

Energy Management Strategy for Triple Hybrid Energy Storage Systems in Smart Grids

Rewaa Elfaham¹, Mostafa Hamad² and Ali Karimi³

¹Mechanical Engineering, Arab Academy for Science, Technology and Maritime Transport, Alexandria, Egypt

²Electrical Engineering, Arab Academy for Science, Technology and Maritime Transport, Alexandria, Egypt

³Department of Energy and Petroleum Engineering, University of Stavanger (UiS), Stavanger, Norway

E-mail: rewaa.mohammad3@gmail.com

Abstract. A triple hybrid storage plant — battery, supercapacitor, and hydrogen fuel cell with electrolyser — runs on the IEEE 33-bus network for a year under eight strategies: a pro-rata baseline, a rule-based controller, four reinforcement learning agents and two model-predictive controllers. The two differ in one thing only: whether their objective prices stored hydrogen. That change moves terminal reserve from 0.001 to 0.935 and costs 1.7 points of annual load deficit (unserved locally, on a feeder whose substation stays available; section 2.4). The variant that ignores the reserve posts the lowest deficit in the study, 20.08%, and empties the tank while issuing no infeasible command, so a feasibility audit alone does not catch it. The reserve-aware variant serves more load and holds more hydrogen than every learned controller, with complete separation on both — though the optimisers see a forecast the learned policies do not, and both are the least voltage-compliant controllers here. Removing the fuel cell during low-renewable spells costs 8 to 9 points of unserved load when the tank is healthy, almost nothing when empty. Report terminal reserve alongside the delivered fraction of commanded energy: an outcome score alone cannot tell a dispatch that ran from one the plant refused.

Nomenclature

Abbreviations. T-HESS, triple hybrid energy storage system; BESS, battery energy storage system; SC, supercapacitor; FC, fuel cell; EMS, energy management system; DRL, deep reinforcement learning; SAC, soft actor-critic; DDPG, deep deterministic policy gradient; TD3, twin delayed DDPG; PPO, proximal policy optimisation; MPC, model-predictive control (MPC-D, deficit objective; MPC-R, reserve-aware); DoD, depth of discharge; pu, per unit; mpu, milli-per-unit.

Symbols. SoC, battery state of charge; SoE, supercapacitor state of energy; SoH_{H₂}, hydrogen tank level; SoH, battery state of health; P_L , P_{res} , load and renewable power; P_{def} , P_{curt} , unserved and curtailed power; P_{cmd} , P_{act} , commanded and delivered power; Δf , frequency deviation; V_{oc} , V_{term} , open-circuit and terminal voltage; R_0 , R_i , C_i , ohmic and RC branch parameters; η_c , η_{elec} , η_{FC} , coulombic, electrolyser and fuel-cell efficiency; $W_\bullet$, reward weights; U , p , Mann–Whitney statistic and p -value.

1 Introduction

Distribution feeders with a lot of solar and wind have an imbalance problem that no single storage technology solves well. The mismatch happens on three timescales at once. Clouds and gusts cause ramps within the minute. The daily cycle moves a midday surplus into an evening peak. And a few times a year a low-renewable spell lasts several days and drains whatever the shorter-duration devices are holding.

A battery covers the middle timescale well and the other two badly. A supercapacitor normally handles ramps at high power for minutes rather than hours. The unit modelled here does not: at 4.44 MWh and 400 kW it is a second medium-duration branch, set apart from the battery by power rating and efficiency rather than by timescale (section 5.3). The sub-minute timescale is therefore something this plant class aims at, not something this sizing reaches. Hydrogen is the only one of the three that spans days, stored as compressed gas and recovered through a fuel cell at a round-trip efficiency near 0.45.

Putting all three together gives a triple hybrid energy storage system, or T-HESS, whose branches are good at different things. That shifts the hard part from sizing to dispatch. The controller has to decide not just how much power to move but which device should move it, and the right answer changes with the state of every reserve.

The objectives conflict. Serving load immediately argues for discharging the fuel cell whenever the battery is low; preserving the hydrogen reserve argues for the opposite. Depletion is also irreversible on any useful horizon, because a tank emptied in March cannot be refilled in time to help in November: the electrolyser runs only on surplus, and surplus is scarce in exactly the conditions where the reserve is wanted. A controller treating the fuel cell as just another power source empties it early in the year and spends the remaining months operating a battery-and-supercapacitor system with an expensive idle stack attached.

This is easy to miss, because it does not show up in the metric most studies report. Annual load deficit rewards a controller for spending reserves it will not replace, and is silent about whether the dispatch that produced it was one the plant could execute. Both appear here directly: the naive pro-rata baseline posts a lower annual deficit than a carefully hand-tuned rule-based controller while emptying the hydrogen tank within the first twenty weeks and issuing fuel-cell commands the plant refuses 44.9% of the time. Any comparison that stops at the annual scalar draws the wrong conclusion.

1.1 *Related work and the gap addressed here*

Hybrid storage combining a battery with a supercapacitor is well established, and the design rules for splitting power between them by frequency content are mature [1]. Adding hydrogen introduces a duration dimension those systems do not have; recent duration-taxonomy work places supercapacitors and flywheels below one hour of discharge and identifies chemical storage as the only generally available option for seasonal-scale mismatch [2].

Deep reinforcement learning has been applied to microgrid energy management with consistent gains over rule-based dispatch [3, 4], and soft actor-critic in particular suits continuous multi-device control [5, 6]. Recent work extends this to hybrid plant: coordinated battery-supercapacitor dispatch under renewable variability [7], and electric-hydrogen systems run through safe or constrained formulations that keep the policy inside operating limits [8, 9, 10]. Table 1 places twelve of these against this study.

The pattern is one of scope. Storage is usually a single asset or a pair, the control comparison stays within one family, and the reported quantity is an aggregate outcome. This paper supplies what that table shows missing: three storage branches separated by power rating and round-trip efficiency, a horizon long enough for depletion to show, eight strategies sharing one environment and one evaluation protocol, and the reserve and feasibility reporting no comparable study provides.

Pricing a stored quantity at the end of a finite horizon is not new: it is the terminal cost of constrained model-predictive control [11]. The same three options appear in recent unit-commitment work

Table 1. Recent reinforcement-learning studies of storage dispatch, and where this work sits among them. The last two columns are what this study adds: terminal reserve reported alongside the aggregate outcome, and commanded energy reconciled against delivered.

Study	Yr	Storage	Control	Reserve	Feas.
[1]	24	battery + SC	review	—	—
[2]	23	multi-technology	duration study	—	—
[3]	25	hybrid ESS	DRL	—	—
[4]	25	hybrid + H ₂	DRL	—	—
[7]	26	battery + SC	DRL	—	—
[8]	25	electric-H ₂	safe RL	—	—
[9]	26	electricity-H ₂	constrained RL	—	—
[10]	24	aggregated VPP	safe RL	—	—
[14]	25	battery	DRL, 4 algs	—	—
[15]	26	battery + H ₂	RL metaheuristic	—	—
[16]	23	PV + battery	shielded MARL	—	—
[17]	26	battery	explainable DRL	—	—
This work	26	battery + SC + H ₂	4 RL, 2 MPC, 2 rules	yes	yes

on storage, where end-of-horizon effects are met by taking no view of the future, by enforcing a terminal state target, or by pricing deviations from one [12] — respectively MPC-D, section 4.3’s restart from the reserve each controller leaves behind, and MPC-R. Receding-horizon dispatch of storage is an active area in its own right [13]. What this paper adds is the reporting around the constraint rather than the constraint: a terminal condition fixes one controller’s behaviour and says nothing about comparing controllers that do not share it, which is the situation whenever learned and optimising policies are placed side by side.

