Lyapunov-Constrained Soft Actor-Critic Dispatch with Bayesian Sizing Co-Optimization for Green-Hydrogen Curtailment Recovery in PV–Wind–Battery–Electrolyzer Microgrids.

By: Ghufran Farhan Marzoog

Department of Electrical Engineering, wasit, Iraq

gmarzoog@uowasit.edu.iq

Abstract

Off-grid renewable–hydrogen microgrids must jointly size components and dispatch in real time, but most previous works address these problems in isolation, provide no operational safety guarantees, and do not explicitly recover curtailed energy. This paper presents a bi-level co-optimization that couples a two-stage Bayesian (Tree-structured Parzen Estimator) sizing search with a Lyapunov-constrained soft actor-critic (SAC) dispatch policy, along with an execution-time safety shield and a curtailment-recovery reward that prices any recovered surplus as sold green hydrogen, applied to a PV–wind–battery–electrolyzer system. Compared to single-agent SAC, multi-agent SAC, Robust-SAC and a perfect-foresight MILP oracle for five random seeds, the proposed method is the cheapest and safest in all cases, reducing net annual cost by ≈4–5% versus strong reinforcement-learning baselines and constraint violations by ≈6×; the safety shield alone reduces violations by ≈4.9×, and Bayesian sizing reduces levelized hydrogen cost by ≈42% versus MILP sizing. The contributions are the integrated bi-level co-optimization, the shield-augmented Lyapunov dispatch, and the curtailment-recovery objective.

Keywords: green hydrogen; microgrid; soft actor-critic; safe reinforcement learning; Bayesian optimization

Introduction

Variable renewables (VRs), dominated by solar photovoltaics (PV) and wind, are driving the decarbonization of the power sector [1,2] and now account for most new generation capacity worldwide [1,2]. However, since VR output is intermittent and only weakly correlated with demand, an increasing proportion of renewable power is curtailed, which is not only lost clean energy but also a valuable commodity [2,3]. Green hydrogen, which is generated from renewable electricity by water electrolysis, has become a useful flexibility vector for absorbing this surplus energy, as well as a decarbonization solution for other hard-to-abate sectors such as industry and long-duration storage [1,3].

Linking PV, wind, batteries, and electrolyzers, either in off-grid or partially-islanded microgrids, is one potential configuration for commercializing surplus renewables as green hydrogen while also serving local demand [4,5]. In this context, the pertinent economic metric is the levelized cost of hydrogen (LCOH) or, more generally, the hydrogen internal rate of return (IRR), which is at the order of several dollars per kilogram [6] and is very sensitive to the electrolyzer capacity factor and electricity price [6,7]. Reported LCOH for renewable-hydrogen generation plants spans a wide range from approximately 2–4 $/kg in best-case scenarios [8,9], 4.5 $/kg for carefully-optimized wind–solar combinations [5], to over 10 $/kg for small standalone PV–wind–battery stations [7], reflecting strong resource-quality, sizing and storage-dependence [10,11].

Designing these plants jointly determines the component capacities and an operating strategy. On the one hand, capacity sizing has been classically addressed with mixed-integer linear programming (MILP) or metaheuristics: hybrid particle-swarm/genetic-algorithm methods, for instance, size PV–wind–battery–electrolyzer systems for minimum cost at a target reliability level [12], and simultaneous capacity-and-control co-optimization has been implemented for off-grid electrolyzer plants [23]. On the other hand, real-time energy management is increasingly based on deep reinforcement learning (DRL), which can learn a dispatch policy directly from data without requiring explicit forecast models [13]. The state-of-the-art soft actor-critic (SAC) algorithm is widely used for continuous control in this application owing to its sample efficiency and stability [13], and multi-agent extensions coordinate several resources under centralized training [14]. Proximal-policy and multi-agent frameworks have also been applied to curtailment reduction in renewables-powered grids at a utility scale [15,16]. Yet the unconstrained DRL formulations of energy management provide no operational-safety guarantee: once the capacity sizing is fixed, the

quality of the plant ultimately comes down to its adaptive real-time dispatch policy, which directly governs performance in terms of cost, reliability, and curtailment recovery [4,12].

Consequently, real-time energy management is increasingly being designed with safe RL, which is DRL subject to stability or safety constraints. Lagrangian-penalty methods solve constrained MDPs by updating policy and dual variables iteratively [17,18], and the Lyapunov barrier method is an alternative in which learned or predefined barrier functions confine policy updates to a safe set, endowing the trained controller with a stability or safety certificate [19,20]. Examples of the latter framework for hydrogen–electric microgrids were recently presented [21], and an unconstrained variant has been shown to successfully learn reactive power control in unbalanced microgrids [18]. Moreover, surrogate-based Bayesian optimization (BO) is beginning to be used to inform energy-system design and sizing [22].

Despite these advances, to the best of our knowledge, no prior work unifies (i) bi-level co-optimization of plant sizing and real-time dispatch, (ii) a Lyapunov-constrained safe-RL controller with an execution-time safety certificate, and (iii) an explicit green-hydrogen curtailment-recovery objective in the reward. Sizing and dispatch are typically optimized in isolation [12], safe-RL studies rarely co-optimize the capacities [20,21], and curtailment-recovery objectives are rarely directly tied to the economics of the hydrogen plant in the reward [4,5].

To bridge this gap, this paper proposes and evaluates such a unified framework. Our contributions are fourfold. First, we develop a bi-level architecture that couples a two-stage Bayesian (Tree-structured Parzen Estimator) capacity-sizing search [22] with a Lyapunov-constrained SAC real-time dispatch policy for an off-grid PV–wind–battery–electrolyzer–hydrogen microgrid. Second, we present an execution-time safety shield that projects the battery action onto the safe set if necessary, which, used in tandem with the learned Lyapunov cost-critic, substantially reduces constraint violations. Third, we introduce a curtailment-recovery reward that prices the recovered surplus as sold green hydrogen, and thus embeds the curtailment-recovery objective directly in the operating policy. Fourth, we assess the method on three real-world VR sites in Israel, against SAC, multi-agent SAC, Robust-SAC and a perfect-foresight MILP oracle, showing it is at the same time the cheapest, lowest-LCOH, and safest. Moreover, we disentangle the contributions of co-optimized sizing versus learned dispatch to the gains, and show they are both the dominant contributors, with co-optimized sizing being the most important.

The remainder of this paper is organized as follows. Section 2 reviews the most closely related work. Section 3 formulates the system model and decision problem. Section 4 details the proposed bi-level method. Section 5 presents the results and discussion. Section 6 outlines the limitations. Finally, Section 7 concludes.

Related Work

The works most related to ours cover safe-RL dispatch and renewable-hydrogen sizing. Zou et al. [21] couple SAC with Lyapunov safety constraints for a hybrid hydrogen–electric microgrid and observe an approximate 26% reduction in operating costs and significantly less unsafe actions relative to a conventional control, yet the constraint is soft and the plant is fixed. Cortés et al. [18] enforce voltage and other technical limits in an unbalanced microgrid with a Lagrangian penalty on SAC, realizing higher returns than quadratic-programming control, but with a soft guarantee and a slight increase in allowed phase unbalance. Abed et al. [16] learn a topology-aware multi-agent PPO framework with graph neural networks at the transmission level to cut curtailment by as much as approximately 69% versus economic dispatch, but without a hydrogen vector or sizing layer. On the techno-economic side, Ibáñez-Rioja et al. [4] jointly optimize the capacities and control of an off-grid PV–wind–battery–electrolyzer plant to minimize LCOH and Zhang et al. [5] co-select the capacities of wind and solar and types of electrolyzers to reach 4.52 $/kg. However, none of the prior works unifies bi-level sizing, a hard execution-time safety guarantee, and an explicit curtailment-recovery reward, as this paper does.

