Quicker to Train and Cheaper to Run: Edge-Ready DRL for Microgrid Resilient Energy Management

1st Mohammad Hossein Nejati Amiri
College of Engineering
Birmingham City Universit
Birmingham, UK
mohammadhossein.nejatiamiri
@mail.bcu.ac.uk

2nd Florimond Gueniat*
College of Engineering
Birmingham City University
Birmingham, UK
florimond.gueniat@bcu.ac.uk

***Abstract*—Resilient microgrid control only delivers value in remote settings if it can be trained quickly and run on resource-constrained hardware. This paper applies Simple Policy Optimisation (SPO) to energy management and benchmarks it against Proximal Policy Optimisation (PPO) and a Model Predictive Control (MPC) baseline under historical Cyclone Laila conditions. A short two-stage curriculum first secures resilience, then trades within the feasible set to extend battery life using either a balanced summation reward or a Lagrangian resilience constraint; goal-based early stopping breaks training once targets are met. Under the same actor–critic scheme, SPO produced steadier updates and stronger battery preservation than PPO, achieving up to 19.7 expected years of battery life (32% longer than MPC and 7% longer than PPO) while maintaining near-perfect resilience (RI = 0.998, compared with 0.993 for PPO Lagrangian). Compared with PPO, training compute is reduced by nearly twelve times with similar inference cost, whereas MPC is around a thousand times heavier at inference. SPO therefore offers an edge-deployable, resilience-first controller that protects batteries without sacrificing continuity of supply.**

***Index Terms*—Microgrid Resilient Operation, Deep Reinforcement Learning, Energy Management, Smart Grid, Proximal Policy Optimisation, Simple Policy Optimisation**

I. Introduction

Microgrids (MGs) are used to keep priority loads on during severe events while integrating variable Renewable Energy Sources (RES) [1]. An Energy Management System (EMS) must balance intermittent generation and limited storage with fast, reliable decisions [2]. Model Predictive Control (MPC) can plan these trade-offs when models and forecasts are accurate, but online solution time grows with horizon length and mixed-integer decisions [3]. Deep Reinforcement Learning (DRL) offers a model-free alternative with light, predictable inference for real-time control [1], [4].

Among DRL methods, actor–critic algorithms are widely used for continuous actions. Proximal Policy Optimisation (PPO) is popular for EMS because it stabilises updates via a clipped surrogate, yet clipping can mute useful gradients once the probability ratio crosses the bound [5]. Simple Policy Optimisation (SPO) replaces hard clipping with a smooth quadratic penalty and keeps all samples gradient-active, improving training stability and efficiency [6]. SPO has not been applied to power system studies. This paper applies SPO to resilient EMS with a two-stage curriculum: Stage 1 secures resilience for priority loads; Stage 2 improves battery life within the feasible region using either a summation reward that balances Resilience Index (RI) and Expected Years (EY) or a Lagrangian constraint that enforces a resilience floor while rewarding longevity, with goal-based early stopping in both cases. PPO and SPO share the same actor–critic scheme and codebase [6], and are compared fairly with an MPC baseline under identical data, metrics (RI and Expected Years, EY), and deterministic evaluation using historical cyclone conditions [3].

A. Literature review

1) DRL for operational EMS: DRL has been used to control charging, discharging, and load allocation under uncertainty and continuous actions [7]. Value-based methods such as DQN/DDQN appear in discrete settings [8], while actor–critic methods handle continuous decisions; examples include DDPG and SAC, with SAC known for stable training [4]. PPO is a standard choice for microgrid scheduling, balancing performance and stability with a clipped objective [1]. However, the clip bound can limit gradient information and slow learning when the ratio saturates [5]. SPO modifies this update by a quadratic penalty around the ratio bound, retaining gradient signal from all samples and giving smoother progress [6]. SPO is a recently introduced method [6], and practical EMS studies applying it for operation have not yet been reported.

Gap (DRL for operation): A direct assessment of SPO for microgrid operation such as its training efficiency versus PPO, stability of returns, and quality of the learned trade-offs, remains missing.

2) Resilience and battery life in operation: Resilience is central to microgrid value: islanded operation, service restoration, and coordinated dispatch during high-impact, low-probability events all depend on it [8], [9]. Networked settings motivate coordinated and hierarchical control, including multi-agent DRL [10], while robustness to uncertainty has been explored with Bayesian DRL to temper value overestimation [11]. In planning, DRL has combined Loss of Load Probability (LOLP)-based resilience with explicit ageing models to guide

capacity and technology choice [12]. At the operational level, studies have begun to include battery terms, yet life preservation is often treated alongside resilience rather than tightly linked in a single learning signal [13], [14].