1.2 Contributions

1. An 8760-hour T-HESS environment on the IEEE 33-bus network — second-order Thevenin battery, series-RC supercapacitor as a second medium-duration branch, ideal-gas hydrogen store, LinDist-Flow power flow, zero-order-hold swing equation — with the energy-balance self-tests that establish its conservation properties (section 2.5).
2. A command-feasibility audit accumulating commanded against delivered energy per device, which finds 44.9% of the baseline’s hydrogen dispatch refused by the plant while no other strategy exceeds 0.2% — a distinction no outcome metric records.
3. An eight-strategy benchmark under one protocol, including two model-predictive controllers that differ only in whether their objective prices the reserve, and on which the non-dominated set for the deficit-reserve pair contains both of them and nothing else.
4. A multi-day counterfactual pricing the hydrogen branch at 8.02 to 8.95 percentage points of avoided unserved load when the reserve is present, stable across three window lengths, and at 2.0 to 2.8 MWh — roughly a seventh as much — once prior dispatch has drawn the tank down.
5. A steady-state comparison restarting each controller from the reserve it leaves behind, identifying annual-scalar reporting as the reason the naive baseline appears competitive.

2 Simulation environment

2.1 Network and plant configuration

The test network is the IEEE 33-bus radial distribution system of Baran and Wu [18], at 12.66 kV on a 10 MVA base, with branch impedances and bus loads from the updated benchmark data of Dolatabadi *et al.* [19]. Total system load is 3715 kW and 2300 kvar. The T-HESS and both renewable sources connect at bus 18, the far end of the main feeder and the weakest point of the network, which makes it the most demanding location for a storage plant.

The plant is arranged as follows. Photovoltaic and wind generation, the three storage branches and the load meet at a common DC bus tied to bus 18 through a single grid-side inverter. Each storage device has its own bidirectional converter, so the energy management system sets three independent power commands per step. Ratings are collected in table 2 and were fixed once, in the environment build, and held constant across every strategy, so that all differences reported later are attributable to dispatch alone.

Table 2. T-HESS plant, generation and network parameters. All values are held constant across every dispatch strategy compared in this paper.

Subsystem	Parameter	Value	Source
Battery (BESS)	Rated power	1200 kW	design
	Rated capacity	8000 Ah at 1200 V (≈ 9.6 MWh)	design
	Ohmic resistance R_0	5 m Ω	[20]
	Fast branch (R_1, C_1)	2 m Ω , 20 kF ($\tau_1 \approx 40$ s)	[20]
	Slow branch (R_2, C_2)	3 m Ω , 80 kF ($\tau_2 \approx 240$ s)	[20]
	Coulombic efficiency	0.98	[22]
	SoC window	[0.20, 0.90]	design
	Cycle life	3000 at 80% DoD	design
Supercapacitor	Rated power	400 kW	design
	Capacitance	50 000 F	[23]
	Maximum cell voltage	800 V	[23]
	Series resistance	0.5 m Ω	[23]
	SoE window	[0.10, 0.95]	[23]
Hydrogen (HFC)	Rated fuel-cell power	1000 kW	design
	Electrolyser efficiency	0.75	[24]
	Fuel-cell efficiency	0.60	[25]
	Higher heating value	141.8 MJ kg $^{-1}$	[26]
	Tank volume	500 m ^{3}	design
	Pressure range	5 to 350 bar	[27]
Generation	PV efficiency, area	0.18, 12 000 m ^{2}	[28]
	Wind rated power	3500 kW	[29]
	Wind rated speed	12 m s $^{-1}$	[30]
Network	Base voltage, power	12.66 kV, 10 MVA	[18]
	Total load	3715 kW, 2300 kvar	[19]
	Voltage limits	[0.95, 1.05] pu	[19]
	Inertia constant H	5 s	[31]
	Damping coefficient D	0.05 pu	[32]

2.2 Storage models

Battery. A second-order Thevenin equivalent circuit gives terminal voltage $V_{\text{term}} = V_{\text{OC}}(\text{SoC}) - IR_0 - V_{C_1} - V_{C_2}$, the open-circuit voltage being a fifth-order polynomial in state of charge fitted to published cell data [21]. The RC time constants, 40 s and 240 s, are far shorter than the one-hour step, so each branch is integrated with its exact zero-order-hold solution rather than by explicit Euler, which would be unstable:

$$V_{C_k}(\Delta t) = IR_k + (V_{C_k,0} - IR_k) e^{-\Delta t/\tau_k}, \quad (1)$$

State of charge follows Coulomb counting, and a cumulative depth-of-discharge counter tracks state of health as $\text{SoH} = \max(0, 1 - \text{DoD}_{\text{cum}}/(2N_{\text{cyc}} \times 0.80))$.

Supercapacitor. A series-RC model gives terminal voltage $V_{\text{SC}} = Q/C_{\text{SC}} - IR_{\text{ESR}}$ and state of energy $\text{SoE} = V_{\text{SC}}^2/V_{\text{max}}^2$.

Hydrogen. The electrolyser converts surplus electrical power to a hydrogen mass flow $\dot{m}_{\text{in}} = \eta_{\text{elec}} P_{\text{elec}}/\text{HHV}_{\text{H}_2}$, and the fuel cell reverses the conversion as $P_{\text{FC}} = \eta_{\text{FC}} \dot{m}_{\text{out}} \text{HHV}_{\text{H}_2}$. Tank pressure follows the ideal gas law at constant temperature, and the hydrogen state of health is the normalised pressure $\text{SoH}_{\text{H}_2} = (P_{\text{H}_2} - P_{\text{min}})/(P_{\text{max}} - P_{\text{min}})$. This single scalar carries the plant's long-duration reserve and is the quantity around which most of the results in this paper turn.

2.3 Network solution and frequency

Bus voltages are obtained from a linearised DistFlow solution over the radial topology,

$$V_j \approx V_i - (r_{ij}P_{ij} + x_{ij}Q_{ij})/V_0, \quad (2)$$

evaluated by a backward-forward sweep, which returns voltage magnitudes at all 33 buses without Newton–Raphson iteration. That matters because the solver runs 8760 times per evaluation episode and roughly 1.3 million times over a training run. Validation against the base-load case gives a minimum bus voltage of 0.9159 pu at bus 18, against 0.913 pu from a full AC solution of the same benchmark; the 0.003 pu offset is the expected bias of a lossless linearisation. The benchmark’s base case is itself outside the [0.95, 1.05] pu band at nominal load, which is part of why it is a standard test feeder.

Frequency deviation is obtained from the swing equation with load damping, $2H\dot{\omega}/\omega_0 = P_m - P_e - D\Delta\omega$, integrated over the hourly step with its exact zero-order-hold solution. The resulting Δf should be read as a quasi-steady indicator of hourly power imbalance, not as a representation of sub-second frequency dynamics, which lie below the resolution of an hourly dispatch model. It is used to compare strategies against the 0.5 Hz ENTSO-E limit [33] on a like-for-like basis; its limitations are taken up in section 5.3.

