Reinforcement Learning–Based Operational Optimization of a Hybrid Wind–CSP System with Thermal and Battery Energy Storage

Monzer Khalid*, Ahmad Al Hanbali*‡, Mohammad AlDurgam*‡, Ahmed M. Ghaithan†‡
*Department of Industrial and Systems Engineering
King Fahd University of Petroleum & Minerals, Saudi Arabia
†Department of Architecture Engineering and Construction Management
King Fahd University of Petroleum & Minerals, Saudi Arabia
‡Interdisciplinary Research Center for Smart Mobility and Logistics
King Fahd University of Petroleum & Minerals, Saudi Arabia

***Abstract*—Hybrid renewable power plants that combine multiple generation technologies with energy storage can improve renewable utilization and reliability, but their operation requires solving a high-dimensional, sequential dispatch problem under uncertainty. This paper presents a reinforcement learning (RL) framework for operational dispatch of a hybrid concentrated solar power (CSP)–wind system–Photo Voltaic (PV) solar system equipped with thermal energy storage (TES) and a battery energy storage system (BESS). We reformulate operational dispatch as a Markov decision process (MDP) with continuous control actions and a feasibility-projection layer that maps agent actions to physically admissible power flows. Proximal policy optimization (PPO) is adopted to learn a closed-loop dispatch policy that can react to variability in demand and renewable output. Using an 8064-hour held-out evaluation horizon, PPO with a nominal loss-of-load penalty improves the shaped return by 4.80% relative to a fixed-priority rule-based baseline, but increases loss of power supply probability (LPSP) from 3.60% to 11.17% due to near-elimination of BESS cycling. Increasing the loss-of-load penalty recovers the baseline reliability (LPSP 3.60%) and curtailment (42.25%), but causes PPO to collapse to the rule-based policy. These results highlight the sensitivity of RL dispatch to reward weights and motivate constrained/safe-RL formulations that enforce reliability targets while optimizing storage usage.**

***Index Terms*—reinforcement learning, hybrid renewable energy systems, energy storage, dispatch optimization, Markov decision process, Proximal Policy Optimization**

I. Introduction

The global transition toward low-carbon electricity is accelerating the deployment of renewable energy resources such as wind and solar. While these technologies reduce emissions and fuel dependence, their variability and partial predictability create operational challenges for balancing supply and demand. As renewable penetration increases, relying solely on dispatchable thermal generation becomes less effective and may increase operating costs.

Hybrid renewable energy systems (HRES) offer an important pathway to improve reliability and utilization by combining complementary resources with energy storage. In particular, CSP with thermal energy storage can provide dispatchable renewable power, while wind generation and battery storage support short-term flexibility. However, coordination of multiple resources requires solving a sequential decision problem: at each time step, the operator must decide how much renewable energy to serve directly to the load, how much surplus to store, and how to discharge storage to cover deficits. These decisions are temporally coupled through storage state-of-charge (SOC) dynamics and are constrained by charging/discharging limits, efficiencies, and reliability targets.

Classical approaches formulate dispatch using deterministic or stochastic optimization, including mixed integer linear programming (MILP) formulations for unit commitment and dispatch, stochastic programming, and robust optimization. Such methods provide optimal schedules under modeling assumptions but can be computationally demanding with the increase of time horizon length, resolution, and uncertainty modeling. Besides, model-based controllers need repeated re-optimization when conditions deviate from forecasts.

Reinforcement learning (RL) provides an alternative paradigm for sequential decision making under uncertainty by learning a control policy through interaction with an environment. Recent surveys and case studies report RL-driven controllers for microgrids, virtual power plants, and hybrid renewable systems, with PPO being a frequent choice for continuous control due to its stability and good performance [8]–[10]. Despite promising results, the practical value of RL for dispatch depends on careful formulation of constraints, reward design, and fair benchmarking against established optimization baselines.

This paper contributes a constrained RL formulation for dispatch of a hybrid CSP–PV–wind plant with both TES and BESS. The key idea is to preserve a realistic dispatch-priority hierarchy (PV/wind/CSP to load first; TES before BESS for deficits) for comparability with a deterministic MILP benchmark, while allowing RL to optimize continuous allocation decisions within that hierarchy. We formalize the problem as an MDP, introduce a feasibility projection that enforces operational constraints, and present a PPO-based learning methodology. We also define an evaluation protocol

and performance indicators for systematic comparison.