A practical difficulty is that resilience (continuity of supply for priority loads) and battery life (sensitivity to Depth-of-Discharge (DoD) and cycling) can pull in opposite directions. Simple weighted-sum rewards may raise one objective at the expense of the other and do not guarantee a resilience base during training. By contrast, a constrained formulation can hold resilience above a target while allowing the agent to trade remaining capacity for longevity.

Gap (Resilience and battery life in operation): The literature offers few operational DRL formulations that jointly optimise resilience and battery life with an explicit mechanism to keep resilience above a threshold; most treatments consider the two aims in parallel (e.g., LOLP with degradation in planning [12]) or without a constraint that protects resilience during learning [1], [13]. This paper addresses that gap by comparing a weighted-sum design with a Lagrangian constraint on an operational RI.

3) Contributions and paper organisation: Based on the gaps identified in the literature, the contributions of this paper are:

- New DRL method for microgrid operation: Applies the recently introduced SPO to resilient microgrid operation, contrasted with PPO.
- Two stage curriculum training: Stage 1 secures resilience; Stage 2 extends battery life using either a summation reward (RI+EY) or a Lagrangian RI constraint, with goal-based early stopping.
- Benchmark: A fair comparison is conducted within a shared codebase and actor–critic scheme, using identical data and metrics (RI, EY, imbalance) under deterministic tests with historical cyclone conditions; MPC is included as a baseline [3].
- Compute efficiency: About $10\times$ lower training cost than PPO while keeping inference computation identical; deployment is far lighter than MPC.

The paper is organised as follows: Section II defines the environment and decisions; Section III reports results (RI, EY, imbalance, compute) against PPO and MPC for the historical Laila cyclone; and the paper concludes in Section IV.

II. Microgrid Environment and Decision Model

The test case builds on our previously published study of a rural microgrid in cyclone-prone Andhra Pradesh, India [3]. Three priority tiers are defined: *essential* (homes, school, clinic), *business* (shops, public lighting), and *agricultural/deferrable* (irrigation and other shiftable loads). HOMER Pro specifies the PV, wind, and BESS sizes as in the earlier work, where operation was managed by an MILP-based MPC. Here, operation is managed by an actor–critic DRL controller trained with SPO [6], with PPO as the baseline [5]. Training and evaluation incorporate wind and availability profiles from Cyclone Laila (2010) to capture resilience in renewable and storage management. Table I summarises the symbols used in this section.

TABLE I: Variables and Parameters Used in the Environment Model.

Symbol	Description	Value (Unit)
$\omega_t^{(k)}$	Raw allocation preference for load k	–
$L_{k,t}\geq 0$	Demand of load tiers $k\in\{1,2,3\}$ (essential, business, agricultural)	– (kW)
$R_t\geq 0$	Renewable generation (PV+WT)	– (kW)
P_t^{net}	Pre-battery net balance $R_t-\sum_{k=1}^{3}L_{k,t}$	– (kW)
$S_{k,t}$	Shortage of load k	– (kW)
SOC_t	State of charge in $[\text{SOC}_{\min},\text{SOC}_{\max}]$	[0.20, 0.90] (p.u.)
$E_{\max}$	Battery energy capacity	780 (kWh)
$P_{b,\max}$	Battery power limit (charge or discharge)	52 (kW)
η_{ch}, η_{dis}	Charge and discharge efficiencies	0.90, 0.95
H	History length	5 (steps)
T	Episode length	216 (steps; 9 days)
$f_t^{(k)}$	Softmax allocation fraction for load k	$\in(0,1)$, $\sum_k f_t^{(k)}=1$
q_k	Priority weight	q_1=7, q_2=2, q_3=1

The following subsections outline the environment in terms of state representation, action space, physical constraints, and reward formulation.

A. State (Observation)

At each step, normalised features summarise storage, supply, and demand:

$$s_t=\big[\,\text{SOC}_t,\ \widehat{L}_{1,t},\ \widehat{L}_{2,t},\ \widehat{L}_{3,t},\ \widehat{R}_t,\ \widehat{P_t^{\text{net}}}\,\big]. \tag{1}$$

Hatted quantities are min–max scaled by scenario-wide maxima:

$$\widehat{X_t}\ =\ \frac{X_t}{X^{\max}},\qquad X_t\in\{L_{1,t},L_{2,t},L_{3,t},R_t,P_t^{\text{net}}\}, \tag{2}$$

with $\widehat{X_t}=0$ if $X^{\max}=0$. This representation provides a consistent, dimensionless state that exposes battery SOC, the three load tiers, renewable availability, and the instantaneous supply–demand balance.