2.4 Load and renewable profiles

Load over the year follows a base level of 2 MW modulated by a diurnal and a seasonal harmonic with additive noise,

$$P_{\text{load}}(t) = 2\text{MW} \left[1 + 0.3 \sin \frac{2\pi(h-6)}{24} + 0.15 \sin \frac{2\pi d}{365} + \varepsilon(t) \right], \quad (3)$$

with $\varepsilon \sim \mathcal{N}(0, 0.05)$, clipped to [0.5, 3.5] MW. The IEEE 33-bus static bus loads are scaled by the ratio of instantaneous to nominal load at each step, so the network solution stays consistent with the temporal profile. Photovoltaic output uses the standard irradiance and cell-temperature relation of IEC 61215 [28], and wind uses the piecewise cubic power curve of IEC 61400-1 [29]. Over the simulated year mean load is 1.999 MW against mean renewable generation of 1.636 MW, an 82% penetration. All profiles are generated once from seed 42 and reused unchanged by every strategy.

Load appears in two unrelated forms here. The 2 MW-mean profile above is a local service obligation the T-HESS and its renewables must meet without further import; every deficit, curtailment and reserve figure is computed against it alone. The Baran and Wu 3715 kW total instead scales the *relative* loading the voltage solve distributes across the feeder, with an infinite slack bus at the substation as in the original benchmark. A reported deficit therefore measures import avoidance against a fixed local target — a functioning substation could always supply the shortfall — while the voltage solve asks a separate question the deficit figure says nothing about: whether the T-HESS’s own injection keeps every bus inside [0.95, 1.05] pu. Section 4.6 already shows dispatch is invariant to which feeder answers that, so the two readings are deliberately decoupled rather than reconciled.

2.5 Corrections to the simulation environment

Six defects were identified in the simulation environment after an earlier version of this study was circulated; table 3 gives each one, and every result reported here is computed on the corrected model. Two affected only the network solution. The base-load figure of 0.9956 pu quoted in an earlier version was a symptom of the DistFlow sign error, not evidence against it; the corrected solver returns 0.9159 pu at bus 18, against 0.913 pu from a full AC solution of the same benchmark. The other four affect the power balance behind every result in section 4.

The correction is quantitative, not just directional. The plant starts the year holding 169.8 MWh of usable inventory and loses 572.6 MWh to hydrogen round-trip inefficiency, so it must be a net energy sink

of at least 402.8 MWh. The previously reported baseline implied +1382.5 MWh, which exceeds that bound by 1785 MWh. That figure was not merely inaccurate but energy-infeasible. The corrected model returns −520.5 MWh and closes its annual balance to within 0.001 MWh, attributing 1362.0 MWh of the change to the hydrogen branch, 539.3 MWh to the supercapacitor and 7.3 MWh to the battery rating. The defect was invisible in annual deficit and visible only in the terminal-reserve trajectory — the diagnostic this paper argues for is what exposed it.

Table 3. The six defects, the model step each affected, and the direction of the error. Defects 1 and 2 move bus voltages only, since load deficit, curtailment, reserve and frequency come from a power balance that never consults the network solve. Defects 3 to 6 move everything.

#	Model step affected	Error	Consequence
1	LinDistFlow backward sweep	child flow added, not subtracted	downstream demand cancelled upstream; all voltages
2	Network injection at bus 18	served load entered twice	all voltages
3	Supercapacitor current clamp	P_{act} from pre-clamp current	full 400 kW from an empty bank
4	Supercapacitor ESR sign	$V + IR$ instead of $V - IR$	terminal voltage rose under discharge
5	Hydrogen tank availability	no check before discharge	1362 MWh delivered from an empty tank
6	Battery converter rating	rating not enforced	baseline ran to 1.75 MW against 1200 kW

3 Dispatch strategies

Three families of controller share this environment. Each receives the same observation and returns three continuous power commands, one per storage device, at every hourly step.

3.1 Pro-rata baseline

The baseline splits any deficit or surplus by fixed fractions across the three devices, allocating 30% of every deficit to the fuel cell regardless of tank state. It uses no state feedback beyond enforcing device limits. Its purpose is to establish what the plant achieves with no intelligence at all, and it proves to be a more interesting reference than expected.

3.2 Two-layer rule-based controller

The second controller adds engineering judgement in two layers. A finite state machine classifies each hour into one of eleven operating states from the net power imbalance, the three storage states and the bus-18 voltage: charging states, discharging states, a hydrogen production state, a voltage support state and an emergency state. Within a state, power is allocated in proportion to each device’s relative headroom,

$$m_k = \text{clip}\left(\frac{x_k - x_{k,\min}}{x_{k,\max} - x_{k,\min}}, 0, 1\right), \quad k \in \{\text{BESS}, \text{SC}, \text{HFC}\}, \quad (4)$$

with base fractions of 0.52, 0.26 and 0.22 tilted by ± 8 percentage points according to how each margin compares with the mean. A dynamic cap tapers fuel-cell rated power linearly once SoH_{H_2} falls below 0.20 and blocks discharge entirely below 0.10, which is the mechanism that stops this controller from draining the tank.

A second layer refines the result near nominal conditions. A 45-rule fuzzy controller maps the net imbalance, battery state of charge, hydrogen state of health and bus voltage onto three per-unit power commands. It uses Mamdani inference and Sugeno weighted-average defuzzification, and its output is blended with the state machine’s at weight 0.40 when the system is in its normal state or voltage is within 2% of a limit. The fuzzy layer is active for 1.1% of the year; the state machine handles the rest. Fourteen unit tests covering every state transition and the fuzzy output directions all pass.

3.3 Deep reinforcement learning

The learning agents see an eight-dimensional normalised state,

$$s(t) = \left[\text{SoC}, \text{SoE}, \text{SoH}_{\text{H}_2}, \frac{P_L}{P_{\text{scale}}}, \frac{P_{\text{res}}}{P_{\text{scale}}}, \frac{\Delta f}{\Delta f_{\text{scale}}}, \sin \frac{2\pi h}{24}, \cos \frac{2\pi h}{24} \right], \quad (5)$$

with $P_{\text{scale}} = 3.0$ MW and $\Delta f_{\text{scale}} = 0.5$ Hz. Hour of day enters as a sine–cosine pair so that hour 23 and hour 0 are adjacent in state space. The action is $a \in [-1, 1]^3$, scaled by each device’s rated power, positive discharging. Device limits are enforced by the environment, so the policy never has to learn hard physical constraints, only how to use the headroom it has. Note that bus voltage is not observed; section 5.3 returns to this.

The reward combines four terms,

$$r = W_{\text{service}} r_{\text{service}} + W_{\text{H}_2} r_{\text{H}_2} + W_{\text{freq}} r_{\text{freq}} + W_{\text{cost}} r_{\text{cost}}, \quad (6)$$

with $W_{\text{service}} = 1.0$, $W_{\text{H}_2} = 2.0$, $W_{\text{freq}} = 0.2$ and $W_{\text{cost}} = 0.3$. All four terms are bounded on a common scale, so the weights set a priority order rather than trading unlike units. Frequency’s weight reflects its role as a training-time smoothing signal rather than an evaluation discriminator: section 4.2 finds it correlates almost perfectly with deficit once policies converge ($r = 0.99$), so a discriminating training signal and a redundant evaluation column are not the same thing. Hydrogen carries the largest weight because reserve preservation is the behaviour under study. Soft actor–critic was retrained at $W_{\text{H}_2} \in \{0.5, 1, 4\}$ against this 2, five seeds at each, sharing the environment, episode budget and evaluation protocol of section 4.2. Mean deficits are 25.74, 25.74, 26.16 and 25.03%, and mean terminal reserves 0.185, 0.189, 0.189 and 0.173. Both ranges are about the size of the seeds’ own spread at each weight, which runs from 0.91 to 2.06 points of deficit and 0.0085 to 0.0239 of reserve. None of the twelve exact Mann–Whitney comparisons between weight pairs separates, and neither objective rises or falls steadily with the weight. The choice of W_{H_2} is therefore not resolvable on this problem at five seeds, and the results below do not rest on a tuned weight. Service and frequency are bounded linear penalties, $r_{\text{service}} = -P_{\text{def}}/P_L$ and $r_{\text{freq}} = -\text{clip}(|\Delta f|/\Delta f_{\text{scale}}, 0, 1)$, both in $[-1, 0]$.