II. Related Work

Hybrid CSP–PV–wind plants with storage have been studied to increase dispatchability and reduce curtailment. Techno-economic analyses show that hybridization can improve utilization, while outcomes depend on storage costs and operating strategies [1]. Operational scheduling of hybrid plants is commonly formulated as an optimization problem, e.g., day-ahead MILP scheduling that jointly considers renewables, storage, and demand response [2]. More broadly, deterministic optimization has a long history in dispatch and optimal power flow [3], while multi-stage stochastic methods remain important for long-horizon planning and operations under uncertainty [4], [5]. Robust optimization approaches hedge against worst-case uncertainty but can be conservative and costly [6], [7].

RL has emerged as a data-driven alternative for energy management and dispatch. Review articles describe RL for battery scheduling, microgrid control, and grid services, highlighting deep RL algorithms such as DDPG/TD3, SAC, and PPO [8], [10]. PPO is widely used in energy applications due to stable policy updates via a clipped objective and support for continuous action spaces [10]. Recent works apply PPO variants for virtual power plant dispatch [11], multi-objective dispatch of hybrid systems [12], and economic dispatch benchmarks [13]. Safety and constraint satisfaction are active research areas, including safe RL methods incorporating risk measures in learning [14].

Motivated by these advances, this paper focuses on a constrained RL formulation for hybrid plant. A comparison with an MILP benchmark is conducted for the same physical system and dispatch-priority assumptions.

III. Hybrid System and Deterministic Benchmark

A. System components

We consider a hybrid renewable system supplying an electrical demand at hourly resolution. The system includes:

- A parabolic-trough CSP plant with solar field, power block, and thermal energy storage (TES).
- A wind farm composed of identical wind turbines.
- A PV solar system.
- A battery energy storage system (BESS).
- A residential (or aggregated) electrical load profile.

The hybrid plant is designed and operated over a representative year, with decisions made at each hour.

B. Modeling of energy resources

The wind farm output depends on hub-height wind speed. Denoting the power of one turbine by P_t^{WT} which is rated as 50 KW and the number of turbines by N_{WT}, the available wind power is

$$P_t^{\mathrm{W}} = N_{\mathrm{WT}} P_t^{\mathrm{WT}}. \tag{1}$$

CSP solar-field thermal power depends on direct normal irradiance (DNI) and the aperture area A_{SF}, which is a design variable. Accounting for efficiencies, the available CSP electrical power is computed from DNI-driven solar-field output.

C. Benchmark MILP formulation

A deterministic MILP benchmark minimizes life-cycle cost (LCC) while meeting demand and a reliability constraint on demand power satisfaction. Following the hybrid CSP–wind sizing literature [15], design variables include the solar-field aperture area A_{SF} and the number of wind turbines N_{WT}. Operational variables include hourly energy flows from wind and CSP to the load, TES and BESS charge/discharge energies, curtailment, and unserved energy.

The objective function minimizes

$$\min \ \mathrm{LCC} = C_{\mathrm{inv}} + C_{\mathrm{O\&M}}, \tag{2}$$

where C_{inv} is the sum of component capital costs and $C_{\mathrm{O\&M}}$ aggregates operation and maintenance costs.

Operational constraints include:

- **Hourly energy balance:** served energy equals demand minus unserved energy, and is supplied by PV, wind, CSP, TES discharge, and BESS discharge.
- **Generation limits:** PV, wind and CSP outputs are bounded by available power.
- **Storage dynamics:** inventories evolve with charge/discharge decisions and efficiencies, subject to SOC bounds and charge/discharge power limits.
- **Reliability:** loss of power supply probability (LPSP) is constrained by an upper bound as follows

$$\mathrm{LPSP} = \frac{\sum_t E_t^{\mathrm{unserved}}}{\sum_t D_t} \leq \mathrm{LPSP}_{\max}. \tag{3}$$

A key feature of the benchmark is a *fixed dispatch-priority policy*: PV, wind and CSP serve the load first; surplus is used to charge storage (subject to capacity limits) and any remaining surplus is curtailed; when there is a deficit, TES is dispatched before BESS. This policy is encoded using additional constraints and binary variables to prevent simultaneous charging and discharging and to implement the priority order [15]. While convenient and tractable, this rule-based dispatch limits adaptivity to changing operating conditions.

IV. MDP Formulation for Operational Dispatch

To capture sequential decision making under variability and uncertainty, we define an MDP $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma \rangle$ with hourly step $\Delta t = 1$ h. Exogenous information are related to weather conditions at time time t such as DNI_t, wind speed at hub height, v_t^{hub}, ambient temperature, T_t^{amb}, and demand load at time t, D_t.