Temporal context is provided by a stacked history,

$$S_t=\big[\,s_t,\ s_{t-1},\ \ldots,\ s_{t-H+1}\,\big]\in\mathbb{R}^{6H}, \tag{3}$$

which allows the policy to observe short-term patterns such as evening peaks or windy nights over the last H steps.

B. Action and Scaling

The controller proposes charge/discharge commands and a soft preference over loads:

$$a_t=\big[\,u_t^{\text{ch}},\ u_t^{\text{dis}},\ \omega_t^{(1)},\ \omega_t^{(2)},\ \omega_t^{(3)}\,\big]\in[-1,1]^5. \tag{4}$$

Power commands are mapped to physical units:

$$P_t^{\text{ch}}=\frac{u_t^{\text{ch}}+1}{2}\,P_{b,\max},\qquad P_t^{\text{dis}}=\frac{u_t^{\text{dis}}+1}{2}\,P_{b,\max}. \tag{5}$$

The first two action components decide battery power; the last three shape how any available supply is divided across loads. Mutual exclusivity follows the pre-battery balance:

$$\begin{cases}P_t^{\text{dis}}=0, & \text{if } P_t^{\text{net}}\geq 0,\\ P_t^{\text{ch}}=0, & \text{if } P_t^{\text{net}}<0.\end{cases} \tag{6}$$

When renewables cover demand, only charging is permitted; when there is a deficit, only discharging is permitted.

C. Physical Constraints

The physical limits of the battery constrain the feasible charging and discharging power:

$$\begin{aligned} 0 \le P_t^{\text{ch}} &\le \min\Big\{P_{b,\max},\ \big(\text{SOC}_{\max} - \text{SOC}_t\big)E_{\max}\Big\}, \\ 0 \le P_t^{\text{dis}} &\le \min\Big\{P_{b,\max},\ \big(\text{SOC}_t - \text{SOC}_{\min}\big)E_{\max}\Big\}. \end{aligned} \tag{7}$$

The state of charge evolves as

$$\begin{aligned} \text{SOC}_{t+1} &= \text{SOC}_t + \frac{\eta_{\text{ch}} P_t^{\text{ch}} - \frac{P_t^{\text{dis}}}{\eta_{\text{dis}}}}{E_{\max}}, \\ \text{SOC}_{t+1} &\leftarrow \text{clip}\big(\text{SOC}_{t+1},\ \text{SOC}_{\min},\ \text{SOC}_{\max}\big). \end{aligned} \tag{8}$$

Charging increases SOC with efficiency η_{ch}, discharging decreases it accounting for η_{dis}; limits prevent overcharge or deep depletion.

After the battery decision, the supply available to loads is

$$P_t^{\text{s}} = R_t + P_t^{\text{dis}} - P_t^{\text{ch}}. \tag{9}$$

This is the *post-battery* power budget to distribute across essential, business, and agricultural demand. Preferences are converted into fractions by a softmax:

$$f_t^{(k)} = \frac{\exp(\omega_t^{(k)})}{\sum_{j=1}^{3} \exp(\omega_t^{(j)})}, \quad k \in \{1,2,3\}. \tag{10}$$

Larger $\omega_t^{(k)}$ yields a larger share for load k, while ensuring $f_t^{(1)} + f_t^{(2)} + f_t^{(3)} = 1$. Delivered power and shortage are then

$$P_{k,t}^{\text{s}} = \begin{cases} f_t^{(k)}\, P_t^{\text{s}}, & P_t^{\text{s}} \ge 0, \\ 0, & P_t^{\text{s}} < 0, \end{cases} \quad S_{k,t} = \max\{0,\ L_{k,t} - P_{k,t}^{\text{s}}\}. \tag{11}$$

When supply is positive, it is split by $f_t^{(k)}$; when supply is negative, all load is unmet. The shortage $S_{k,t}$ measures unserved demand in each group.

D. Rewards Shaping

To reflect the relative importance of different loads, fixed priority weights are assigned as q_k (Table I). The weighted demand and shortage defined as

$$D_t = \sum_{k=1}^{3} q_k L_{k,t}, \qquad N_t = \sum_{k=1}^{3} q_k S_{k,t}. \tag{12}$$

A per-step resilience reward are

$$R_t^{\text{step}} = \begin{cases} 1, & D_t = 0, \\ 1 - N_t/D_t, & D_t > 0, \end{cases} \qquad r_t = \frac{\beta}{T} R_t^{\text{step}}, \tag{13}$$

R_t^{step} equals one when all priority-weighted demand is met and declines linearly with the weighted fraction unserved; the small scale factor ($\frac{\beta}{T}$) keeps per-step rewards comparable across stages.