3.3.1 The hydrogen term and why its shape matters. A hard-clipped penalty on SoH_{H_2} produces a dead zone: below the threshold every level of depletion scores alike, so an agent that emptied the tank gets no local signal that refilling would help, and recovery falls to slow long-horizon credit assignment. The design adopted here is instead a log barrier with a C^1 linear tail. With $x = \text{SoH}_{\text{H}_2}/\text{SoH}_{\text{H}_2, \text{target}}$ and a target of 0.10,

$$r_{\text{H}_2}(x) = \begin{cases} \min(\ln x, 0.5), & x \geq e^{-1}, \\ -1 + e(x - e^{-1}), & x < e^{-1}. \end{cases} \quad (7)$$

The pieces match in value and slope at the knee. The function is continuous, strictly increasing and bounded below at -2 , and near total depletion its recovery slope is $e/0.10 \approx 27$ per unit of SoH_{H_2} , so every increment of hydrogen improves the next reward. The cap at $+0.5$ prevents hoarding at the expense of load service.

3.3.2 Operating cost. The fourth term prices the dispatch itself, so technical targets are not met through wasteful actuation. Per-unit costs are 0.0625 \$/kWh of battery throughput, 0.05 of fuel-cell output, 0.02 of electrolyser input, 0.002 of supercapacitor throughput and 0.04 of curtailed energy, each amortised over the relevant rated life. The term is $r_{\text{cost}} = -\min(\text{cost}/C_{\text{ref}}, 1)$ with $C_{\text{ref}} = 150$ \$/h. Lost load at 5 \$/kWh is kept out of the reward, since unserved energy is already penalised by r_{service} , but is retained for the cost reporting in section 4.5.

3.3.3 Agents and training. Four algorithms were trained: DDPG [34], TD3 [35], PPO [36] and soft actor–critic [5, 6]. All share the environment, reward, state encoding and action decoding, and all use two hidden layers of width 256 with ReLU, learning rate 3×10^{-4} , discount 0.99, batch size 256 and 150 episodes of 8760 steps. The off-policy methods share a 150 000-transition replay buffer and Polyak coefficient 0.005; soft actor–critic adds twin critics, LayerNorm and automatic entropy tuning to a target of -3 nats. The three off-policy agents pre-fill their buffer with one year of rule-based trajectories before the first gradient step, which is what allows convergence within 150 episodes; PPO, being on-policy, cannot. Evaluation always uses the deterministic policy at the end of training, scored with the routine used for the rule-based controllers.

3.4 Model-predictive control

A receding-horizon controller was added so the learned policies face a classical constrained optimiser, not only hand-tuned rules. Each hour it solves a linear program over a 48-hour horizon, applies the first action and re-solves. Its internal model is an energy balance — each device a state, a charge and a discharge power and a per-unit-energy conversion, bounded by the ratings and state windows of table 2 — while it is scored on the same nonlinear plant as every other strategy, so the model mismatch is real rather than assumed away.

Two objectives are used, and the difference between them is the point. Both minimise unserved energy over the horizon under a small curtailment penalty. *MPC-D* stops there, optimising the very metric this paper argues is misleading. *MPC-R* adds a linear terminal value on stored hydrogen, priced at the electrical energy the reserve can return, 331.7 MWh per unit of SoH_{H_2} ; this price is swept from 0 to $2 \times$ in figure 1, which finds a threshold rather than a smooth trade-off. The reward the learned agents maximise places a logarithmic barrier on hydrogen, which no linear program can represent; MPC-R’s terminal value is a surrogate for that barrier, not a copy of it.

Forecast quality is varied rather than assumed away. The reported results use an error growing as $0.08\sqrt{h}$ on renewables and $0.02\sqrt{h}$ on load, which reaches day-ahead accuracies typical of operational practice. Perfect foresight and a naive next-day persistence forecast bracket it.

4 Results

4.1 Annual benchmark

Table 4 reports the headline indicators for all eight strategies. Load deficit is annual unserved energy as a fraction of annual demand, and hydrogen state of health is the terminal tank level.

Table 4. Annual benchmark over 8760 hours. Every strategy is evaluated on the same five weather years (seeds 42–46) and reported as mean $\pm$ sample standard deviation. An earlier version scored the deterministic controllers on one year and the learned ones on five, which was not like-for-like. Model-predictive figures use the realistic forecast of section 3.4. The final column is the fraction of commanded hydrogen energy the plant refused because the tank state could not supply it, $(\sum |P_{\text{cmd}}| - \sum |P_{\text{act}}|) / \sum |P_{\text{cmd}}|$ over the electrolyser and fuel-cell branch. Frequency is the mean absolute deviation; the maximum saturates at the 1 Hz clamp in fifteen of twenty learned runs and is not comparable. Dominance is shown in figure 1.

Strategy	n	Load deficit (%)	$ \Delta f $ mean (Hz)	SoH_{H_2} (final)	Curtailment (%)	H_2 refused (%)
Model-predictive, deficit	5	20.08 $\pm$ 0.51	0.126 $\pm$ 0.003	0.001 $\pm$ 0.002	0.05 $\pm$ 0.01	0.0
Pro-rata baseline	5	21.35 $\pm$ 0.49	0.138 $\pm$ 0.003	0.000 $\pm$ 0.000	0.28 $\pm$ 0.06	44.9
Model-predictive, reserve	5	21.80 $\pm$ 0.53	0.139 $\pm$ 0.003	0.935 $\pm$ 0.001	0.06 $\pm$ 0.02	0.2
Rule-based EMS	5	22.36 $\pm$ 0.46	0.154 $\pm$ 0.003	0.104 $\pm$ 0.003	1.82 $\pm$ 0.10	0.0
TD3	5	25.88 $\pm$ 0.85	0.177 $\pm$ 0.008	0.194 $\pm$ 0.025	1.98 $\pm$ 0.40	0.0
Soft actor–critic	5	26.16 $\pm$ 1.07	0.180 $\pm$ 0.009	0.189 $\pm$ 0.018	2.18 $\pm$ 0.30	0.0
DDPG	5	26.31 $\pm$ 0.85	0.180 $\pm$ 0.006	0.164 $\pm$ 0.002	1.96 $\pm$ 0.31	0.0
PPO	5	27.34 $\pm$ 1.53	0.192 $\pm$ 0.014	0.183 $\pm$ 0.009	2.89 $\pm$ 0.76	0.0

Eight strategies span 20.1 to 27.3% deficit, and model-predictive control takes both ends of the useful range: MPC-D posts the lowest deficit at 20.08%, MPC-R the highest terminal reserve by an order of magnitude at 0.935. The deterministic rule sets follow at 21.35 and 22.36%, the four learned controllers at 25.88–27.34%. That 1.46-point spread among the learned controllers matches their own across-seed standard deviations, 0.85 to 1.53, so algorithm choice is not resolved here.

The reserve column runs the other way, and the contrast between the two model-predictive variants is the clearest result in the table. They are the same controller with the same information, differing only in whether the objective prices stored hydrogen; that one change moves terminal reserve from 0.001 to 0.935 and costs 1.7 points of deficit, and both execute very nearly every command they issue. No measure built on served load alone separates them in proportion to the states they leave behind.