A. State space

The state $s_t \in \mathcal{S}$ at time t collects observable exogenous variables, time features, storage SOC, and the previous action a_{t-1}:

$$\begin{aligned} s_t = \big(&\mathrm{POA}_t, \mathrm{DNI}_t, T_t^{\mathrm{amb}}, V_t^{\mathrm{hub}}, D_t, \\ &\phi_t^{\mathrm{hr}}, \phi_t^{\mathrm{mo}}, \mathrm{SOC}_{\mathrm{ThB},t}^{A}, \mathrm{SOC}_{\mathrm{B},t}^{A}, a_{t-1}\big). \end{aligned} \tag{4}$$

Here ϕ_t^{hr} and ϕ_t^{mo} encode hour-of-day and month-of-year to account for seasonality; the state of charge of thermal energy

SOC_t^{Th} and battery storage SOC_t^B are normalized by their capacities. Including a_{t-1} helps discourage oscillatory control actions.

B. Action space

To enable a fair comparison with the benchmark while keeping decisions physically meaningful, we preserve the same dispatch-priority structure of the benchmark in [15] and let RL optimize the power allocation *within* it. The action is a continuous vector $a_t \in [0,1]^3$:

$$a_t = (u_t^{\mathrm{RE}}, u_t^{\mathrm{CSP}}, u_t^{\mathrm{def}}), \tag{5}$$

where u_t^{RE} is the fraction of PV+wind surplus sent to BESS charging, u_t^{CSP} is the fraction of CSP surplus sent to TES charging, and u_t^{def} is the fraction of the remaining deficit requested from TES (with BESS covering the residual).

C. Feasibility projection and transition dynamics

The feasibility-projection layer Π acts as a hard-coded physical filter between the RL policy and the environment. It guarantees admissibility by: (1) enforcing a strict power-balance equality (Eq. 6–9), (2) clipping requested storage actions against instantaneous power limits (P_{dis}^{max}) and available energy headroom (H_{TES}, H_{BESS}), and (3) preventing simultaneous charging and discharging via the sequential priority logic. Consequently, even if the neural network outputs an physically impossible action, the projection maps it to the nearest feasible power flow, ensuring the system never violates SOC or power-rating constraints. Applying directly a_t as power allocation can violate constraints. Therefore, we introduce a deterministic feasibility projection Π that maps (s_t, a_t) to admissible power flows that satisfy power balance, priority order, storage limits, and efficiency dynamics.

Let P_t^{RE} and P_t^{CSP} denote available PV+wind and CSP electrical power at time t. PV+Wind and CSP are dispatched directly to the load first:

$$P_t^{\mathrm{RE}\to\mathrm{load}} = \min\left(D_t, P_t^{\mathrm{RE}}\right), \tag{6}$$

$$D_t^{(1)} = D_t - P_t^{\mathrm{RE}\to\mathrm{load}}, \tag{7}$$

$$P_t^{\mathrm{CSP}\to\mathrm{load}} = \min\left(D_t^{(1)}, P_t^{\mathrm{CSP}}\right). \tag{8}$$

The post-renewables deficit is

$$\mathrm{Def}_t = D_t - \left(P_t^{\mathrm{RE}\to\mathrm{load}} + P_t^{\mathrm{CSP}\to\mathrm{load}}\right) \geq 0. \tag{9}$$

TES is dispatched before BESS to cover deficits, subject to discharge power limits and available energy:

$$P_t^{\mathrm{TES}\to\mathrm{load}} = \min\left(u_t^{\mathrm{def}}\,\mathrm{Def}_t,\, P_{\max}^{\mathrm{TES,dis}},\, \eta_{\mathrm{dis}}^{\mathrm{TES}} E_t^{\mathrm{TES}}\right), \tag{10}$$

$$\mathrm{Def}_t' = \mathrm{Def}_t - P_t^{\mathrm{TES}\to\mathrm{load}}, \tag{11}$$

$$P_t^{\mathrm{BESS}\to\mathrm{load}} = \min\left(\mathrm{Def}_t',\, P_{\max}^{\mathrm{BESS,dis}},\, \eta_{\mathrm{dis}}^{\mathrm{BESS}} E_t^{\mathrm{BESS}}\right), \tag{12}$$

where E_t^{TES} and E_t^{BESS} denote stored energy (or equivalently SOC-scaled energy) at time t.

TABLE I
MDP ELEMENTS FOR HYBRID CSP–WIND DISPATCH.