Resilience over the episode is

$$\text{RI}_{\text{ep}} = \begin{cases} 1, & \sum_t D_t = 0, \\ 1 - \dfrac{\sum_t N_t}{\sum_t D_t}, & \sum_t D_t > 0. \end{cases} \tag{14}$$

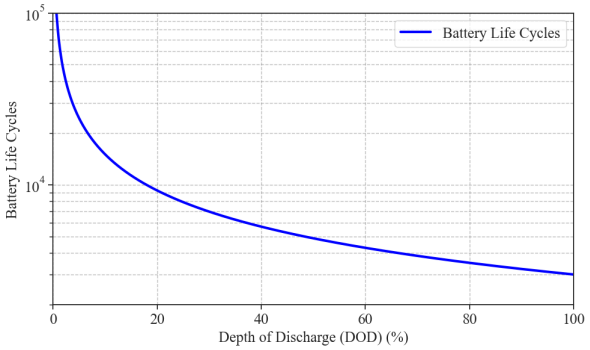


Fig. 1: Battery life cycles as a function of depth of discharge [3].

The index reports the priority-weighted fraction of demand served across the horizon. Battery stress is mapped to expected life through a Battery Life Cycle (BLC)–DoD relation, illustrated in Figure 1. At each step, the instantaneous depth of discharge is computed as

$$\text{DOD}_t = \frac{P_t^{\text{dis}}}{E_{\max}} \quad (\text{if } P_t^{\text{dis}} > 0), \tag{15}$$

A harmonic aggregator converts variable-depth cycles to expected years:

$$C_{\text{harm}} = \Big(\sum_{t\in\mathcal{D}} \tfrac{1}{\text{BLC}(\text{DOD}_t)}\Big)^{-1}, \qquad \text{EY} = \frac{n_{days}}{365}\, C_{\text{harm}}, \tag{16}$$

with normalised score

$$\text{EY}_{\text{rw}} = \min\{\,\text{EY}/15,\ 1\,\}. \tag{17}$$

Deep cycles contribute fewer equivalent full cycles; the harmonic mean captures this, and the $n_{days}/365$ factor converts cycles to years at a nominal daily usage rate.

Two complementary constructions are used to form the final episodic reward from resilience and battery longevity.

Let the per-step resilience be $R_t^{\text{step}} \in [0,1]$ and define the *normalised* average

$$\bar{R} \;=\; \frac{1}{T}\sum_{t=0}^{T-1} R_t^{\text{step}} \;\in\; [0,1], \tag{18}$$

which removes any stage-dependent scale factors used during training. Let the episode metrics be $\text{RI}_{\text{ep}} \in [0,1]$ and $\text{EY}_{\text{rw}} \in [0,1]$ (the lifetime score after normalisation). Set $B = \frac{1}{2}\big(\text{RI}_{\text{ep}} + \text{EY}_{\text{rw}}\big) \in [0,1]$ as a balanced terminal summary.

(i) Summation (weighted-sum) mode. The final return blends the trajectory quality ($\bar{R}$) with the terminal summary (B):

$$R_{\text{sum}} \;=\; \tfrac{1}{2}\,\bar{R} \;+\; \tfrac{1}{2}\,B. \tag{19}$$

The return lies in $[0,1]$ and increases when the policy both maintains high per-step service (especially for essential loads) and achieves strong terminal resilience and battery life.

(ii) Lagrangian (constraint) mode. Resilience is enforced as a minimum threshold $\mathrm{RI}_{\mathrm{ep}} \geq \rho$ (target ρ=0.99). Define the violation

$$v \;=\; \max\{0,\, \rho - \mathrm{RI}_{\mathrm{ep}}\}, \tag{20}$$

and update the dual (penalty) by projected ascent

$$\lambda \;\leftarrow\; \Pi_{[0,\lambda_{\max}]}\big(\lambda + \alpha\, v\big), \tag{21}$$

with step size $\alpha > 0$ and cap $\lambda_{\max} > 0$. The terminal Lagrangian term rewards longevity and penalises any shortfall,

$$L \;=\; \max\{\, 0,\, \mathrm{EY}_{\mathrm{rw}} - \lambda\, v \,\}, \tag{22}$$

and the normalised final return is

$$R_{\mathrm{lag}} \;=\; \bar{R} \;+\; \big(\mathrm{EY}_{\mathrm{rw}} - \lambda v\big), \quad v = \max\{0, \rho - \mathrm{RI}_{\mathrm{ep}}\}. \tag{23}$$

This return is unnormalised and may exceed 1 when the resilience constraint is met (v=0). If the constraint is met, the terminal contribution reduces to L=$\mathrm{EY}_{\mathrm{rw}}$ and learning focuses on extending battery life. If the threshold is violated, the penalty λv reduces L and the dual increases, making future violations costlier; the projection ensures stability and boundedness.