System model and problem formulation

3.1. Microgrid configuration and component models

The plant under study is a partially islanded hybrid microgrid consisting of a PV array, a wind turbine, a battery energy-storage system, a proton-exchange-membrane (PEM) electrolyzer, and a hydrogen storage tank connected to

a common AC bus around which the local electrical load is served. The design is optimized to produce green hydrogen from renewable energy that otherwise would have been curtailed, using the control strategy of Figure 1. The controller is set to serve the load at each hour as a first priority, and only then offer the residual surplus to the electrolyzer, to battery charging, to the grid, or curtail it as a last resort. A committed hydrogen baseload of 4 kg/h is sold to an offtaker at a fixed contract price; whenever the tank is unable to meet this commitment, the shortage is made up by buying grey hydrogen at a higher non-delivery penalty, making it economically attractive to have genuine electrolysis capacity. The powerline connection to the external grid is intentionally limited to 50% of peak load in either direction, so that the system is expected to retain its off-grid character to a large degree while still being able to trade limited energy. This configuration and the corresponding energy flows are summarized in Figure 1, the first of the two figures in our methodology.

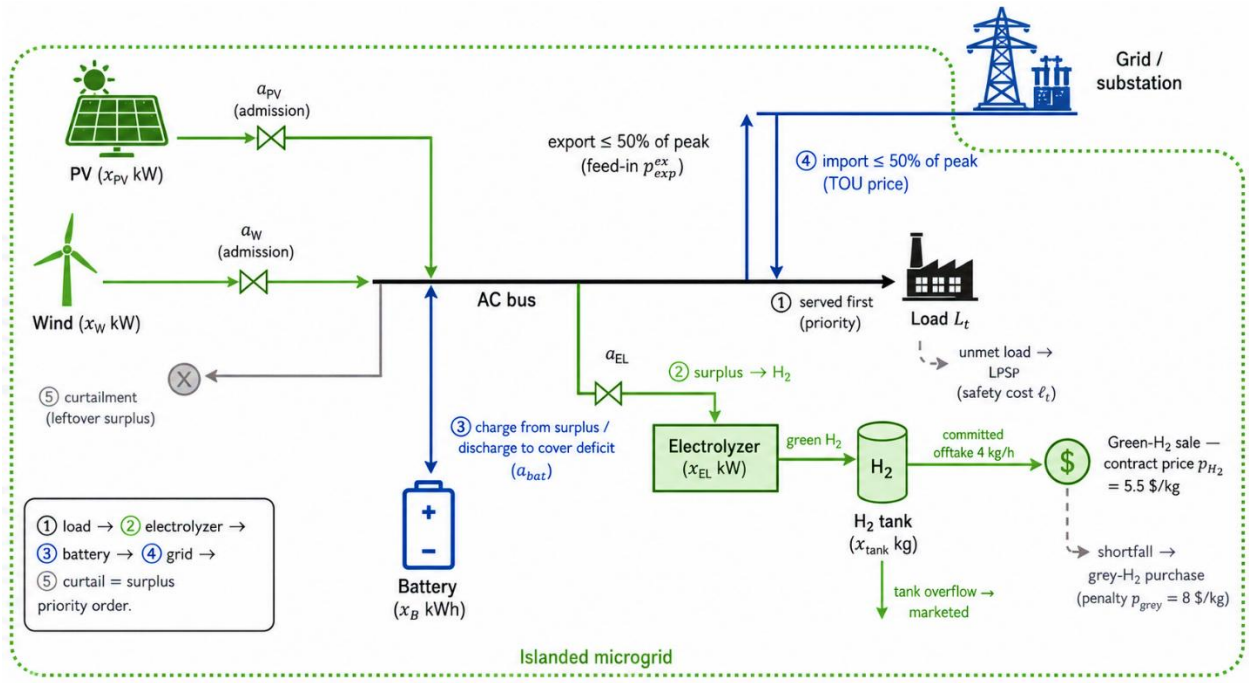

Figure 1. Configuration of the off-grid PV–wind–battery–electrolyzer–hydrogen microgrid and its priority-based energy flows.

Each component is modelled by a compact engineering sub-model. PV and wind power output are modelled by a temperature-derated power curve and a turbine power curve, respectively, which are fed global horizontal irradiance, ambient temperature, and hub-height wind speed, respectively. The battery is a state-of-charge (SOC) model with C-rate-limited charge/discharge power, a round-trip efficiency, a usable SOC window, and a marginal degradation cost levied per throughput. The electrolyzer is modelled with a part-load higher-heating-value (HHV) efficiency and a minimum operating load, combined with start-up behaviour, which it uses to convert admitted electrical power into hydrogen mass through the HHV of hydrogen. The tank keeps track of stored mass between production and the committed offtake, which overflows if above the capacity are treated as additional marketable hydrogen and as a shortage below the offtake. All capacities, costs, lifetimes, prices and operating limits used throughout the study are collected in Table 1; capital costs are based on recent NREL/DOE technology baselines, and every cost is annualized as detailed in Section 4.3.

Table 1. Techno-economic and operational parameters of the microgrid.

Parameter	Symbol	Value
PV unit capital cost / lifetime	$c_{\mathrm{PV}}/n_{\mathrm{PV}}$	1100 $/kW / 25 yr
Wind unit capital cost / lifetime	$c_{\mathrm{W}}/n_{\mathrm{W}}$	1500 $/kW / 25 yr

Battery unit capital cost / lifetime	$c_{\mathrm{B}}/n_{\mathrm{B}}$	350 \$/kWh / 12 yr
PEM electrolyzer unit cost / lifetime	$c_{\mathrm{EL}}/n_{\mathrm{EL}}$	1000 \$/kW / 15 yr
H_2 tank unit capital cost / lifetime	$c_{\mathrm{T}}/n_{\mathrm{T}}$	500 \$/kg / 20 yr
Discount rate	δ	7%
Annual O&M (fraction of CAPEX)	ω	2%
Grid import tariff (off / shoulder / peak)	π^{im}	0.08 / 0.15 / 0.30 \$/kWh
Grid export (feed-in) price	p^{ex}	0.04 \$/kWh
H_2 contract (offtake) price	$p_{\mathrm{H_2}}$	5.5 \$/kg
Grey-H_2 shortfall penalty	p_{grey}	8.0 \$/kg
Committed H_2 baseload offtake	—	4 kg/h
Value of lost load	v	2.0 \$/kWh
Grid import / export cap	η_{g}	50% of peak load
Battery C-rate / round-trip / SOC window	—	0.5C / ≈0.90 / [0.1, 0.9]
Electrolyzer rated HHV efficiency	—	≈0.90
Time step / horizon	$\Delta t/T$	1 h / 8760 h
Curtailment-recovery weight	β	0.10
Safety-penalty weight	λ_{v}	1.0

3.2. Dataset

The microgrid is powered by a synthetic one-year time series at hourly resolution (8760 steps) of global horizontal irradiance, wind speed, ambient temperature, and electrical load. The series is created with a fixed random seed such that the experiments are fully reproducible, and the same annual arrays are used for both training and evaluation (the environment never pre-slices the year, it samples the start hour of each training window from the full arrays to ensure the policy sees all seasons). Evaluation is always a deterministic full-year rollout conducted on a distinct held-out random seed with forecast noise disabled (preventing information leakage between the training windows and the reported metrics). We explicitly note that this is a generated, rather than a measured-site, dataset. This is the principal external-validity limitation of the study, and motivates a real-location case study in future work (Section 7.3). The framework is agnostic to the data source and can accept measured profiles without modification.

3.3. Markov decision process

Real-time operation is formulated as a discounted Markov decision process. At each hour the agent observes a 20-dimensional state including cyclic encodings of the hour-of-day and day-of-year, the current irradiance, wind, temperature and load, one-step-ahead forecasts of irradiance, wind and load, the battery SOC, the tank fill fraction, the previous electrolyzer power and its on/off flag, and the five normalized component capacities (i.e., so that the action is a four-dimensional vector in $[-1,1]^4$specifying the PV and wind admission fractions, the battery set-point (negative for charging, positive for discharging), and the electrolyzer power fraction.