Element	Description
State s_t	Weather $(\mathrm{POA}_t, \mathrm{DNI}_t, T_t^{\mathrm{amb}}, v_t^{\mathrm{hub}})$, demand D_t, time features, storage SOCs, previous action a_{t-1}.
Action a_t	Continuous fractions $(u_t^{\mathrm{W}}, u_t^{\mathrm{CSP}}, u_t^{\mathrm{def}}) \in [0,1]^3$ controlling charging from surplus and TES share of deficit.
Transition $\mathcal{P}$	Deterministic projection enforcing power balance, priority order, storage limits, and efficiencies.
Reward r_t	Negative weighted sum of O&M cost, curtailment, BESS throughput, unserved energy, and action variation.

Surplus power of PV+wind and CSP after serving the load are:

$$P_t^{\mathrm{RE,sur}} = \max\left(P_t^{\mathrm{RE}} - P_t^{\mathrm{RE}\to\mathrm{load}}, 0\right), \tag{13}$$

$$P_t^{\mathrm{CSP,sur}} = \max\left(P_t^{\mathrm{CSP}} - P_t^{\mathrm{CSP}\to\mathrm{load}}, 0\right). \tag{14}$$

Charging is constrained by headroom and power limits. Let the available headroom of BESS and TES be

$$H_t^{\mathrm{BESS}} = \min\left(P_{\max}^{\mathrm{BESS,ch}},\, E_{\max}^{\mathrm{BESS}} - E_t^{\mathrm{BESS}}\right), \tag{15}$$

$$H_t^{\mathrm{TES}} = \min\left(P_{\max}^{\mathrm{TES,ch}},\, E_{\max}^{\mathrm{TES}} - E_t^{\mathrm{TES}}\right), \tag{16}$$

then the charging flows are

$$P_t^{\mathrm{RE}\to\mathrm{BESS}} = \min\left(u_t^{\mathrm{W}} P_t^{\mathrm{RE,sur}},\, H_t^{\mathrm{BESS}}\right), \tag{17}$$

$$P_t^{\mathrm{CSP}\to\mathrm{TES}} = \min\left(u_t^{\mathrm{CSP}} P_t^{\mathrm{CSP,sur}},\, H_t^{\mathrm{TES}}\right). \tag{18}$$

Any remaining surplus is curtailed. Storage states are updated using charge/discharge efficiencies:

$$E_{t+1}^{\mathrm{TES}} = E_t^{\mathrm{TES}} + \eta_{\mathrm{ch}}^{\mathrm{TES}} P_t^{\mathrm{CSP}\to\mathrm{TES}} \Delta t - \frac{P_t^{\mathrm{TES}\to\mathrm{load}}}{\eta_{\mathrm{dis}}^{\mathrm{TES}}} \Delta t, \tag{19}$$

$$E_{t+1}^{\mathrm{BESS}} = E_t^{\mathrm{BESS}} + \eta_{\mathrm{ch}}^{\mathrm{BESS}} P_t^{\mathrm{RE}\to\mathrm{BESS}} \Delta t - \frac{P_t^{\mathrm{BESS}\to\mathrm{load}}}{\eta_{\mathrm{dis}}^{\mathrm{BESS}}} \Delta t. \tag{20}$$

Served and unserved energies follow from the total supplied power. Table I summarizes the MDP elements.

D. Reward design

The reward function was designed using a weighted-sum scalarization approach commonly adopted in RL-based energy management studies. The coefficient selection was guided by the relative operational importance of reliability, renewable utilization, and storage degradation. In particular, λ_{LOL} was treated as the primary tuning parameter because it directly governs the reliability-efficiency trade-off The reward at time t penalizes operational costs and undesirable outcomes:

$$\begin{aligned} r_t = -(&C_t^{\mathrm{O\&M}} + C_t^{\mathrm{cyc}} E_t^{\mathrm{BESS,thru}} + c_{\mathrm{curt}} E_t^{\mathrm{curt}} \\ &+ \lambda_{\mathrm{LOL}} E_t^{\mathrm{unserved}} + C_{\mathrm{sw}} |a_t - a_{t-1}|) \end{aligned} \tag{21}$$

Here $E_t^{\mathrm{BESS,thru}}$ is BESS energy throughput (charge plus discharge) as a proxy for cycling/degradation, and λ_{LOL} is the loss-of-load penalty weight.

The cost factors were fixed at $C_t^{\mathrm{O\&M}} = 0$, $c_{\mathrm{cyc}} = 2.5$, $c_{\mathrm{curt}} = 2 \times 10^{-4}$, and $c_{\mathrm{sw}} = 10^{-2}$. For numerical stability the reward was additionally scaled by `reward_scale` $= 10^{-3}$. We then sweep $\lambda_{\mathrm{LOL}} \in \{5, 10, \ldots, 50\}$ to quantify how reward shaping changes the learned dispatch.