E. Policy Optimisation: PPO vs. SPO and Training Procedure

Both controllers are on-policy actor–critic methods with an identical training procedure (vectorised rollouts, GAE, mini-batch updates). They differ only in the *policy loss*: PPO uses a clipped surrogate; SPO replaces clipping with a smooth quadratic penalty that keeps all samples contributing gradients.

Let $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)}$, $\hat{A}_t$ be the GAE advantage, and ε the ratio bound.

The PPO loss, based on the clipped surrogate objective, is defined as [5]:

$$\mathcal{L}_\pi^{\mathrm{PPO}} \;=\; -\,\mathbb{E}\Big[\min\Big(r_t(\theta)\hat{A}_t,\; \mathrm{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\,\hat{A}_t\Big)\Big]. \tag{24}$$

In contrast, the SPO loss can be expressed as [6]:

$$\mathcal{L}_\pi^{\mathrm{SPO}} \;=\; -\,\mathbb{E}\left[r_t(\theta)\hat{A}_t \;-\; \frac{|\hat{A}_t|}{2\varepsilon}\,(r_t(\theta)-1)^2\right]. \tag{25}$$

The quadratic penalty acts as a soft regulariser: for $\hat{A}_t > 0$ it pulls r_t upward towards $1+\varepsilon$, and for $\hat{A}_t < 0$ it pulls r_t downward towards $1-\varepsilon$. Unlike PPO, which sets the gradient to zero once r_t crosses the clipping range, SPO preserves informative gradients from all samples, ensuring smoother and more stable updates.

F. Curriculum Learning and Goal-Based Early Stopping

The controller is trained with a short, two-stage curriculum that first stabilises resilience and then trades within the feasible set to improve battery lifetime. The stages differ only in their reward shaping and stopping targets. The overall flow is summarised in Figure 2, while hyperparameter settings are reported in Table II.

TABLE II: Training hyperparameters used in the custom SPO/PPO trainer.

Category	Parameter	Value
General	Vectorised envs N_{env}	12
	Episode length T_{ep}	216 steps (9 days)
	Rollouts per chunk	1000 (per stage chunk)
Policy & Value	Actor (depth)	7 layers: 256,256,128,128,64,64, out
	Actor nonlinearity	tanh
	Critic (depth)	2×128, tanh
Optimisation	Algorithm	PPO or SPO (selectable)
	Optimiser	Adam
	Learning rate η	3×10^{-4} (adaptive throttle)
	Discount γ	0.995
	GAE λ	0.95
	PPO clip ϵ (PPO only)	0.20
Reward / Curriculum	Stage 1 scaling β	0.5
	Stage 2 scaling β	1.0
	Stage 2 reward mode	sum / lagrangian
	Lagrange step α	0.05 (default)
	Lagrange cap $\lambda_{\max}$	10.0 (default)
Evaluation/Stop	Eval policy	deterministic (mean action)
	Stop target (Stage 1)	RI $\geq$ 0.998
	Stop target (Stage 2)	RI $\geq$ 0.9956, EY $\geq$ 15.9 yr

a) Stage 1 (stabilise resilience): The per-step shaping places a stronger emphasis on served energy for priority loads in the first stage. The scale factor is set to β=0.5 so that step rewards remain normalsied during exploration. Training proceeds in chunks of 1000 rollouts. After each chunk, the deterministic policy is evaluated on a held-out dataset to compute the episode RI. Stage 1 stops as soon as the goal $\mathrm{RI} \geq 0.998$ is met on the evaluation set or the iteration limit is reached. This ensures that Stage 2 begins from a policy that already protects the resilience of essential and business loads.

b) Stage 2 (optimise longevity under a resilience baseline): Stage 2 uses the same rollouts/updates but switches the terminal objective to one of two modes: (i) *Summation*, which averages episode resilience and the normalised lifetime score, or (ii) *Lagrangian*, which enforces $\mathrm{RI}_{\mathrm{ep}} \geq \rho$ via a dual variable λ (step size α and cap $\lambda_{\max}$) while rewarding battery life. The per-step scale is set to β=1.0. After each training chunk, the policy is evaluated and Stage 2 stops once both deployment targets are satisfied: $\mathrm{RI} \geq 0.9956$ and $\mathrm{EY} \geq 15.9$ years, or when the maximum number of iterations is reached. This *goal-based early stopping* avoids over-training and shortens wall-clock time.