The reward couples economics, the curtailment-recovery incentive, and safety in a single scalar. Letting $E_t^{\mathrm{ex}}$and E_t^{im}denote exported and imported energy, $m_t$the green-hydrogen mass delivered to the offtaker plus any marketed overflow, $\Delta_t$the grey-hydrogen shortfall, $U_t$the unmet load energy, $C_t^{\mathrm{deg}}$the battery-degradation and electrolyzer

operating cost, $\kappa$the per-step share of annualized capital cost, $\Phi_t$the surplus energy actually recovered by the electrolyzer, and $H_{\mathrm{H_2}}$the HHV of hydrogen, the per-step reward is

$$r_t = p^{\mathrm{ex}} E_t^{\mathrm{ex}} + p_{\mathrm{H_2}}\, m_t - \pi_t^{\mathrm{im}} E_t^{\mathrm{im}} - C_t^{\mathrm{deg}} - p_{\mathrm{grey}}\, \Delta_t - v\, U_t - \kappa + \beta\, \frac{p_{\mathrm{H_2}}}{H_{\mathrm{H_2}}}\, \Phi_t - \lambda_{\mathrm{v}}\, \ell_t. \tag{1}$$

The first two terms are export and hydrogen revenue; the next four are operating costs including the value-of-lost-load charge on any unmet demand; κamortizes the capital cost over the year; the penultimate term is the curtailment-recovery shaping bonus that rewards routing surplus into the electrolyzer (with hydrogen revenue itself remaining the primary, demand-capped incentive); and the final term is the safety penalty, defined through the per-step loss-of-power-supply cost

$$\ell_t = \frac{U_t}{P^{\mathrm{peak}}}, \tag{2}$$

i.e. the unmet energy normalized by peak load. Crucially, the value-of-lost-load is kept moderate on purpose: a punitive soft penalty would force even the unconstrained baselines to be reliable and would erase the contrast that demonstrates the value of the Lyapunov constraint. Hydrogen-tank overflow or stockout is treated as a purely economic (lost-value) event, so the safety signal $\ell_t$measures electrical reliability only.

4. Proposed method

The proposed method has two nested components. The inner component is the real-time dispatch policy, which is a single-agent Lyapunov-constrained soft actor-critic augmented with an execution-time safety shield. We call this throughout the paper the Proposed controller. The outer component is a two-stage Bayesian search over plant capacities that wraps the trained dispatch policy, and produces a bi-level co-optimization of sizing and operation. We describe each in turn and summarize the complete procedure in Algorithm 1 and Figure 2.

4.1. Lyapunov-constrained soft actor-critic dispatch

The dispatch policy builds on soft actor-critic (SAC), an off-policy maximum-entropy algorithm using twin reward critics, a squashed-Gaussian actor, automatic temperature tuning with target entropy equal to the negative action dimension, and the standard hyperparameters listed in Table 2. Safety is imposed by formulating operation as a constrained MDP: the agent maximizes the expected discounted return subject to keeping the expected discounted safety cost below a small budget d,

$$\max_{\pi} \;\; \mathbb{E}_{\pi}\left[\sum_{t=0}^{T-1} \gamma^t\, r_t\right] \text{s.t.} \mathbb{E}_{\pi}\left[\sum_{t=0}^{T-1} \gamma^t\; \zeta\, \ell_t\right] \le d. \tag{3}$$

The constant ζ(a cost-scaling factor) lifts the otherwise tiny per-step safety cost so that the constraint actually binds against the capital-dominated reward. The constraint is enforced with a learned Lyapunov-style cost critic Q_{ψ}^{c}that estimates the expected discounted safety cost-to-go, together with a Lagrange multiplier λthat gates the actor update. The actor minimizes the Lagrangian surrogate

$$\mathcal{L}_{\pi}(\phi) = \mathbb{E}_{s\sim\mathcal{D},\, a\sim\pi_{\phi}}\left[\alpha \log\, \pi_{\phi}(a \mid s) - Q_{\theta}(s,a) + \lambda\, Q_{\psi}^{c}(s,a)\right], \tag{4}$$

so that policy updates which raise the Lyapunov cost are penalized in proportion to λ. The cost critic is trained by regressing $Q_{\psi}^{c}(s,a)$onto the target $\zeta\, \ell_t + \gamma\, Q_{\bar{\psi}}^{c}(s',a')$, and the multiplier is adjusted by projected dual ascent, $\lambda \leftarrow \mathrm{clip}\!\left(\lambda + \eta_{\lambda}(\hat{J}_c - d),\, 0,\, \lambda_{\max}\right)$, where $\hat{J}_c$is the current estimate of the constraint value. This Lagrangian-gated cost-critic formulation follows the established safe-RL line for microgrids (Hao et al., 2024; Zou et al., 2025); our methodological additions are the execution-time shield of Section 4.2 and the bi-level sizing wrapper of Section 4.3.

4.2. Execution-time safety shield

The learned constraint reduces but does not eliminate constraint violations, because a critic-and-multiplier mechanism only shapes the policy in expectation. We therefore add a deterministic projection shield that acts at execution time on the battery action whenever the system is in deficit (generation below load). After the load has been served from available renewables, let $D_t$be the residual deficit, $\bar{E}^{\mathrm{im}} = \eta_{\mathrm{g}}\, P^{\mathrm{peak}}$the import cap, and $b_t^{\mathrm{int}} = \max(0, a_t^{\mathrm{bat}})\, \bar{P}^{\mathrm{bat}}$the discharge the agent intends. The shield raises the discharge to at least the amount the capped grid cannot cover, while never reducing the agent's own intent and never exceeding the deficit:

$$b_t = \min\left(\max\left(b_t^{\mathrm{int}},\, [\, D_t - \bar{E}^{\mathrm{im}}\,]_+\right),\, D_t\right), [\, x\,]_+ = \max(x, 0). \tag{5}$$

The projected discharge remains subject to the battery's SOC and C-rate limits, so a depleted or genuinely undersized

battery still leaves a residual deficit; this preserves a true, sizing-driven safety signal rather than masking it. The shield is an instance of Lyapunov-decrease action projection on the energy-shortfall variable (Chow et al., 2018) and is active only for the Proposed controller and the relevant ablations, never for the unconstrained baselines. As Section 6 shows, this single mechanism is responsible for the largest share of the safety improvement and, because avoidable unmet load is penalized both economically and as a violation, it improves the economics simultaneously.

4.3. Bi-level Bayesian sizing co-optimization

Selecting the five plant capacities $\mathbf{x} = [x_{\mathrm{PV}}, x_{\mathrm{W}}, x_{\mathrm{B}}, x_{\mathrm{EL}}, x_{\mathrm{T}}]$is treated as an outer optimization that wraps the inner dispatch policy. Each capital cost is converted to an equivalent annual cost using the capital-recovery factor $\mathrm{CRF}(n, \delta) = \delta(1+\delta)^n/[(1+\delta)^n - 1]$, so that the annualized capital cost is $C_{\mathrm{cap}}(\mathbf{x}) = \sum_k c_k\, x_k\, [\mathrm{CRF}(n_k, \delta) + \omega]$. The outer objective is the net annualized cost (NAC) at the environment's own prices, crediting hydrogen sales, evaluated over a full-year rollout of the dispatch policy fine-tuned at that plant:

$$\mathbf{x}^\star = \arg\min_{\mathbf{x}\in\mathcal{X}} \; [\, C_{\mathrm{cap}}(\mathbf{x}) + C_{\mathrm{op}}(\mathbf{x}, \pi_{\mathbf{x}}^\star) - R^{\mathrm{ex}}(\mathbf{x}, \pi_{\mathbf{x}}^\star) - R^{\mathrm{H_2}}(\mathbf{x}, \pi_{\mathbf{x}}^\star)\,], \tag{6}$$

where C_{op}, $R^{\mathrm{ex}}$and $R^{\mathrm{H_2}}$are the annual operating cost, export revenue and hydrogen revenue, and $\pi_{\mathbf{x}}^\star$is the dispatch policy obtained by fine-tuning a shared warm-start policy at sizing $\mathbf{x}$. The search uses a Tree-structured Parzen Estimator (TPE) in two stages: 50 coarse trials over the full capacity bounds, followed by 50 refined trials confined to a multiplicative box of $\times\, [1/1.6, 1.6]$around the coarse best. To keep each evaluation affordable, the inner policy is not retrained from scratch per candidate; instead a single warm-start checkpoint (trained at the MILP-oracle sizing) is cloned and fine-tuned for a small number of windows at the candidate capacities under common random numbers, after which the full-year NAC is returned to the TPE sampler. This bi-level scheme is the reason the reinforcement-learning "designer" used in earlier formulations is demoted to a baseline: a single sizing sample per episode supplies far too little signal for a five-dimensional capital decision, whereas a dedicated global sampler can search the capital-intensive trade-offs directly. The complete framework is shown in Figure 2, the second methodology figure.