V. PPO-Based Learning Methodology

We learn a stochastic policy $\pi_\theta(a_t|s_t)$ using PPO. The actor network outputs the mean and diagonal variance of a Gaussian distribution over actions; samples are mapped into $[0, 1]^3$ and then passed to the feasibility projection layer. The critic estimates the state value $V_\psi(s_t)$. PPO updates maximize a clipped surrogate objective to improve training stability and avoid large destructive policy steps [10].

The learning environment implements the same physical models used by the benchmark study (renewable availability, storage dynamics, and constraints), but produces closed-loop trajectories in response to actions. Training can incorporate:

- **Reward normalization and gradient clipping** to stabilize optimization.
- **Domain randomization** over weather and demand profiles to improve generalization.
- **Warm-start policies** consistent with the priority hierarchy to accelerate learning.

After training, the learned policy can be deployed for fast online decision making, since inference avoids solving an optimization problem at each hour.

A. Training configuration and reproducibility

We implemented PPO using the Stable-Baselines3 (SB3) library with a Gymnasium environment wrapper and a feasibility-projection layer implemented inside the environment step function. Table II lists the PPO hyperparameters and network architecture used in the experiments. The actor and critic neural networks are multilayer perceptrons (MLPs) with two hidden layers of 128 units each (separate π and V heads). The SB3's default tanh activation is used. Parameter updates are performed with the Adam optimizer. A single environment instance is wrapped by DummyVecEnv; the random seed is set to 42 for reproducibility.

The rollout length per PPO update is $n_{\mathrm{steps}} = 2688$ (approximately two 1344-hour episodes), with 16 mini-batches (batch size 168) and 8 epochs per update. Training is run for 500 updates, totaling 1.344×10^6 environment interactions. During training, evaluation is performed every 5 PPO updates on three fixed evaluation seeds (123, 456, 789) using deterministic action selection.

TABLE II
PPO TRAINING AND NETWORK HYPERPARAMETERS (SB3 IMPLEMENTATION).

Parameter	Value
Policy	`MlpPolicy` (Gaussian actor–critic)
Actor/Critic MLP	2 layers $\times$ 128 units (pi/vf)
Activation	`tanh` (SB3 default)
Optimizer	Adam, learning rate 3×10^{-4}
Discount factor γ	0.997
GAE parameter λ	0.95
Clip range ϵ	0.2
Entropy coefficient	10^{-3}
Value-function coefficient	0.5
Max grad norm	1.0
Rollout length n_{steps}	2688
Mini-batches / batch size	16 / 168
Epochs per update	8
Total updates / timesteps	500 / 1,344,000

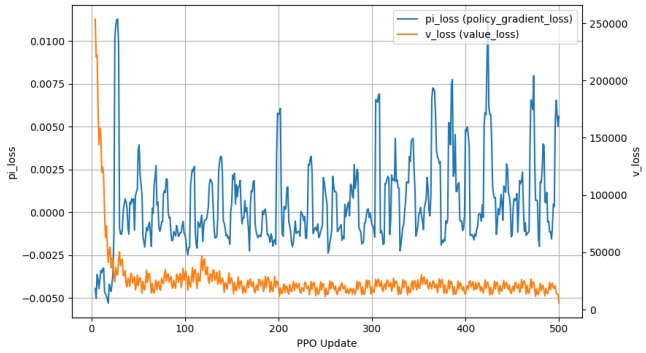

Fig. 1. PPO learning convergence diagnostics: policy-gradient loss (π-loss, left axis) and critic value loss (V-loss, right axis) over training updates. The dual-axis presentation improves readability because the two losses differ by several orders of magnitude.

VI. Evaluation Protocol and Performance Metrics

To compare RL and optimization in a controlled manner, we propose:

1) **Benchmark sizing and parameters:** solve the deterministic MILP to determine design parameters (e.g., A_{SF}, N_{WT}, storage capacities) and to define the dispatch hierarchy.
2) **RL training with matched constraints:** train PPO in simulation using the same physical models and operational constraints, and preserve the same priority order.
3) **Out-of-sample testing:** evaluate both approaches on identical held-out weather/load profiles and perturbations (e.g., forecast errors) to assess robustness.

We report metrics that capture reliability, sustainability, and economic performance:

- **Reliability:** LPSP, total unserved energy, and distribution of loss-of-load events.
- **Renewable utilization:** total curtailment and served renewable fraction.
- **Operational cost proxies:** O&M cost and storage cycling penalties (or full operating cost if available).