Against the learned controllers the comparison is one-sided on these two objectives: MPC-R serves more load and holds more reserve than every one of them, with complete separation in all four pairwise tests ($U = 0$ and $U = 25$, $p = 0.008$; section 4.2), so the non-dominated set has two members on deficit and reserve alone. This is not fully symmetric — the optimisers see a 48-hour forecast the agents’ state (equation 5) carries none of — and while the gap does not narrow under naive next-day persistence (deficit 23.12%, reserve 0.956), that forecast is still information the agents cannot use, so this bounds the asymmetry rather than resolving it. Nor is the front the whole picture: section 4.6 finds both model-predictive variants are the least voltage-compliant controllers tested, so a three-objective reading including voltage does not leave them unconditionally non-dominated.

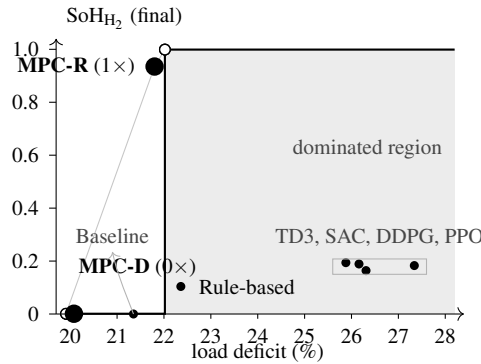


Figure 1. The eight strategies on the two competing objectives: annual load deficit (lower is better) against terminal hydrogen reserve (higher is better). The heavy staircase is the Pareto front and the shaded area beneath it is the dominated region. MPC-D and MPC-R name two points of an eight-point sweep of the reserve’s terminal price, 0 to $2\times$ its physical value (331.7 MWh per unit SoH_{H_2} ; the rest are drawn small). The response is a threshold: in every one of five seeds, reserve stays under 0.002 at or below $0.75\times$ and exceeds 0.998 by $1.25\times$, with the paper’s own price inside that transition at 0.935. Pushing past the knee is close to free: $1.25\times$ buys the last 0.064 of reserve for 0.22 points of deficit, and every price above it lands within 0.01 points and 0.001 of reserve of that same corner — so the choice is not a tuned dial. Every rule-based and learned strategy lies inside the dominated region, separated by less than 0.20 in reserve, a range the axis has to compress to show the sweep at all. The plot omits a third objective: section 4.6 finds both model-predictive variants are the least voltage-compliant strategies tested, so a three-objective reading including voltage does not leave the front unqualified. Points are means over the same five weather years.

The final column of table 4 is the first sign that the deficit column does not measure what it appears to. Of the hydrogen energy the pro-rata baseline commands over the year, 44.9% is refused because the tank cannot supply it: the dispatch it nominally follows is one the plant never executes. Every other controller commands hydrogen the plant can deliver.

That test is necessary but not sufficient, and MPC-D is the demonstration. It refuses nothing, issues no instruction the plant cannot meet, is optimal against its stated objective, and still ends the year with an empty tank. A feasibility audit catches a controller that promises energy it does not hold; only the

reserve state catches one that spends energy it will not replace. The two diagnostics are complementary, and neither is implied by annual deficit.

Voltage needs more care than a single compliance figure conveys. The IEEE 33-bus benchmark does not satisfy the $[0.95, 1.05]$ pu band at nominal load — its base-case minimum is 0.9159 pu at bus 18 — so full compliance is unattainable on this feeder by any dispatch strategy, and the corrected solve does not report it.

Frequency follows deficit closely: across the eight strategies the correlation between deficit and mean $|\Delta f|$ is $r = 0.99$, so the frequency column adds little beyond the deficit column. MPC-D achieves the lowest mean deviation, 0.126 Hz. Two caveats apply. Δf comes from a zero-order-hold solution of the swing equation at the hourly step, so it measures quasi-steady imbalance rather than sub-second dynamics and is not a grid-code assessment; and only the mean is reported, because the maximum saturates at the model’s 1 Hz clamp in fifteen of the twenty learned runs. Where a compliance figure is wanted, the fraction of hours inside the 0.5 Hz band is uncensored.

4.2 Seed repetition and statistical testing

All four learning algorithms were retrained from scratch on five seeds, 42 to 46, with the environment, reward and hyperparameters fixed. The seed also selects the load and renewable realisation, so each run draws its own weather year and the spread reported here bounds environment and training stochasticity together. Algorithms are compared primarily with exact Mann–Whitney U tests, since five observations per group cannot support a normality assumption; a parametric test is reported alongside in one case as a corroborating check, not as the basis for a conclusion.

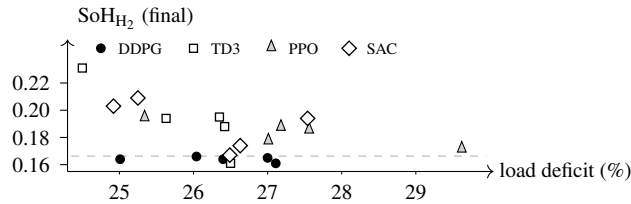


Figure 2. Every individual training run behind the $n = 5$ rows of table 4: four algorithms, five seeds each, on the two competing objectives. The four clouds overlap on both axes, which is why no deficit comparison and only two of twelve reserve comparisons separate. The exception is visible as a gap rather than a test statistic: every DDPG run falls below the dashed line at 0.1663 and every soft actor–critic run above it.

Four algorithms give six pairwise comparisons on each indicator. Of the twelve tests only two reach significance, and both are the same contrast: DDPG holds less terminal reserve than soft actor–critic ($U = 25$, two-sided $p = 0.008$) and less than PPO ($U = 0$, $p = 0.008$). No deficit comparison separates any pair. Neither survives a Bonferroni correction across the full family of twelve ($0.008 \times 12 \approx 0.10$), and none could: $p = 0.008$ is already the smallest two-sided value attainable at five observations per group, so no comparison in this section could reach correction-adjusted significance regardless of the underlying effect. That sharpens rather than weakens the reading above — algorithm choice is not resolved by this experiment at this sample size.

The DDPG–soft actor–critic contrast is the firmest because the samples separate completely: the largest DDPG value, 0.1663, lies below the smallest soft actor–critic value, 0.1667, giving the smallest one-sided p attainable at five runs per group, $1/252 = 0.004$, and a Welch t -test corroborates ($t = 3.04$, $p = 0.019$). That is a statement about ordering, not effect size: the difference in means is 0.025 and the margin between samples 0.0004, narrow enough that a sixth seed could close it.

What distinguishes DDPG is less that it holds the least reserve — the gap to PPO’s mean is 0.019 — than that it does so with almost no variation, a standard deviation of 0.002 against 0.009 for PPO, 0.018 for soft actor–critic and 0.025 for TD3. It converges reliably to a low-reserve policy where the others

reach 0.18 to 0.19 with varying success. TD3 records both the highest mean reserve and the widest spread: its best seed reaches 0.231, its worst falls to 0.161, below every DDPG run. Soft actor–critic is the only algorithm all of whose seeds clear all of DDPG’s.

Neither curtailment nor frequency separates soft actor–critic from DDPG: DDPG curtails less on the means, 1.96% against 2.18% ($U = 19$, $p = 0.92$), and frequency is indistinguishable at 0.180 Hz ($U = 15$, $p = 0.73$). An earlier version reported both as separating in soft actor–critic’s favour; neither survives. On deficit the samples are thoroughly interleaved ($U = 12$, $p = 0.50$), and no pairwise deficit comparison across the four approaches separates. Against the baseline’s own five-seed distribution ($21.35 \pm 0.49\%$, table 4), soft actor–critic is 4.82 points on the five-seed means worse with complete separation — every SAC seed lies above every baseline seed, the same exact test used throughout this section ($U = 25$, the maximum attainable at five per group, one-sided $p = 0.004$).