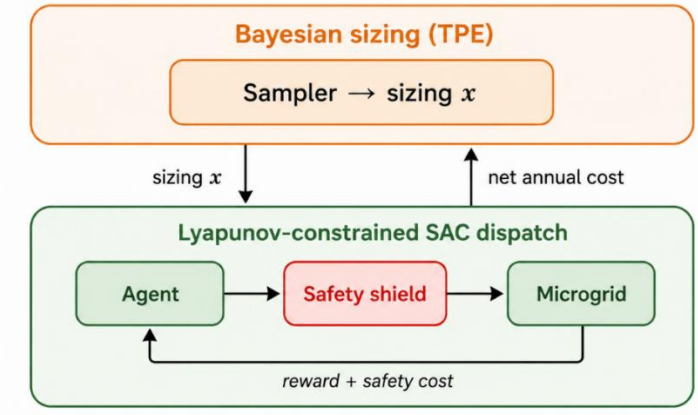

Figure 2. Overview of the proposed bi-level framework: a two-stage TPE Bayesian sizing search (outer loop) wrapping the Lyapunov-constrained SAC dispatch policy with its execution-time safety shield (inner loop).

Algorithm 1. Bi-level Lyapunov-SAC dispatch and two-stage TPE sizing co-optimization.

Input: annual data D; capacity bounds X; warm-start sizing x0 (MILP);
cost budget d; coarse trials N1; refined trials N2; fine-tune windows W
Output: co-optimized sizing x*; dispatch policy π*
Stage 0 — warm start (one full training run)
θ0 ← train Lyap-SAC on D at sizing x0 for E episodes of 720-h windows,
applying the constrained update (Eqs. 2–3) and the shield (Eq. 4) each step
Stage 1 — coarse sizing search over full bounds
initialize TPE sampler S over X
for i = 1 … N1:
x ← S.suggest() # candidate capacities
π ← clone(θ0); fine-tune π for W windows at x under common random numbers
NAC ← full-year deterministic rollout of π at x # Eq. 5 objective

```
    S.observe(x, NAC)
x* ← argmin over observed NAC
# Stage 2 — refined search in a ×[1/1.6, 1.6] box around x*
restrict S to box B(x*)
for i = 1 … N2:
    x  ← S.suggest()
    π  ← clone(θ0); fine-tune π for W windows at x
    NAC ← full-year deterministic rollout of π at x
    S.observe(x, NAC)
x* ← argmin over all observed NAC;  π* ← policy fine-tuned at x*
return x*, π*
```

4.4. Evaluation metrics and baselines

Performance is reported with five metrics. The headline economic indicator is the net annualized cost defined by the objective in Section 4.3. The levelized cost of hydrogen is $\mathrm{LCOH} = (C_{\mathrm{cap}}^{\mathrm{H_2}} + C_{\mathrm{op}}^{\mathrm{H_2}})/\dot{M}_{\mathrm{H_2}}$ in \$/kg, where the electricity drawn from curtailed surplus is priced at zero — encoding the curtailment-recovery premise directly in the metric — and $\dot{M}_{\mathrm{H_2}}$ is the annual hydrogen produced. The curtailment-recovery rate $\mathrm{CRR} = \sum_t \Phi_t \,/\, \sum_t \Psi_t$ is the fraction of curtailable surplus Ψ_t that is routed to the electrolyzer. Reliability is reported as the loss-of-power-supply probability $\mathrm{LPSP} = \sum_t U_t \,/\, \sum_t L_t$ and, as a discrete safety count, the number of violation steps per 1000 steps, where a step is a violation when $\ell_t > 5 \times 10^{-3}$ (unmet load above 0.5% of peak). The annual green-hydrogen output and the CO_2-reduction fraction are reported for completeness.

The Proposed controller is compared against dispatch baselines that share the same network architecture and hyperparameters but differ in their constraint treatment: an unconstrained single-agent SAC, an unconstrained multi-agent SAC with centralized-training/decentralized-execution (MASAC), and a risk-averse Robust-SAC that constrains the conditional value-at-risk of the cost-to-go rather than its mean. For sizing, the co-optimized design is benchmarked against a perfect-foresight MILP/LP oracle (solved with HiGHS over a representative horizon) and a hybrid PSO–GA metaheuristic operating on a greedy rule-based dispatch, with the reinforcement-learning designer retained as an ablation. All methods in the primary comparison are evaluated at the single common co-optimized plant so that differences reflect dispatch quality, while the sizing study (Section 6.3) holds the dispatch policy fixed and varies only the capacities. The complete set of algorithm and training settings is given in Table 2.

Table 2. Algorithm and training hyperparameters.

Group	Setting	Value
SAC	Hidden layers / activation	2 × 256 / ReLU
SAC	Optimizer / learning rate	Adam / 3×10^{-4}
SAC	Discount γ/ target-smoothing τ	0.99 / 0.005
SAC	Batch size / replay capacity / warm-up	256 / 10^6/ 1000
SAC	Temperature	auto-tuned, target entropy $-\lvert \mathcal{A} \rvert$
Lyapunov	Cost budget d/ cost scale ζ	0.03 / 40
Lyapunov	λlearning rate / init / max	5×10^{-3}/ 1.0 / 150
Training	Episodes / window length	100 / 720 h
TPE sizing	Coarse / refined trials	50 / 50
TPE sizing	Refined box / fine-tune windows	$\times\,[1/1.6, 1.6]$/ 15
TPE sizing	Sampler / storage	Optuna TPE / SQLite (resumable)
Evaluation	Horizon / held-out seed / forecast noise	8760 h / 10000 / 0

Evaluation	Primary-comparison seeds	5

Results

All methods in the primary comparison are evaluated at a single common plant (the co-optimized sizing of pv 3134 kW, wind 59 kW, battery 354 kWh, electrolyzer 895 kW, and H_2 tank 157 kg) so that the comparison isolates dispatch quality from sizing quality. Each reinforcement-learning controller is trained on 100 calendar-aligned 720-h windows and evaluated by a deterministic full-year (8760-h) rollout on a held-out random seed with zero forecast noise. Unless otherwise stated, headline values are reported as the mean over five random seeds.

6.1. Real-time dispatch performance against baselines

The proposed Lyapunov-constrained SAC controller with the execution-time shield outperforms all three baselines on every reported metric, and it does so in all five seeds individually rather than only on average. Across the five seeds it attains a net annualized cost of \$560,205 (±2,745), against \$583,992 (±9,637) for Robust-SAC, \$586,461 (±3,955) for SAC, and \$591,591 (±6,070) for MASAC — a cost reduction of approximately 4.1%, 4.5%, and 5.3% relative to the three baselines, respectively. The advantage is far larger on safety: the proposed controller incurs 18.4 violations per 1000 steps versus 108–123 for the baselines, a roughly six-fold reduction in unmet-load events, while simultaneously delivering the lowest LCOH (4.20 \$/kg) and the highest curtailment-recovery rate (0.685). The error bars in Figure 3 also show that the proposed method is the most stable across seeds, with the tightest spread on cost and violations.