A. Implementation and experiment setup

The system is modeled for a site in Dhahran, Saudi Arabia, using meteorological data (POA, DNI, wind speed, temperature) at 1-hour resolution. The sizing of the system as follow: PV size =1 MWp, $A_{SF} = 15,000$ m^2, $N_{WT} = 20$ turbines, TES capacity of 5 MWh, and BESS capacity of 2 MWh.

We implemented the environment and PPO agent using the Stable-Baselines3 (SB3) library. The simulator operates

at hourly resolution and enforces the feasibility projection described in Section IV. Weather inputs consist of hourly POA, DNI and wind-speed profiles obtained from the PVGIS database for the considered location, while the demand trajectory corresponds to an aggregated residential load profile. Training and evaluation were performed using non-overlapping temporal windows to reduce dependence on a specific seasonal realization. The 8064-hour evaluation horizon corresponds to six independent 1344-hour test windows aggregated for reporting. The rule-based benchmark represents a fixed-priority "greedy" dispatch logic. Setting $a_t = (1, 1, 1)$ for all t implies: (i) 100% of PV+wind surplus is directed to BESS charging, (ii) 100% of CSP surplus is directed to TES charging, and (iii) the system always requests the maximum available energy from TES to cover deficits before engaging the BESS.

To study reward sensitivity, we compare two reward configurations that differ only in the loss-of-load penalty λ_{LOL}: a *nominal* setting (moderate reliability emphasis) and a *high* setting (strong reliability emphasis). Since the episode return aggregates weighted penalty terms, its magnitude is not directly comparable across different λ_{LOL} settings; therefore, we interpret performance primarily through physical metrics such as LPSP, curtailment, and storage throughput. The reward coefficients were selected to reflect operational trade-offs commonly considered in hybrid renewable systems. The cycling coefficient c_{cyc} acts as a proxy for battery degradation cost, while the curtailment coefficient c_{curt} encourages renewable utilization without dominating the reward. The switching penalty c_{sw} discourages oscillatory control actions that may be operationally undesirable. Since reliability is the primary operational objective, sensitivity analysis was performed over $\lambda_{LOL} \in \{5, 10, \ldots, 50\}$ to evaluate how reliability weighting influences the learned dispatch behavior. The objective of this study is not to optimize reward coefficients exhaustively, but rather to investigate the sensitivity of RL dispatch performance to reliability penalization.

VII. Numerical Results and Discussion

Table III reports representative operating outcomes for three settings of the loss-of-load penalty λ_{LOL} while keeping the remaining coefficients in eq. (21) fixed. Note, the episode return is a *shaped* objective, therefore, changing λ_{LOL} rescales the reward. The rewards for different λ_{LOL} values are not directly comparable. We therefore analyze the impact on key performance indicators (KPIs)—LPSP, curtailment share (Curt.), and storage throughput. Fig. 2 summarizes the impact of the full range of$\lambda_{\mathrm{LOL}} \in \{5, 10, \ldots, 50\}$.

At low penalty ($\lambda_{\mathrm{LOL}} = 5$), PPO slightly improves the shaped return relative to the rule-based baseline (from -1889.49 to -1797.93), but it does so by sacrificing reliability: LPSP increases from 3.597% to 11.166% and curtailment rises from 42.25% to 47.21% of raw generation. The drastic reduction in BESS throughput (525.8 MWh $\rightarrow$ 5.7 MWh) indicates that the agent largely avoids battery cycling and instead accepts energy-not-served when it is weakly penalized. This behavior is consistent with eq (21): when the loss-of-load term is underweighted relative to the cycling penalty, the optimal policy under the scalarized objective can prefer "do-nothing" storage actions even if they increase outages.

Increasing λ_{LOL} systematically shifts this trade-off. As shown in Fig. 2, PPO's LPSP drops sharply from $\sim$11% toward the rule-based level ($\approx$3.6%) as λ_{LOL} grows, with curtailment simultaneously decreasing toward the baseline. In the mid-range (e.g., $\lambda_{\mathrm{LOL}} = 25$ in Table III), PPO re-engages storage (throughput $\approx$508.6 MWh), nearly matches baseline curtailment, and closes most of the reliability gap. At high penalty ($\lambda_{\mathrm{LOL}} = 50$), PPO matches the rule-based policy's physical KPIs exactly, indicating that strong reliability shaping drives the learned policy to behave conservatively and leaves little room for improvement over the hand-crafted baseline under the chosen state/action representation. Practically, the curve suggests diminishing returns beyond roughly $\lambda_{\mathrm{LOL}} \approx 40$ for this system: further penalization changes the shaped return but does not materially improve LPSP or curtailment.