4.3 *Why the baseline’s low deficit is misleading*

Among the controllers that do not optimise a horizon, the pro-rata baseline reports the lowest annual deficit, 21.35%, against 22.36% for the rule-based controller and 25.88% for the best learned one — which at face value says neither the engineering effort nor the training bought anything. Three measurements say otherwise, none visible in the deficit column.

Nearly half of the baseline’s hydrogen dispatch is never executed. It allocates 30% of every deficit to the fuel cell irrespective of tank state. Over the year it commands 3037 MWh from the hydrogen branch, the plant delivers 1675 MWh, and the remaining 1362 MWh — 44.9% of the command — is refused because the tank cannot supply it. No other controller exceeds 0.2%. Annual deficit cannot distinguish a command that was met from one that was refused: it records only the load left unserved, never the instruction that failed.

Its advantage is partly an opening endowment. Every run starts the tank half full, an inventory of 165.9 MWh that the baseline spends and never replaces, finishing at $\text{SoH}_{\text{H}_2} = 0.000$. That makes 21.31% a first-year figure. Restarting each controller from the reserve it actually bequeaths, the second year is a fixed point for every one of them — terminal reserve equalling initial reserve to four decimal places — so these are steady states rather than a further transient. On the seed-42 year the baseline moves from 21.31 to 22.26% and ends at $\text{SoH}_{\text{H}_2} = 0.000$ having delivered 468.4 MWh of fuel-cell energy; the rule-based controller from 22.24 to 23.01% at 0.103 and 407.1 MWh; soft actor–critic from 26.18 to 26.70% at 0.189 and 889.0 MWh.

The baseline loses 0.95 percentage points, nearly twice the 0.52 that soft actor–critic loses, because it alone had been drawing on inventory it never rebuilt, and its margin over the rule-based controller narrows from 0.93 points to 0.75 without reversing. The claim is not that the baseline loses, but that the number by which it wins moves once the horizon extends past the endowment, with no indication in the scalar that it would.

An empty tank leaves the plant electrolyser-limited. In steady state the baseline can deliver only the hydrogen it manufactures within the same year, 468.4 MWh. Soft actor–critic, entering each year with roughly a fifth of a tank, delivers 889.0 MWh — close to twice the fuel-cell energy from identical hardware, indefinitely. A preserved reserve is not idle capital; it is what allows the stack to be used at all when it matters, which section 4.4 measures directly.

None of this makes the learned policy’s deficit the price of the reserve it keeps. Decomposing the 631 MWh by which soft actor–critic’s unserved load exceeds the baseline’s on the seed-42 year, the energy identity closes exactly on two terms, neither of them withheld hydrogen: 291 MWh of extra curtailment and 340 MWh of extra round-trip loss. It in fact delivers more fuel-cell energy than the baseline that year, 844 against 634 MWh, by running the electrolyser harder, which is where the extra loss arises. Its deficit gap is waste rather than restraint.

An earlier version of this study reported a single soft actor–critic run at 10.21% deficit and concluded

the learned policy undercut the baseline outright. Neither figure survives the correction of section 2.5, and the repetition does not support the conclusion in any case: no seed reproduces it and, on the seed-42 year both figures come from, the mean is 4.85 points the other way. That is this paper’s own argument applied to itself. A metric blind to command feasibility, biased against reserve preservation, and noisier than the margins it decides should not be the quantity on which a dispatch strategy is selected. Terminal reserve is the better one, and the refused-command fraction of table 4 is the cheapest check that a reported dispatch was physically realisable at all.

4.4 Does the hydrogen branch earn its place?

The argument so far is that the reserve matters, which needs a falsifiable test rather than an appeal to plausibility. A counterfactual was run in which the fuel-cell command is forced to zero after the controller has produced it, leaving the battery and supercapacitor allocation untouched and letting the fuel cell’s share go unserved. Zeroing at the output rather than inside the allocator is deliberate: it tests what happens if hydrogen were unavailable, not what a controller that never knew about hydrogen would do, and only the first is the correct comparison for duration-limited storage. The five worst non-overlapping five-day renewable windows were identified by rolling cumulative sum and both configurations run through each, starting from the same state reached by running the full controller from hour 0 without scoring.

Table 5. Multi-day counterfactual. “Extra” is the additional unserved energy when hydrogen is removed, in MWh and in percentage points of window load. The best-hydrogen column is the window with the healthiest starting reserve and is the meaningful comparator, because the worst-renewable window is often one where prior dispatch has already drawn the tank down.

Window length	Worst-renewable window		Best-hydrogen window		
	Extra (MWh)	Extra (pp)	Extra (MWh)	Extra (pp)	SoH _{H₂} start
3 days (72 h)	13.41	8.95	13.41	8.95	0.329
5 days (120 h)	10.20	4.92	20.85	8.38	0.347
7 days (168 h)	13.10	4.52	27.94	8.02	0.365

When the tank is healthy at the start of a low-renewable spell, removing hydrogen costs between 8.02 and 8.95 percentage points of additional unserved load, and that figure is stable across three-, five- and seven-day windows, a spread of 0.93 points. The stability is the point: a contribution appearing at only one window length would be an artefact of the window, whereas one holding across three is a structural property of the plant’s duration architecture. The battery and supercapacitor are exhausted within roughly 24 to 48 hours of sustained deficit, and past that point only the fuel cell can serve load. No resizing of the two short-duration devices within their rated power would change this, because the limitation is energy duration, not power.

The worst-renewable column varies much more, from 4.52 to 8.95 points, and the reason is instructive: the single worst window falls late in the year, after the reserve has already been drawn down, so hydrogen’s measured contribution there is understated by nearly a factor of two. Across the five candidate windows the dependence on starting reserve is sharper still. Where the tank holds between 0.24 and 0.35 at the window start, removing the fuel cell costs 18.8 to 20.9 MWh; where prior dispatch has drawn it to roughly 0.10, the same removal costs 2.0 to 2.8 MWh. The hydrogen branch is worth eight to nine percentage points when there is hydrogen in the tank and approximately nothing when there is not. That is the quantity section 4.3 shows the annual scalar cannot see, measured directly: the baseline’s terminal reserve of 0.000 is not a bookkeeping detail but the difference between a plant that can absorb a multi-day drought and one that cannot.

4.5 Seasonal behaviour

Annual aggregates hide when strategies diverge. Four representative days were selected by independent criteria — winter as the highest load-to-renewable ratio, spring as the median net-load day, summer as the highest combined throughput, autumn as the most volatile — and all three main strategies run on each, each warm-started from its own policy-consistent state so the reserve it carries into an arbitrary day is directly observable.

The SoH_{H_2} -start column dates the depletion. On the winter day, hour 384, all three controllers still hold reserve: 0.286 for the baseline, 0.315 for the rule-based controller, 0.175 for the learned policy. By the spring day at hour 3312 the baseline is at 0.000, and it stays there through summer and autumn at 0.003 and 0.001. The tank therefore empties by day 37, as figure 3 dates precisely, and never refills.