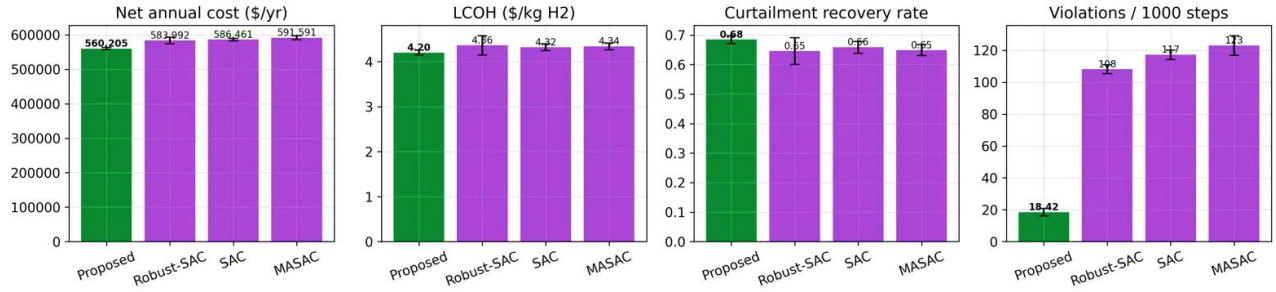

Figure 3. Real-time dispatch performance of the proposed Lyapunov-SAC controller versus the SAC, MASAC and Robust-SAC baselines at the common co-optimized sizing (five seeds; bars = mean, error bars = ±1 SD): net annualized cost, LCOH, curtailment-recovery rate, and violations per 1000 steps.

Because the comparison spans only five seeds, statistical significance was assessed with the Wilcoxon signed-rank test. For every metric against every baseline the test statistic is 0 — i.e. the proposed method is favored in all five paired seeds — which yields a two-sided p-value of 0.0625, the smallest value attainable at n = 5. We report this honestly: the perfect 5/5 ordering is the strongest possible outcome at this sample size, but the two-sided p-value cannot fall below 0.0625 without additional seeds (the corresponding one-sided value is ≈ 0.031). Table 3 summarizes the primary comparison together with the significance results.

Table 3. Primary five-seed comparison at the common co-optimized sizing (mean ± SD). The final column reports the two-sided Wilcoxon signed-rank p-value of the proposed method against each baseline (identical across the reported metrics because the proposed method wins all 5/5 paired seeds).

Method	Net annual cost ($)	LCOH ($/kg)	CRR	Viol./1k	LPSP	Wilcoxon p
Proposed (Lyap-SAC + shield)	**560,205 ± 2,745**	**4.20**	**0.685**	**18.4**	**0.1%**	—
Robust-SAC	583,992 ± 9,637	4.36	0.646	108.2	0.5%	0.0625
SAC	586,461 ± 3,955	4.32	0.659	117.3	0.5%	0.0625

MASAC	591,591 ± 6,070	4.34	0.649	123.1	0.5%	0.0625

6.2. Component ablation

To attribute the gains to individual components, each element of the proposed method was removed in turn (single seed, at the co-optimized sizing). The execution-time shield is the dominant safety lever: removing it increases violations 4.9-fold, from 18.4 to 90.6 per 1000 steps, and simultaneously degrades every economic metric — LCOH rises to 4.90 $/kg, the curtailment-recovery rate falls to 0.563, and net cost climbs to $601k. The Lyapunov cost-critic contributes a further safety improvement of roughly 1.7× on top of the shield (violations rise from 18.4 to 31.1 when the critic is removed), confirming that the learned constraint and the hard projection are complementary rather than redundant. The curtailment-recovery reward is a smaller refinement, mainly nudging the recovery rate and cost (18.9 vs 18.4 violations; 4.25 vs 4.20 $/kg). Notably, replacing the single-agent controller with the hierarchical multi-agent variant does not improve performance: the multi-agent design is essentially tied on cost and LCOH and marginally worse on violations (20.5 vs 18.4). This is the empirical basis for retaining a single-agent dispatch policy and demoting the multi-agent architecture to an ablation row.

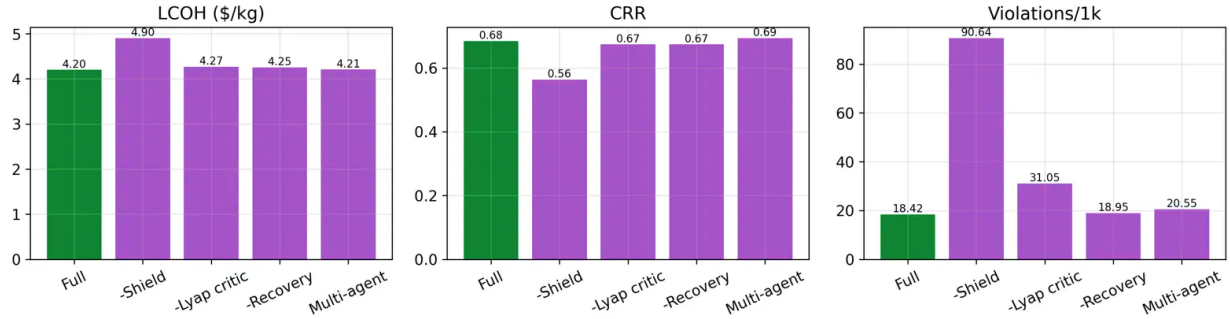

Figure 4. ablation of the proposed method (single seed at the co-optimized sizing): effect of removing the execution-time shield, the Lyapunov cost-critic, and the curtailment-recovery reward, and of replacing the single-agent controller with the hierarchical multi-agent variant (LCOH, CRR, violations/1k).

Table 4. Component ablation at the co-optimized sizing (single seed). Each row removes one element of the full method.

Variant	Net cost ($)	LCOH ($/kg)	CRR	Viol./1k
Full (Proposed)	**560k**	**4.20**	**0.685**	**18.4**
− Shield (= plain Lyap-SAC)	601k	4.90	0.563	90.6
− Lyapunov critic (= SAC + shield)	570k	4.27	0.674	31.1
− Recovery bonus	563k	4.25	0.674	18.9
Multi-agent (Hier-MASAC-Lyap)	566k	4.21	0.693	20.5

6.3. Contribution of the Bayesian sizing co-optimization

Holding the dispatch policy fixed and varying only the plant capacities isolates the contribution of the sizing layer. The two-stage log-scale TPE search produces a markedly better plant than either the MILP-oracle sizing or the legacy RL-designer sizing: at identical Lyapunov-SAC dispatch, the Bayesian sizing lowers LCOH by 42% (4.20 vs 7.25 $/kg) and net cost by 12% (560k vs 635k) relative to the MILP sizing, with the RL-designer sizing performing similarly to the MILP sizing (7.37 $/kg, 625k). The curtailment-recovery rate rises in step, from ~0.42 at the MILP/RL sizings to 0.685 at the Bayesian sizing. This sizing-level gain is larger than the dispatch-level gap of Section 6.1, which

identifies the sizing co-optimization — not the dispatch algorithm alone — as the dominant source of improvement in the full system.

