliable.

V. Conclusion

A consistent structural finding emerged across 150 runs: off-policy algorithms with replay buffers reliably solve the long-horizon dispatch problem, while on-policy methods exhibit systematic and diverse failure modes. DDPG achieves the best reliability-fuel trade-off (7.5 kWh/yr unserved, 20,007 L/yr diesel), with SAC/TQC achieving comparable reliability at a 23% fuel penalty due to fixed-throttle satisficing. On-policy failure modes—PPO gradient collapse, A2C battery over-cycling, RPPO bimodal

convergence—are structural, not incidental, and suggest that reward shaping alone cannot compensate for the credit assignment limitations of on-policy updates over 17,520-step annual episodes.

Future work should investigate hybrid architectures combining replay-based learning with recurrent policy networks, multi-site transfer learning to reduce per-site training cost, and integration with real hardware controllers for field validation. The environment, training pipeline, and all 150 run results are released openly to support reproducible RL research in energy access.

Acknowledgements

The authors thank the University of Waterloo for computational resources. Climate data were obtained from the NASA Langley Research Center POWER Project funded through the NASA Earth Science/Applied Science Program.

References

[1] International Energy Agency, "Africa energy outlook 2022," IEA, Tech. Rep., 2022. [Online]. Available: https://www.iea.org/reports/africa-energy-outlook-2022

[2] World Bank Group, "Mini grids for half a billion people: Market outlook and handbook for decision makers," World Bank, Tech. Rep., 2023. [Online]. Available: https://www.worldbank.org/en/topic/energy/publication/mini-grids-for-half-a-billion-people

[3] T. A. Nakabi and P. Toivanen, "Deep reinforcement learning for energy management in a microgrid with flexible demand," Sustainable Energy, Grids and Networks, vol. 25, p. 100413, 2021.

[4] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. MIT Press, 2018.

[5] U. Inayat, M. F. Zia, H. Mahmood, M. Tariq, and A. U. Rehman, "Learning-based methods for energy management in microgrid: A comprehensive review," Energies, vol. 16, no. 1, 2023.

[6] X. Qi, Y. Luo, G. Wu, K. Liao, and H. Tian, "Deep reinforcement learning-based optimal scheduling of iot-driven isolated microgrid," IEEE Internet of Things Journal, vol. 7, no. 10, pp. 10 172–10 183, 2020.

[7] Q. Sang, G. Wang, C. Wang, Y. Xu, and W. Shi, "Deep reinforcement learning-based energy management of hybrid energy storage systems in electric vehicles," Journal of Energy Storage, vol. 54, p. 105336, 2022.

[8] R. Agarwal, M. Schwarzer, P. S. Castro, A. Courville, and M. G. Bellemare, "Deep reinforcement learning at the edge of the statistical precipice," in Advances in Neural Information Processing Systems (NeurIPS), vol. 34, 2021, pp. 29 304–29 320.

[9] G. Dulac-Arnold, N. Levine, D. J. Mankowitz, J. Li, C. Paduraru, S. Gowal, and T. Hester, "Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis," Machine Learning, vol. 110, pp. 2419–2468, 2021.

[10] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," Journal of Machine Learning Research, vol. 22, no. 268, pp. 1–8, 2021.

[11] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in Proceedings of the 35th International Conference on Machine Learning (ICML), 2018, pp. 1861–1870.

[12] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in Proceedings of the 4th International Conference on Learning Representations (ICLR), 2016.

[13] A. Kuznetsov, P. Shvechikov, A. Grishin, and D. Vetrov, "Controlling overestimation bias with truncated mixture of continuous distributional quantile critics," in Proceedings of the 37th International Conference on Machine Learning (ICML), 2020, pp. 5556–5566.

[14] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.

[15] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Harley, T. P. Lillicrap, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in Proceedings of the 33rd International Conference on Machine Learning (ICML), 2016, pp. 1928–1937.

[16] M. Towers, J. K. Terry, A. Kwiatkowski, J. U. Balis, G. de Cola, T. Deleu, M. Goulão, A. Kallinteris, A. KG, M. Krimmel, R. Perez-Vicente, A. Pierré, S. Schulhoff, J. J. Tai, A. T. J. Shen, and O. G. Younis, "Gymnasium: A standard interface for reinforcement learning environments," arXiv preprint arXiv:2407.17032, 2024.

[17] B. Xu, A. Oudalov, A. Ulbig, G. Andersson, and D. S. Kirschen, "Modeling of lithium-ion battery degradation for cell life assessment," IEEE Transactions on Smart Grid, vol. 9, no. 2, pp. 1131–1140, 2018.

[18] S. D. Downing and D. F. Socie, "Simple rainflow counting algorithms," International Journal of Fatigue, vol. 4, no. 1, pp. 31–40, 1982.

[19] M. C. Peel, B. L. Finlayson, and T. A. McMahon, "Updated world map of the Köppen-Geiger climate classification," Hydrology and Earth System Sciences, vol. 11, no. 5, pp. 1633–1644, 2007.

[20] NASA Langley Research Center, "POWER: Prediction of worldwide energy resources," NASA Applied Sciences Program, Tech. Rep., 2023. [Online]. Available: https://power.larc.nasa.gov