Separating "make it resilient" (Stage 1) from "make it last" (Stage 2) mirrors operational priorities in cyclone-prone microgrids: resilience is non-negotiable, then longevity is optimised inside that feasible region. The Lagrangian mode makes this logic explicit by increasing λ whenever RI dips below ρ, automatically refocusing the agent on continuity of supply; once the constraint is met, the penalty vanishes and the policy is free to maximise lifetime. All optimiser, network, and LR/KL controls are exactly those listed in Table II.

III. Results and Discussion

This section presents the deterministic evaluation of the proposed SPO controller against PPO and MPC baselines. Performance is compared on RI, EY, load imbalance, and computational effort.

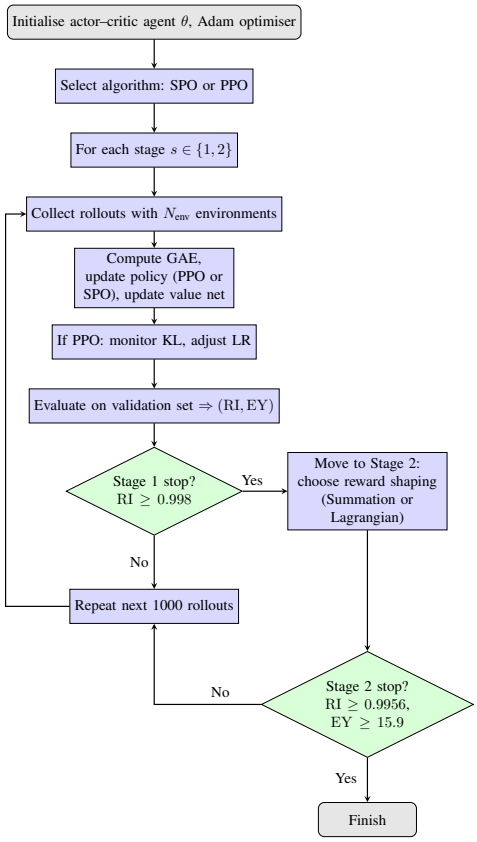


Fig. 2: Flowchart of the two-stage SPO/PPO curriculum learning.

A. Deterministic Performance

Table III summarises the results under certain conditions. The MPC baseline reached near-perfect resilience (RI $=$ 0.9991) with a 14.44-year battery life. However, its SOC profile (Fig. 3) shows sharp discharges, reflecting deeper cycling and explaining its lower lifetime. A similar pattern is observed in PPO Stage 1 (summation), which also reached very high resilience (RI $=$ 0.9998) but at the cost of shorter lifetime (14.95 years).

PPO in Lagrangian mode extended battery life (18.34 years) but with reduced resilience (RI $=$ 0.9928) and greater imbalance. By contrast, SPO consistently preserved higher SOC across the horizon, reducing DOD stress. In Lagrangian mode, SPO achieved near-perfect resilience (RI $=$ 0.9981) with the longest lifetime (19.65 years). In summation mode, SPO also maintained resilience at unity, though the SOC trajectory shows deeper discharges comparable to MPC, indicating that the agent prioritised resilience by allowing more stress on the battery. Nevertheless, SPO summation resulted in no imbalance, unlike PPO summation.

MPC further illustrates its forward-looking nature: although it maintained lower SOC levels during operation, it strategically avoided high charging cycles and consistently returned to SOC levels comparable with DRL controllers at the end of each horizon (e.g. day 20–21 dip followed by recovery by day 23–24). This highlights MPC's forecast advantage, which DRL must approximate through training.