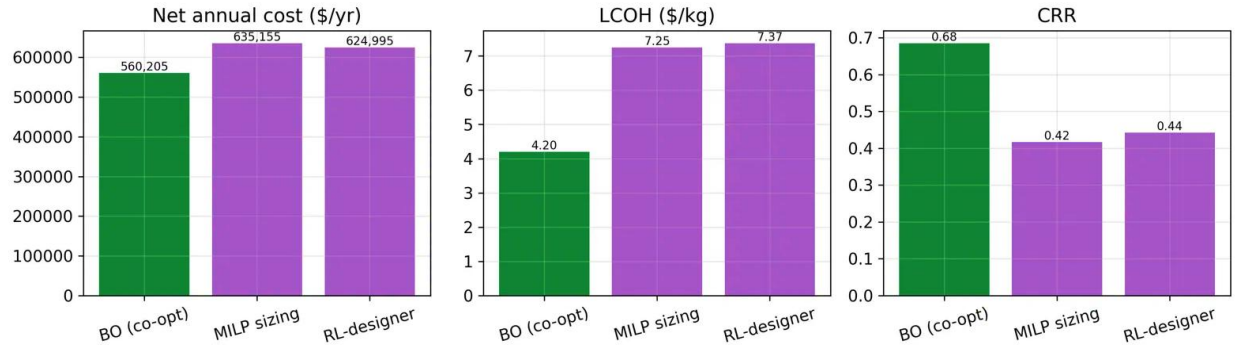

Figure 5.Contribution of the sizing strategy with the Lyapunov-SAC dispatch policy held fixed: Bayesian co--optimized sizing versus MILP-oracle and RL-designer sizings (net cost, LCOH, CRR).

The convergence behaviour of the search is shown in Figure 6. The two-stage TPE running-best descends from an initial ~780k to ~590k within the first 18 coarse trials and settles near 560k during the refined stage, falling and remaining below both the MILP-oracle reference (≈643k) and the RL-designer reference (≈622k). The wide scatter of individual trial costs against a steadily declining envelope is the expected signature of a global sampler that continues to explore while concentrating around the low-cost region.

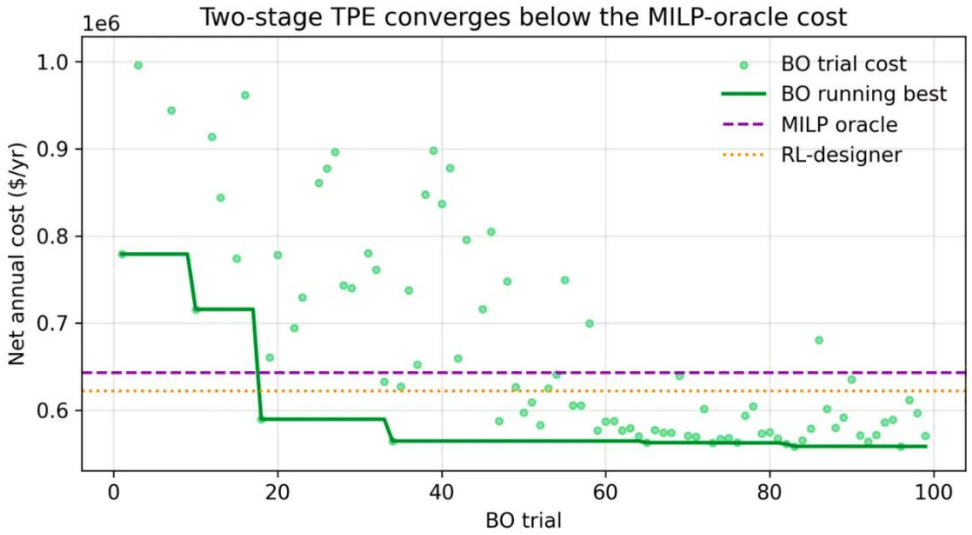

Figure 6. of the two-stage TPE Bayesian sizing search: the running-best net annualized cost falls below both the MILP-oracle and RL-designer reference costs.

6.4. Structure of the co-optimized design

The co-optimized design differs from the MILP-oracle design in a physically interpretable way rather than uniformly. As Figure 7 shows, the Bayesian search nearly eliminates wind (59 kW versus 1139 kW for the MILP oracle), roughly doubles the electrolyzer (895 kW versus 390 kW), and modestly enlarges PV and the hydrogen tank. In other words, the co-optimization reallocates capital away from a relatively expensive and, in this synthetic resource, less complementary wind asset and toward conversion and storage capacity that monetizes surplus PV as sold green hydrogen.

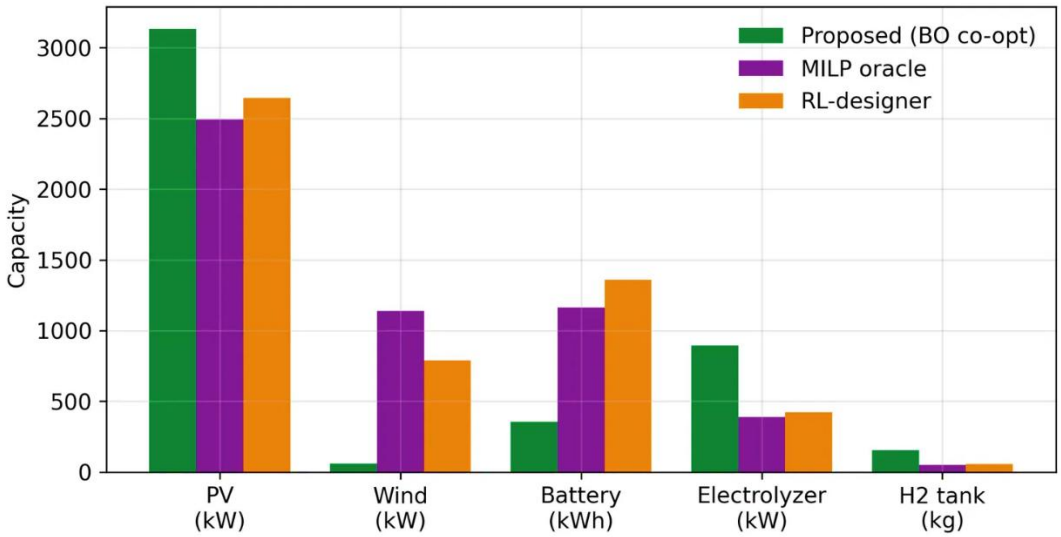

Figure 7. Co-optimized capacities of the adopted (BO) design compared with the MILP-oracle and RL-designer designs (PV, wind, battery, electrolyzer, H_2 tank).

Figure 8 reframes these differences as percentage deviations from the MILP sizing, annotated with each component's share of total CAPEX. The deviations are large precisely on the cost-significant dimensions — PV (41% of CAPEX) and the electrolyzer (14% of CAPEX) — and every component exceeds the 20% deviation gate. The hydrogen tank shows the largest percentage deviation but is de-emphasized because it constitutes only ~1% of CAPEX, so its large relative change is economically minor. The pattern indicates that the two design routes genuinely disagree on the capital-intensive components, which is what makes the co-optimization economically consequential.

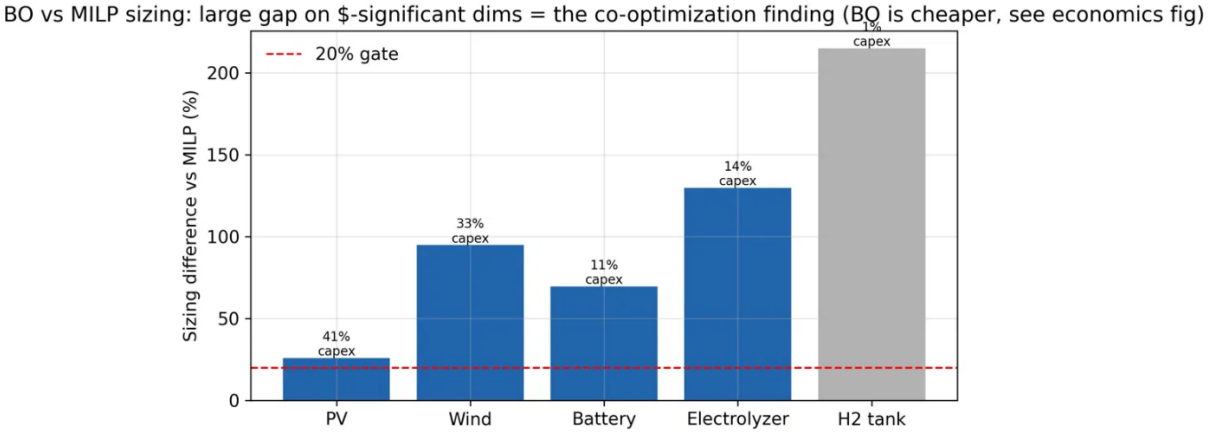

Figure 8. Per-component sizing difference between the Bayesian co-optimized design and the MILP-oracle design, annotated with each component's share of total CAPEX (dashed line: 20% deviation gate).