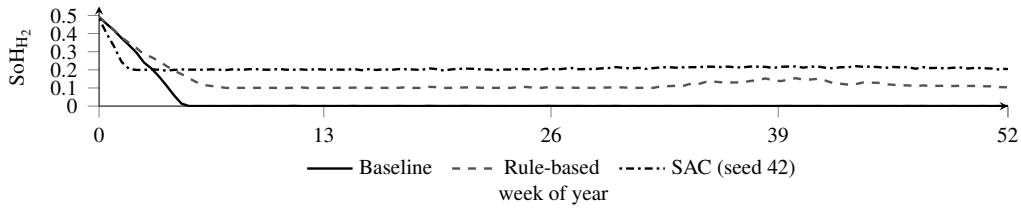


Figure 3. Terminal hydrogen reserve over the simulated year, daily means, seed 42. The baseline reaches $\text{SoH}_{\text{H}_2} \approx 0$ by day 37 and never recovers; the rule-based controller settles on its 0.10 floor by roughly day 80; soft actor-critic falls to about 0.20 by day 25 and then trends upward, the only one that rebuilds reserve. The four representative days examined in this section are independently warm-started and are not points on this trace.

The rule-based controller settles onto its 0.10 floor; the learned policy carries 0.163 to 0.167 into every non-winter day and is the only strategy that raises its reserve within the day in three of the four seasons.

This resolves the apparent paradox of section 4.3: on every non-winter day the baseline posts the lowest deficit — 29.80%, 5.41%, 3.86% — but from an already-empty tank, the accounting shadow of a depletion no single-day scalar attributes to its cause. Winter, sampled before depletion, is the exception that confirms it: the baseline wins by the largest margin in the study, 4.4 points against 0.93 on the seed-42 year (1.01 on the five-seed means of table 4), while draining reserve fastest (0.022 against 0.016 and 0.009). It leads while it still has reserve to spend, and spends it quickest.

Preservation costs show in two places: curtailment rises to 3.28 MWh in summer and 2.73 MWh in autumn as the learned policy spills surplus rather than force-charging; and winter frequency deviation is higher, 0.282 Hz against 0.225 Hz, since withholding devices makes for a lumpier dispatch.

Costing the same logs adds little. With every device’s throughput recomputed per seed over the five weather years, operating totals are \$268 686, \$314 665 and \$346 809, but lost load at 5 \$/kWh exceeds every operating term by two orders of magnitude, so the grand totals of \$18.97M, \$19.91M and \$23.27M restate the deficit column in dollars. A fifteen-year net present value leaves the ranking unchanged at any discount rate from 4 to 8%. A longer horizon does not help, because the omission that matters is structural: the costing gives a maintained reserve no value in any year.

4.6 A second distribution network

The dispatch results above are network-invariant by construction: the power balance is taken at the point of common coupling, $P_{\text{net}} = P_{\text{gen}} - P_L$, and never consults the network solve, while load is drawn independently of the feeder rating. Deficit, curtailment, reserve and frequency are therefore identical on any radial network, so a second network tests the voltage result alone. Reporting the same dispatch figures under two network headings would present one experiment as two.

Repeating the voltage solve on the IEEE 69-bus feeder, with the plant at bus 65 — its weakest bus, as bus 18 is on the 33-bus network — changes the result substantially. Injecting 1 MW raises the plant

bus by 49.4 mpu against 72.5 mpu on the 33-bus feeder, because the longer feeder is the weaker one. That lower sensitivity keeps the same dispatch inside the $[0.95, 1.05]$ pu band far more often (figure 4). The mean plant-bus voltage falls from about 1.06 pu, above the upper limit, to about 1.03 pu, inside it.

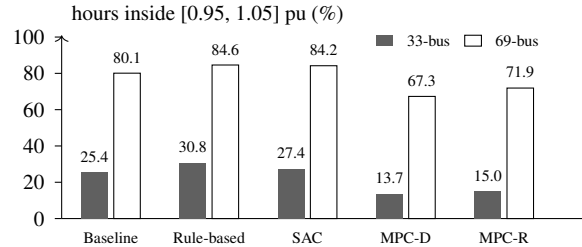


Figure 4. Voltage compliance of the same dispatch on both feeders. Every controller gains between 54 and 57 points on the 69-bus network, whose weaker plant bus turns 1 MW of injection into 49.4 mpu rather than 72.5 mpu. That the five gains fall in so narrow a band is the substance of the claim: the compliance figure of section 4.1 is a property of the 33-bus feeder, not of any strategy. The model-predictive controllers are worst on both — the cost of moving the most energy through a bus no controller observes, and the one indicator on which the learned policies do better.

5 Discussion

5.1 What determines success on this problem

The objective decides the outcome, not the algorithm class. MPC-D and MPC-R share a solver, a horizon, a forecast and a plant, and differ only in whether stored hydrogen carries a terminal value. That one difference is worth 0.93 of terminal reserve, while every distinction among the four learning algorithms is worth less than 0.03. A comparison that varies the algorithm while holding the objective fixed is measuring the smaller of the two effects.

Hydrogen depletion is path-dependent and one-directional over the operating year: refilling needs surplus, and surplus is absent exactly when the reserve is wanted. An optimiser with a horizon long enough to see the shortage coming, and an objective that prices the reserve, steps around the trap directly. An agent that has to discover it by exploration may not, and if it converges first it stays trapped.

Among the learning algorithms the residual pattern is value overestimation. The deterministic policy gradient accumulates critic bias with a direction here — discharging always reduces immediate deficit — so a critic that overvalues it drives the tank down irreversibly. TD3’s clipped double- Q targets, soft actor-critic’s entropy regularisation and PPO’s on-policy updates each avoid it: all three reach 0.18 to 0.19 against DDPG’s 0.164, and DDPG alone converges at a standard deviation of 0.002, failing the same way every time. An earlier version attributed preservation to entropy regularisation specifically; repeating TD3 and PPO withdraws that, since TD3 is not entropy regularised and holds the highest reserve.

5.2 Choosing the evaluation metric

The most consequential methodological choice here was refusing to rank strategies on annual load deficit, which places the naive baseline above a competent rule-based controller and would have produced the wrong recommendation. Reporting terminal reserve alongside deficit as a Pareto pair fixes it. The lesson generalises: whenever a system holds a resource that depletes irreversibly on a timescale longer than the reporting interval, aggregate metrics favour policies that spend it, and the fix is an explicit state variable in the results table rather than a better weighting.

A second failure generalises further. Annual deficit never asks whether the dispatch producing it was executable: 44.9% of the hydrogen the baseline commands is refused, and the metric reads the same either way. Accumulating commanded against delivered energy per device costs nothing and exposed a

defect that repeated reading of the deficit column had not. Feasibility alone is not sufficient either, and MPC-D is the counterexample: it issues no infeasible command, is optimal against its objective, and still ends the year empty. The two checks ask different questions — whether a controller promised energy it did not hold, and whether it spent energy it would not replace — and long-duration storage needs both.

5.3 Limitations

Nine limitations bound these results. The frequency model is quasi-steady at hourly resolution (section 2.3), so the supercapacitor’s sub-second capability appears only as ramp and peak-power headroom. Five seeds support an exact rank test but not a precise variance estimate, and section 4.2 finds most differences between learned controllers smaller than the seed spread itself. Each run regenerates its weather year from its seed, so that spread bounds weather variation and training stochasticity together and cannot separate them. The learned controllers are dominated on both objectives by model-predictive control, but the optimisers receive a 48-hour forecast the agents never see, and how much of the gap is horizon and how much is method is not separated here. Load and renewable profiles are synthetic, and because the power balance is taken at the point of common coupling the dispatch results are network-invariant: section 4.6 shows a second feeder changes the voltage result alone. The grid-side converter is modelled at unity power factor with no tap changer, regulator or switched capacitor, so the reported overvoltage is an upper bound a volt-var law would materially reduce. The supercapacitor is a second medium-duration branch rather than a short-duration one (section 1), so the three branches are separated by power rating and efficiency rather than by duration. The battery ages by cumulative depth-of-discharge counting rather than rainflow analysis, with no thermal model. Finally, the economic comparison prices only operating cost and lost load, assigning no value to a maintained reserve, so it cannot express the quantity section 4.4 measures.