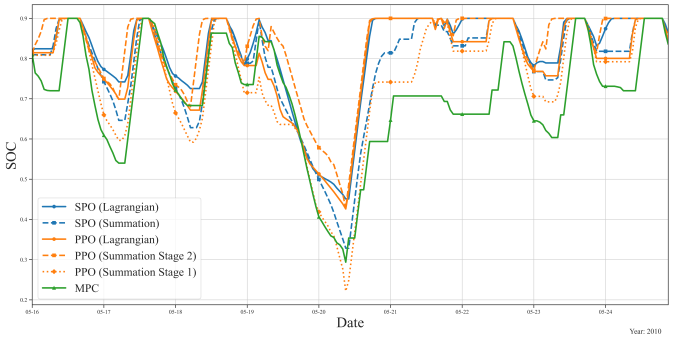


Fig. 3: Battery SOC trajectories for MPC, PPO, and SPO under deterministic operation. Initial SOC is set near 0.8 for fair comparison.

Another observation is that the summation reward is not guaranteed to preserve resilience in PPO: Stage 1 achieved a strong RI, but Stage 2 degraded sharply. The imbalance values confirm this instability: while all methods fully supported essential loads (Load 1), PPO summation Stage 2 failed, with 171.6 kWh unserved essential demand. This reinforces that the Lagrangian approach provides a more reliable trade-off between resilience and lifetime.

B. Computational Cost

Table IV compares computation across a 15-year horizon. PPO and SPO have essentially the same inference cost, while MPC is about a thousand times heavier per step and per episode. For training, SPO reduces compute by about 93% compared with PPO (almost thirteen times lighter). When considering lifetime cost over the full horizon, SPO remains around twelve times cheaper than PPO and more than twenty-five times cheaper than MPC, all while keeping inference requirements unchanged. These results underline SPO's training efficiency and deployment practicality (real-time feasible on edge hardware), in contrast to MPC's prohibitive runtime. The reduced training cost also benefits from the goal-based early stopping built into our two-stage curriculum, which terminates training once target resilience and lifetime thresholds are achieved.

Figures 4a and 4b show the episodic reward curves during curriculum training. SPO converges to high returns within a fraction of the roll-outs required by PPO, which continues to oscillate in Stage 2. This early convergence explains the order-of-magnitude lower training FLOPs reported for SPO in Table IV, while inference costs remain essentially unchanged. Together, the table and figures confirm SPO's efficiency and stability advantage

IV. Conclusion

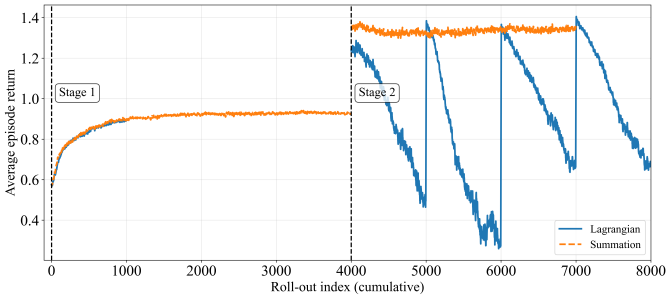
SPO, paired with a resilience-first curriculum and goal-based early stopping, delivered the best resilience–lifetime trade-off at a fraction of PPO's training cost and far lower

TABLE III: Comparison of MPC, PPO, and SPO under deterministic operation. Imbalance values are disaggregated by load tier.

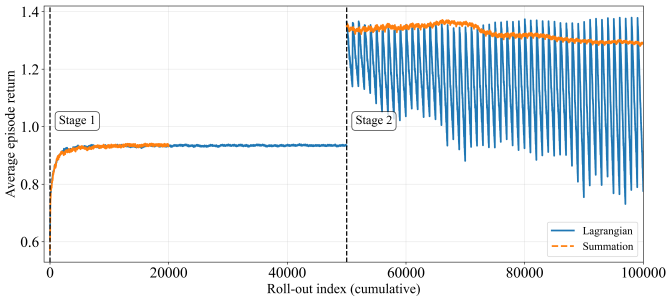
Method	Reward Mode	RI	EY (yr)	Load 1 (kWh)	Load 2 (kWh)	Load 3 (kWh)	Total Imb. (kWh)
MPC	–	0.9991	14.44	0.00	0.00	14.00	14.00
PPO	Lagrangian	0.9928	18.34	0.00	60.24	1.10	61.33
SPO	Lagrangian	0.9981	19.65	0.00	8.50	15.17	23.67
PPO	Summation (Stage 1)	0.9998	14.95	0.00	0.00	2.12	2.12
PPO	Summation (Stage 2)	0.9252	17.33	171.60	3.97	57.65	233.22
SPO	Summation	1.0000	16.06	0.00	0.00	0.00	0.00

TABLE IV: Computation-time comparison of MPC, PPO, and SPO controllers over a 15-year horizon (all values are in FLOPs).