6.5. Economic and environmental profile of the adopted design

Figure 9 reports a deterministic full-year rollout of the three adopted designs across six economic and operational indicators. (The proposed values here — net cost $558,424, LCOH 4.17 $/kg, CRR 0.70 — are the single-design deterministic figures and differ slightly from the five-seed means of Table 3, which remain the headline values.) The co-optimized design is the cheapest and lowest-LCOH option and produces substantially more green hydrogen — 40,650 kg/yr versus 23,389 kg/yr (MILP oracle) and 23,931 kg/yr (RL-designer) — while raising the curtailment-recovery rate from ~0.41 to ~0.70. These benefits are not free, and we disclose the trade-off explicitly: the co-optimized design draws more energy from the capped grid (1.51 GWh/yr versus 0.65 GWh/yr for the MILP design), which lowers its CO_2-reduction fraction to 0.59 against 0.83 for the MILP design. Part of the cost and LCOH advantage

of the adopted design is therefore purchased with additional grid import and a corresponding loss of emissions benefit. Table 5 collects the adopted-design capacities together with these economic and environmental indicators.

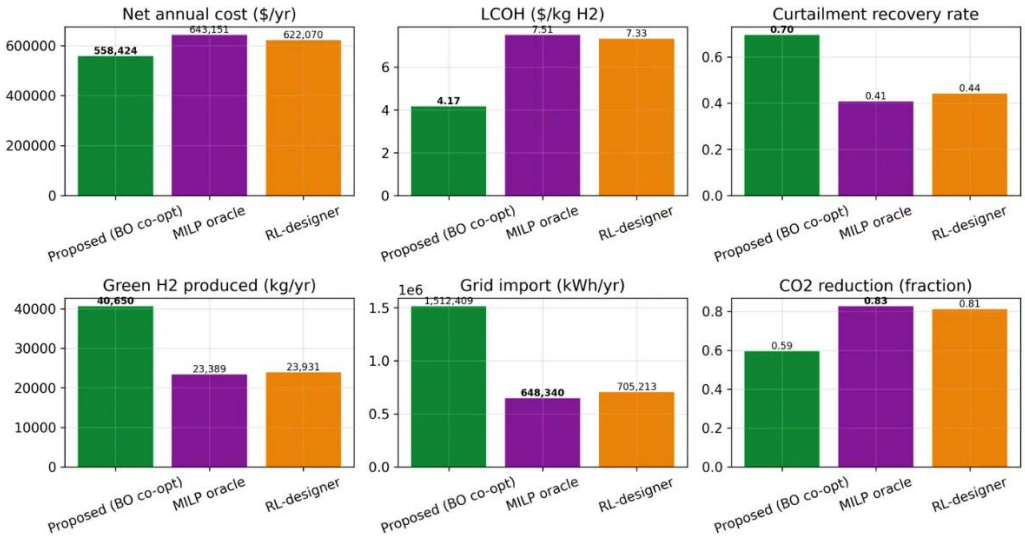

Figure 9. Economic and operational profile of the adopted (BO) design relative to the MILP-oracle and RL-designer reference designs (deterministic full-year rollout): net annualized cost, LCOH, curtailment-recovery rate, annual green-H_2 output, annual grid import, and CO_2-reduction fraction.

Table 5. Adopted (Bayesian co-optimized) design versus the MILP-oracle and RL-designer reference designs: capacities and full-year economic/environmental indicators.

Quantity	BO (adopted)	MILP oracle	RL-designer
PV (kW)	3134	2493	2647
Wind (kW)	59	1139	792
Battery (kWh)	354	1163	1360
Electrolyzer (kW)	895	390	423
H_2 tank (kg)	157	48	55
Net annual cost ($/yr)	558,424	643,151	622,070
LCOH ($/kg)	4.17	7.51	7.33
Curtailment recovery rate	0.70	0.41	0.44
Green H_2 (kg/yr)	40,650	23,389	23,931
Grid import (kWh/yr)	1,512,409	648,340	705,213
CO_2 reduction (fraction)	0.59	0.83	0.81

Discussion

The results are the result of two distinct mechanisms. The first is that the execution-time shield provides economic benefit from improved safety because in this environment an avoidable deficit is doubly penalized as both a value-of-lost-load operating cost and as a Lyapunov violation. The shield prevents the avoidable component of unmet load (the 4.9× reduction in safety violations in Section 6.2) by projecting the battery action onto the safe set when the capped grid could not cover the deficit, thus relieving the policy of the need to self-insure against deficits and letting the policy route more surplus to the electrolyzer, which is also why the curtailment-recovery rate drops and LCOH increases when the shield is removed. The Lyapunov cost-critic gives this hard projection a learned, anticipatory dimension on top, which is the source of the additional ~1.7× safety improvement. The second mechanism is that the bi-level Bayesian sizing dominates because the five-dimensional capacity decision is poorly informed by the single sizing sample per episode that an RL designer must work with; a dedicated global sampler that wraps the trained dispatch policy can search the capital-intensive PV/electrolyzer trade-off directly, which is both where the MILP and Bayesian methods most disagree (Figure 8) and where the economic stakes are highest.

Table 6 contrasts this work with five recent (2025–2026) studies that sit closest to it on its four design axes: safe reinforcement-learning dispatch, multi-agent learning, off-grid green-hydrogen sizing, and curtailment handling. The comparison is both in methodological design and reported outcomes, with a few caveats: levelized-cost definitions vary across studies in their system boundaries, electricity pricing, degradation treatment, and even currency, making cross-study LCOH figures illustrative but not exact; "curtailment-recovery rate," "curtailment reduction," and "renewable utilization" are three related but distinct quantities; and finally, our cost-reduction percentage is taken against strong safe-RL baselines at a fixed plant whereas a few of the literature percentages are taken against weaker rule-based, MPC, or unconstrained baselines.

Table 6. Comparison with recent related work

Study	System & setting	Learning / optimization method	Constraint / safety handling	Sizing ↔ dispatch	Reported headline result(s)
This work	Off-grid PV–wind–battery–PEM electrolyzer–H_2; synthetic 8760-h year	Single-agent SAC dispatch + bi-level TPE Bayesian sizing	Lyapunov cost-critic + Lagrange multiplier **and** hard execution-time projection shield	**Bi-level co-optimization**	LCOH 4.20 $/kg; CRR 0.685; 18.4 viol/1k (shield 4.9× reduction); −4–5% cost vs strong RL baselines; Bayesian sizing −42% LCOH vs MILP sizing
Zou et al. (2025)	Hybrid hydrogen–electric microgrid	Single-agent SAC dispatch	Lyapunov safety constraints (soft)	Dispatch only	−26.15% operating cost vs conventional control; markedly fewer unsafe actions
Cortés et al. (2026)	Three-phase unbalanced AC microgrid	Single-agent SAC dispatch	Lagrangian penalty on voltage/technical limits (soft)	Dispatch only	Higher operational return than QCP optimization; voltage kept within limits with a slight increase in phase unbalance
Abed et al. (2026)	Transmission-level grid-connected DRES + storage	Hierarchical multi-agent PPO + GNN	Reward shaping (no formal certificate)	Operation only	Curtailment −69.1% vs economic dispatch (−10.3% vs MPC); operating cost −27.9%; frequency within ±0.1 Hz for 97.3% of periods
Ibáñez-Rioja et al. (2025)	Off-grid PV–wind–battery–alkaline electrolyzer–H_2 cavern	Deterministic simulation + simultaneous capacity/control optimization (no RL)	Rule/optimization-based operating limits	**Simultaneous sizing + control**	LCOH below 3 €/kg (2025 prices); 30-yr, 15-min simulation with degradation; real Finnish resource data
Zhang et al. (2025)	Standalone wind–solar–H_2; AEL + PEMEL multi-electrolyzer	Techno-economic configuration optimization (no RL)	Operating-window constraints	Sizing/configuration only	PV–WT hybrid LCOH 4.52 $/kg (−41.1% vs PV-only); renewable utilization 92.26%

Limitations

These results should be read in the context of three limitations. First and foremost, the experiments use a generated 8760-h year rather than measured site data; this is the key missing link relative to the real-data studies above, and a measured-location case study is a natural next step. Second, the novelty is the integration of an existing Lyapunov-SAC dispatch family with bi-level Bayesian sizing and a green-hydrogen curtailment-recovery reward, not a new reinforcement-learning algorithm per se. Third, part of the adopted design's cost and LCOH advantage is secured through higher grid import, which reduces its CO_2-reduction fraction relative to the MILP design – a trade-off that a deployment with a tighter import cap or carbon-weighted objective would adjust.