Learning stability is evaluated via training diagnostics. Fig. 1 shows the critic loss rapidly decreasing from a high initial value and then remaining bounded, while the policy-gradient loss oscillates around a small magnitude. These trends indicate a stable training regime in which PPO updates do not diverge. Importantly, combining Fig. 1 and Fig. 2 highlights a practical insight: stable convergence does not guarantee acceptable physical performance—reward weights must be calibrated to reflect the desired reliability–efficiency trade-off, and sensitivity analysis (as done here) is a lightweight way to select penalty ranges before moving to more principled constrained or multi-objective formulations.

A. Generalization and robustness considerations

The PPO policy was trained and evaluated on multiple non-overlapping temporal windows with varying demand and renewable conditions. Since the state includes weather, demand, and storage SOC variables, the learned policy adapts its dispatch decisions to changing operating conditions.

Although this study considers a fixed system sizing , the proposed MDP and feasibility-projection framework can be retrained for other plant scales and storage capacities without modifying the RL architecture.

The 8064-hour evaluation horizon provides an initial out-of-sample assessment across multiple seasonal windows; however, testing under forecast uncertainty and multi-year datasets remains future work.

B. Insights and implications

The observed behavior provides practical guidance for designing RL-based dispatch controllers. *First*, a single weighted-sum reward induces a Pareto trade-off between reliability, curtailment, and storage degradation proxies. If reliability is a hard requirement (as in the MILP benchmark through an LPSP constraint), then a penalty-only reward may be insufficient: moderate λ_{LOL} can produce policies with attractive return but unacceptable reliability. *Second*, increasing λ_{LOL}

TABLE III
REPRESENTATIVE AGGREGATED EVALUATION RESULTS OVER AN 8064-HOUR HELD-OUT HORIZON FOR THREE LOSS-OF-LOAD PENALTY SETTINGS.

Controller	λ_{LOL}	Episode return	Demand (MWh)	Unserved (MWh)	LPSP (%)	Curt. (% of raw gen)	BESS throughput (MWh)
Rule-based	5	-1889.49	3193.9	114.9	3.597	42.25	525.8
PPO (SB3)	5	-1797.93	3193.9	356.6	11.166	47.21	5.7
Rule-based	25	-4187.10	3193.9	114.9	3.597	42.25	525.8
PPO (SB3)	25	-4349.23	3193.9	123.1	3.854	42.41	508.6
Rule-based	50	-7059.12	3193.9	114.9	3.597	42.25	525.8
PPO (SB3)	50	-7059.12	3193.9	114.9	3.597	42.25	525.8

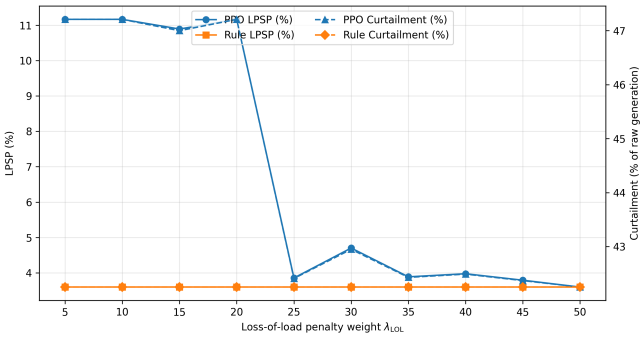

Fig. 2. Sensitivity of operating reliability and curtailment to the loss-of-load penalty λ_{LOL} in the reward.

can enforce reliability, but may remove the incentive for PPO to discover strategies beyond the fixed-priority baseline. In our setting, the rule-based action $a_t = (1, 1, 1)$ is already near-optimal for minimizing unserved energy under the imposed dispatch hierarchy and feasibility projection, so PPO naturally converges to it.

These findings motivate two extensions: (i) *constrained or safe RL* (e.g., Lagrangian methods or risk-sensitive objectives) that enforces an LPSP target while optimizing secondary metrics such as curtailment and cycling; and (ii) *richer action parameterizations* that allow RL to improve beyond a fixed hierarchy (e.g., additional degrees of freedom for allocating surplus across both storage devices or for shaping discharge aggressiveness).