Method	Training	Inference per Step	Inference per Episode	Lifetime (15yr)
MPC	–	$(7.594 \pm 21.840) \times 10^{11}$	$(1.650 \pm 0.048) \times 10^{14}$	1.004×10^{17}
SPO	$(3.594 \pm 1.268) \times 10^{15}$	$(5.362 \pm 0.525) \times 10^{8}$	$(1.161 \pm 0.113) \times 10^{11}$	3.665×10^{15}
PPO	$(4.556 \pm 0.912) \times 10^{16}$	$(5.425 \pm 0.309) \times 10^{8}$	$(1.175 \pm 0.067) \times 10^{11}$	4.563×10^{16}



(a) SPO training reward.



(b) PPO training reward.

Fig. 4: Curriculum training reward trajectories for SPO and PPO across two stages.

runtime than MPC. In Lagrangian mode it achieved near-perfect RI with the longest expected life, while the summation mode retained unity resilience without imbalance; PPO showed instability in summation. Inference cost for SPO and PPO is similar and lightweight, enabling edge deployment, whereas MPC remains orders of magnitude heavier. Overall, SPO is a practical alternative for cyclone-prone microgrids: quicker to train, cheaper to run, and kinder on batteries. Future work will extend to multi-microgrid coordination and richer ageing models.

References

[1] M. H. N. Amiri, F. Annaz, M. De Oliveira, and F. Gueniat, "Deep reinforcement learning with local interpretability for transparent microgrid resilience energy management," *arXiv preprint arXiv:2508.08132*, 2025.

[2] L.-L. Li, B.-X. Ji, Z.-T. Li, M. K. Lim, K. Sethanan, and M.-L. Tseng, "Microgrid energy management system with degradation cost and carbon trading mechanism: A multi-objective artificial hummingbird algorithm," *Applied Energy*, vol. 378, p. 124853, 2025.

[3] M. H. N. Amiri, S. Dhundhara, F. Annaz, M. De Oliveira, and F. Gueniat, "Two-stage microgrid resilience and battery life-aware planning and operation for cyclone prone areas in india," *Sustainable Cities and Society*, vol. 124, p. 106290, 2025.

[4] R. Kumar and M. De, "Advancement in power system resilience through deep reinforcement learning: A comprehensive review," *Renewable and Sustainable Energy Reviews*, vol. 222, p. 115951, 2025.

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

[6] Z. Xie, Q. Zhang, and R. Xu, "Simple policy optimization," *arXiv preprint arXiv:2401.16025*, 2024.

[7] N. F. P. Dinata, M. A. M. Ramli, M. I. Jambak, M. A. B. Sidik, and M. M. Alqahtani, "Designing an optimal microgrid control system using deep reinforcement learning: A systematic review," *Engineering Science and Technology, an International Journal*, vol. 51, p. 101651, 2024.

[8] H. Momen and S. Jadid, "Resilience enhancement of power distribution system using fixed and mobile emergency generators based on deep reinforcement learning," *Engineering Applications of Artificial Intelligence*, vol. 137, p. 109118, 2024.

[9] D. Kumar and A. Kumar, "A reliable hybrid autoregressive integrated moving average and deep reinforcement machine learning strategy for resiliency enhancement in microgrid," *Sustainable Energy, Grids and Networks*, vol. 39, p. 101424, 2024.

[10] D. Qiu, Y. Wang, T. Zhang, M. Sun, and G. Strbac, "Hierarchical multi-agent reinforcement learning for repair crews dispatch control towards multi-energy microgrid resilience," *Applied Energy*, vol. 336, p. 120826, 2023.

[11] T. Zhang, M. Sun, D. Qiu, X. Zhang, G. Strbac, and C. Kang, "A bayesian deep reinforcement learning-based resilient control for multi-energy micro-gird," *IEEE Transactions on Power Systems*, vol. 38, no. 6, pp. 5057–5072, 2023.

[12] K. Pang, J. Zhou, S. Tsianikas, and Y. Ma, "Deep reinforcement learning for resilient microgrid expansion planning with multiple energy resource," *Quality and Reliability Engineering International*, vol. 40, no. 1, pp. 34–56, 2024.

[13] L. Tightiz and H. Yang, "Resilience microgrid as power system integrity protection scheme element with reinforcement learning based management," *IEEE Access*, vol. 9, pp. 83963–83975, 2021.

[14] J. Cao, D. Harrold, Z. Fan, T. Morstyn, D. Healey, and K. Li, "Deep reinforcement learning-based energy storage arbitrage with accurate lithium-ion battery degradation model," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4513–4521, 2020.