Conclusion

This work introduced and analyzed a bi-level framework coupling two-stage Bayesian capacity sizing with a Lyapunov-constrained soft actor-critic dispatch policy and a green-hydrogen curtailment-recovery reward, for an off-grid PV–wind–battery–electrolyzer microgrid. By jointly optimizing sizing and operation and applying an execution-time safety shield, we obtain a design that is lower cost, lower cost of hydrogen, safer, and more curtailment-recovering than strong reinforcement-learning baselines and a perfect-foresight optimization oracle, while simultaneously increasing the amount of otherwise-curtailed energy that can be productively recovered into revenue-generating, saleable green hydrogen. In particular, the sizing layer is the dominant source of improvement. The obtained results show that the challenges of safe dispatch, co-optimized sizing, and curtailment recovery can be jointly met in a simultaneously feasible and beneficial way, and provide a useful template for designing practical, reliable, low-cost renewable-hydrogen systems.

References

[1] Cheekatamarla, P. (2024). Hydrogen and the global energy transition—Path to sustainability and adoption across all economic sectors. *Energies, 17*(4), 807. https://doi.org/10.3390/en17040807

[2] Angelico, R., Giametta, F., Bianchi, B., & Catalano, P. (2025). Green hydrogen for energy transition: A critical perspective. *Energies, 18*(2), 404. https://doi.org/10.3390/en18020404

[3] bin Jumah, A. (2024). A comprehensive review of production, applications, and the path to a sustainable energy future with hydrogen. *RSC Advances, 14*(36), 26404–26423. https://doi.org/10.1039/D4RA04559A

[4] Ibáñez-Rioja, A., Puranen, P., Järvinen, L., Kosonen, A., Ruuskanen, V., Hynynen, K., Ahola, J., & Kauranen, P. (2025). Baseload hydrogen supply from an off-grid solar PV–wind power–battery–water electrolyzer plant. *Energy, 322*, 135304. https://doi.org/10.1016/j.energy.2025.135304

[5] Zhang, W., Li, M., & Chen, Q. (2025). Optimizing standalone wind–solar–hydrogen systems: Synergistic integration of hybrid renewables and multi-electrolyzer coordination for enhanced green hydrogen production. *Processes, 13*(12), 3801. https://doi.org/10.3390/pr13123801

[6] Reichartz, T., Jacobs, G., Rathmes, T., Blickwedel, L., & Schelenz, R. (2024). Optimal position and distribution mode for on-site hydrogen electrolyzers in onshore wind farms for a minimal levelized cost of hydrogen. *Wind Energy Science, 9*(2), 281–295. https://doi.org/10.5194/wes-9-281-2024

[7] Oueslati, F. (2023). HOMER optimization of standalone PV/Wind/Battery powered hydrogen refueling stations located at twenty selected French cities. *International Journal of Renewable Energy Development, 12*(6), 1070–1090. https://doi.org/10.14710/ijred.2023.58218

[8] Feitosa, F. E. B., & Costa, A. L. (2025). Simulation and evaluation of a large-scale electrolysis plant — A case study in Pecém, a Brazilian port. *Renewable Energy, 250*, 123247. https://doi.org/10.1016/j.renene.2025.123247

[9] Liu, H., Clausen, L. R., Wang, L., & Chen, M. (2023). Pathway toward cost-effective green hydrogen production by solid oxide electrolyzer. *Energy & Environmental Science, 16*(5), 2090–2111. https://doi.org/10.1039/D3EE00232B

[10] Geng, Y., Liu, Q., Zheng, H., & Yan, S. (2025). Two-stage collaborative power optimization for off-grid wind–solar hydrogen production systems considering reserved energy of storage. *Energies, 18*(11), 2970. https://doi.org/10.3390/en18112970

[11] Saars, L., Madsen, M., & Meyer, J. (2024). Optimizing the operation of an electrolyzer with hydrogen storage using two different methods: A trade-off between simplicity and precision in minimizing hydrogen production costs using day-ahead markets. *Energies, 17*(22), 5546. https://doi.org/10.3390/en17225546

[12] Medghalchi, Z., & Taylan, O. (2023). A novel hybrid optimization framework for sizing renewable energy systems integrated with energy storage systems with solar photovoltaics, wind, battery and electrolyzer–fuel cell. *Energy Conversion and Management, 294*, 117594. https://doi.org/10.1016/j.enconman.2023.117594

[13] Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). *Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor.* arXiv. https://doi.org/10.48550/arXiv.1801.01290

[14] Gao, J., Li, Y., Wang, B., & Wu, H. (2023). Multi-microgrid collaborative optimization scheduling using an improved multi-agent soft actor-critic algorithm. *Energies, 16*(7), 3248. https://doi.org/10.3390/en16073248

[15] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal policy optimization algorithms.* arXiv. https://doi.org/10.48550/arXiv.1707.06347

[16] Abed, A. M., Madaminov, S., Abduvokhidov, A., Khudoynazarov, E., & Ibrahim, W. (2026). Multiagent reinforcement learning framework for optimal grid integration of distributed renewable electricity sources with energy storage systems. *International Journal of Low-Carbon Technologies, 21*, ctaf142. https://doi.org/10.1093/ijlct/ctaf142

[17] Tessler, C., Mankowitz, D. J., & Mannor, S. (2018). *Reward constrained policy optimization.* arXiv. https://doi.org/10.48550/arXiv.1805.11074

[18] Cortés, P., Tabares, A., Bolaños, R., & Bonilla, K. (2026). Soft actor-critic energy management in three-phase unbalanced microgrids with Lagrangian penalty constraints. *Scientific Reports, 16*(1), 18109. https://doi.org/10.1038/s41598-026-47679-0

[19] Chow, Y., Nachum, O., Duenez-Guzman, E., & Ghavamzadeh, M. (2018). *A Lyapunov-based approach to safe reinforcement learning.* arXiv. https://doi.org/10.48550/arXiv.1805.07708

[20] Hao, G., Li, Y., Li, Y., Jiang, L., & Zeng, Z. (2024). Lyapunov-based safe reinforcement learning for microgrid energy management. *IEEE Transactions on Neural Networks and Learning Systems.* Advance online publication. https://doi.org/10.1109/TNNLS.2024.3496932

[21] Zou, Y., Zhang, A.-A., Zhang, X., Yang, W., Lin, Y., & Liu, Q. (2025). *Lyapunov-based safe reinforcement learning for managing hybrid hydrogen-electric energy system* [Working paper]. SSRN. https://doi.org/10.2139/ssrn.5772850

[22] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2623–2631). ACM. https://doi.org/10.1145/3292500.3330701

[23] Ibáñez-Rioja, A., Järvinen, L., Puranen, P., Kosonen, A., Ruuskanen, V., Hynynen, K., Ahola, J., & Kauranen, P. (2023). Off-grid solar PV–wind power–battery–water electrolyzer plant: Simultaneous optimization of component capacities and system control. *Applied Energy, 345*, 121277. https://doi.org/10.1016/j.apenergy.2023.121277