VIII. CONCLUSION

We presented a constrained RL framework for dispatch of a hybrid PV–wind–CSP power plant with both thermal and battery storage. The approach reformulates dispatch as an MDP and learns a PPO policy whose actions are mapped to feasible power flows through a projection layer enforcing operational constraints and a fixed dispatch-priority hierarchy. Using an 8064-hour held-out evaluation horizon, we showed that PPO can learn markedly different storage-usage patterns, but performance is highly sensitive to reward weights: with a nominal loss-of-load penalty, PPO reduces BESS cycling and slightly improves shaped return, but yields a higher LPSP than the rule-based baseline; with a high loss-of-load penalty, PPO recovers baseline reliability but collapses to the rule-based behavior. Future work will focus on constrained/safe-RL formulations that explicitly enforce reliability targets while optimizing curtailment and degradation proxies, and on extending the action space to allow improvements beyond fixed dispatch hierarchies.

REFERENCES

[1] A. Zurita, C. Mata-Torres, C. Valenzuela, C. Felbol, J. M. Cardemil, A. M. Guzmán, and R. A. Escobar, "Techno-economic evaluation of a hybrid CSP+PV plant integrated with thermal energy storage and a large-scale battery energy storage system for base generation," *Solar Energy*, vol. 173, pp. 1262–1277, 2018.

[2] Z. Zhao, Y. Zhang, Y. Wang *et al.*, "Optimal scheduling of the CSP–PV–wind hybrid power generation system considering demand response," *Frontiers in Energy Research*, vol. 12, p. 1357406, 2024.

[3] S. Frank, I. Steponavice, and S. Rebennack, "Optimal power flow: A bibliographic survey. Part I: Formulations and deterministic methods," *Energy Systems*, vol. 3, no. 3, pp. 221–258, 2012.

[4] C. Füllner and S. Rebennack, "Stochastic dual dynamic programming and its variants: A review," *SIAM Review*, vol. 67, no. 3, pp. 415–539, 2025.

[5] C. M. McDonald and C. T. Maravelias, "Stochastic programming models for long-term energy transition planning," *Systems Control Transactions*, vol. 3, pp. 519–526, 2024.

[6] Y. Wang and J. Li, "Optimizing multi-energy systems with enhanced robust planning for cost-effective and reliable operation," *International Journal of Electrical Power and Energy Systems*, vol. 161, p. 110178, 2024.

[7] M. Wedemeyer, E. Cramer, A. Mitsos, and M. Dahmen, "Robust energy system design via semi-infinite programming," *arXiv preprint* arXiv:2411.14320, 2025.

[8] A. T. D. Perera and P. Kamalaruban, "Applications of reinforcement learning in energy systems," *Renewable and Sustainable Energy Reviews*, vol. 137, p. 110618, 2021.

[9] S. Stavrev and I. Ginchev, "Reinforcement learning techniques in optimizing energy systems," *Electronics*, vol. 13, no. 8, p. 1459, 2024.

[10] A. Hanif, G. Pillai, L. Dong, H. Punit, and M. Azmi, "The solar AI nexus: Reinforcement learning shaping the future of energy management," *Wiley Interdisciplinary Reviews: Energy and Environment*, vol. 14, p. e70012, 2025.

[11] Z. Gao, W. Kang, X. Chen, S. Gong, Z. Liu, D. He, S. Shi, and X.-C. Shangguan, "Optimal economic dispatch of a virtual power plant based on gated recurrent unit proximal policy optimization," *Frontiers in Energy Research*, vol. 12, p. 1357406, 2024.

[12] C. Wang, Y. Ma, J. Xie, and Q. Ouyang, "Multi-objective energy dispatch with deep reinforcement learning for wind-solar-thermal-storage hybrid systems," *Journal of Energy Storage*, vol. 105, p. 114635, 2025.

[13] A. Rizki, A. Touil, A. Echchatbi, R. Oucheikh, and M. Ahlaqqach, "A reinforcement learning-based proximal policy optimization approach to solve the economic dispatch problem," *Engineering Proceedings*, vol. 97, no. 1, p. 24, 2025.

[14] J. Feng, Z. Ren, and W. Li, "A robust safe reinforcement learning-based operation method for hybrid electric-hydrogen energy system risk-based dispatch considering dynamic efficiency characteristics of electrolysers," *Renewable Energy*, vol. 254, p. 123761, 2025.

[15] A. M. Ghaithan, A. Al Hanbali, A. Mohammed, and M. Abdel-Aal, "A MILP optimization model for sizing a hybrid concentrated solar power-wind system considering energy allocation," *Process Integration and Optimization for Sustainability*, vol. 8, no. 5, pp. 1527–1544, 2024.