Three practical points bear on deployment. Inference is negligible: 70 150 actor parameters and 103 μ s per deterministic action on one CPU thread, a duty cycle of 3×10^{-8} against the hourly interval. Telemetry is modest — three storage states, load and renewable power, frequency deviation and a clock — though no controller observes bus voltage, so the voltage results are an uncontrolled consequence of dispatch. The real barrier is that behaviour outside the training distribution is untested. A deployment would need a supervisory envelope around the policy, with the rule-based controller as the natural fallback. That changes the design question from which policy performs best to which one fails most safely.

6 Conclusion

A triple hybrid storage plant was operated on the IEEE 33-bus network for a full simulated year under eight dispatch strategies. Two of them, differing only in their objective, frame the result. A model-predictive controller minimising unserved load records the lowest deficit in the study, 20.08%, and ends the year with an empty tank. The same controller with a terminal value on stored hydrogen gives up 1.7 points of deficit and ends at 0.935. Neither issues a command the plant cannot execute, so the feasibility audit that exposes the pro-rata baseline — 44.9% of its hydrogen dispatch refused — does not separate them. Only the reserve state does.

The reserve-aware controller serves more load and holds more reserve than all four learned controllers on these two objectives, with complete separation on both indicators in every pairwise test. That comparison is not entirely fair: the optimisers see a 48-hour forecast the agents’ state never carries, and while the gap does not close as forecast quality degrades to next-day persistence, a weak forecast is still information the agents cannot use at all, so this bounds the asymmetry rather than resolving it. Nor is the reserve-aware controller unconditionally dominant: section 4.6 finds it is the least voltage-compliant strategy in the study, so the two-member non-dominated set holds on deficit and reserve, not on a three-objective reading that includes voltage. Among the learned controllers the experiment separates almost

nothing: of twelve pairwise tests only two reach significance, both distinguishing DDPG on terminal reserve.

The hydrogen branch earns its place quantitatively. Removing the fuel cell over the year's low-renewable windows costs 8.02 to 8.95 percentage points of additional unserved load when the tank is healthy, stable across three window lengths, and 2.0 to 2.8 MWh when prior dispatch has already emptied it.

Two methodological findings carry beyond this system. Aggregate outcome metrics systematically reward policies that spend an irreversibly depleting reserve, so any study of long-duration storage reporting only such aggregates risks recommending the wrong controller — and an optimal controller optimising that metric is the clearest demonstration of it rather than an exception. Second, an outcome metric cannot distinguish a dispatch the plant executed from one it refused, and accumulating commanded against delivered energy per device costs nothing to instrument. The two checks are complementary: one catches a controller promising energy it does not hold, the other one spending energy it will not replace.

Further work should give the learning agents the forecast the optimisers already receive, which is the likeliest explanation for the gap between them. It should also replace the synthetic profiles with a measured weather year. That single change would do most to extend these results, because it would also separate weather from training stochasticity, which the seed currently confounds (section 5.3). The curtailment gap of section 4.3 is the remaining loose end.

Statements and declarations

Data availability. Simulation code, trained checkpoints, training logs and the scripts generating every figure and table are available from the authors on request, together with an extended version of this paper that tabulates the per-seed deficit and terminal reserve behind figure 2 for all twenty runs. The released environment is the corrected version described in section 2.5; it carries the Baran and Wu bus loads totalling 3715 kW and 2300 kvar, and includes the self-test suite used to verify the corrections.

Funding. This research received no specific grant from any funding agency. **Conflict of interest.** The authors declare no competing financial or personal interests.

References

- [1] Muralee Gopi C V and Ramesh R 2024 *Results Eng.* **24** 103598
- [2] Beyers I *et al* 2023 *J. Energy Storage* **73** 109097
- [3] Kumar K, Kwon S and Bae S 2025 *J. Energy Storage* **108** 114936
- [4] Zhu W *et al* 2025 *J. Mar. Sci. Eng.* **13** 225
- [5] Haarnoja T, Zhou A, Abbeel P and Levine S 2018 *Proc. 35th Int. Conf. on Machine Learning* pp 1861–70
- [6] Haarnoja T *et al* 2018 Soft actor-critic algorithms and applications *arXiv* 1812.05905
- [7] Big-Alabo A 2026 *J. Energy Storage* **172** 123033
- [8] Feng J, Ren Z and Li W 2025 *Renew. Energy* **254** 123761
- [9] Zhou C *et al* 2026 *Energy* **344** 139876
- [10] Yi Z, Xu Y and Wu C 2024 *IEEE Trans. Power Syst.* **39** 3153–68
- [11] Mayne D Q, Rawlings J B, Rao C V and Sokaert P O M 2000 *Automatica* **36** 789–814
- [12] Schrock J, Chen Y and Ostrowski J 2026 *IEEE Access* **14** 106730–7
- [13] Huang Y, Gordon D and Scott P 2023 *Appl. Energy* **352** 121856
- [14] Sage M *et al* 2025 *J. Energy Storage* **115** 115428
- [15] Singh C *et al* 2026 *Int. J. Hydrogen Energy* **205** 153275
- [16] Chen P *et al* 2023 *IEEE Trans. Smart Grid* **14** 2656–67
- [17] Nejati Amiri M H *et al* 2026 *Energy Convers. Manage.* **353** 121215
- [18] Baran M E and Wu F F 1989 *IEEE Trans. Power Deliv.* **4** 1401–7
- [19] Dolatabadi S H, Ghorbanian M, Siano P and Hatziaargyriou N D 2021 *IEEE Trans. Power Syst.* **36** 2565–72
- [20] Peng J *et al* 2022 *eTransportation* **12** 100164
- [21] LeBel F A *et al* 2022 *J. Energy Storage* **54** 105255
- [22] Obeid H *et al* 2023 *IEEE Trans. Veh. Technol.* **72** 4448–60
- [23] Pattnaik M 2023 *Energy Storage* **5** e404
- [24] US Department of Energy Hydrogen Program 2025 *DOE Hydrogen and Fuel Cells Program Record* (Washington DC: DOE)
- [25] Wang Z, Zheng Y, Cheng M and Fan S 2022 *J. Power Sources* **536** 231245
- [26] National Institute of Standards and Technology 2022 *NIST Chemistry WebBook SRD 69*
- [27] Sdanghi G *et al* 2024 *Hydrogen* **5** 293–311
- [28] IEC 2021 *IEC 61215-1: Terrestrial Photovoltaic Modules* (Geneva: IEC)
- [29] IEC 2021 *IEC 61400-1: Wind Energy Generation Systems AMD1* (Geneva: IEC)
- [30] Li Q *et al* 2023 *Renew. Sustain. Energy Rev.* **187** 113723
- [31] Yegon L, Vasquez S and Guerrero J 2023 *IEEE Access* **11** 1782–97
- [32] Salem M *et al* 2023 *Sustainability* **15** 8
- [33] ENTSO-E 2016 *Network Code on Requirements for Grid Connection (RfG)* (Brussels: ENTSO-E) operative 2023
- [34] Lillicrap T P *et al* 2016 Continuous control with deep reinforcement learning *Proc. ICLR*
- [35] Fujimoto S, van Hoof H and Meger D 2018 *Proc. 35th Int. Conf. on Machine Learning* pp 1587–96
- [36] Schulman J *et al* 2017 Proximal policy optimization algorithms *arXiv* 1707.06347