Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Risk-Aware Hybrid Reinforcement Learning and Robust Optimization for Reliable PV–Battery Scheduling Under Uncertainty

Bareq Kawwam Ahmed
University of Tabriz
Mehrdad Tarafdar Hagh
tarafdar@tabrizu.ac.ir

University of Tabriz
Seyyed Hossein Hosseini
University of Tabriz

Article

Keywords: Photovoltaic–battery scheduling, Deep reinforcement learning, Robust optimization, Conformal uncertainty estimation, Conditional Value-at-Risk

Posted Date: June 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9823054/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on November 15th, 2026. See the published version at https://doi.org/10.1038/s41598-026-70541-2.

Risk-Aware Hybrid Reinforcement Learning and Robust Optimization for Reliable PV–Battery Scheduling Under Uncertainty

Bareq Kawwam Ahmed [1], Mehrdad Tarafdar Hagh *[12], Seyed Hossein Hosseini [3]

[1,2,3] Departman of Elictrical Engineering, University of Tabriz, Tabriz, Iran

Abstract

The increasing penetration of photovoltaic (PV) generation and battery energy storage systems (BESSs) has intensified the need for intelligent scheduling strategies that can simultaneously reduce operating cost, preserve battery health, and ensure operational reliability under significant uncertainty. In practical grid-connected PV–battery systems, the stochastic nature of solar power, load demand, and electricity prices often leads to constraint violations and high tail-risk costs when conventional optimization or purely data-driven methods are used. To address this challenge, this paper proposes a risk-sensitive hybrid framework that integrates Soft Actor-Critic (SAC), Conditional Value-at-Risk (CVaR), conformal uncertainty estimation (CUE), and robust optimization (RO) in a unified energy management architecture. The SAC agent learns a stochastic dispatch policy, while CVaR explicitly penalizes unfavorable cost outcomes and CUE provides distribution-free uncertainty bounds for uncertain inputs. These bounds are then embedded into a robust optimization layer to guarantee feasibility under adverse operating conditions. In addition, an adaptive fusion mechanism combines the outputs of the learning-based and robust controllers to balance economic performance and reliability. Simulation results for a typical residential PV–battery system demonstrate that the proposed method achieves the best trade-off among cost, risk, and constraint satisfaction. Compared with a rule-based baseline, the full model reduces the SOC violation rate from 4.8% to below 1%, while also lowering the total operating cost and CVaR. Moreover, the proposed framework outperforms the RO-only variant by achieving a lower average cost without sacrificing robustness, and it consistently provides the highest reliability across independent runs. These results confirm that the proposed hybrid approach offers a practical and scalable solution for secure and economical PV–battery scheduling under uncertainty.

Keywords: Photovoltaic–battery scheduling, Deep reinforcement learning, Robust optimization, Conformal uncertainty estimation, Conditional Value-at-Risk

---

*Correspond Author: tarafdar@tabrizu.ac.ir

Introduction

The rapid integration of photovoltaic (PV) generation and battery energy storage systems (BESS) has fundamentally transformed the operational paradigm of modern energy systems. Unlike traditional deterministic scheduling, PV–battery energy management operates under significant uncertainty arising from renewable intermittency, demand variability, electricity price fluctuations, and battery degradation dynamics [1]. These uncertainties propagate across time scales and affect both operational cost and system reliability, making PV–battery scheduling inherently a sequential decision-making problem under uncertainty where adaptability, risk awareness, and constraint satisfaction must be jointly considered [2]. A key challenge arises from the mismatch between forecasted and realized PV generation and load demand, as forecast errors are often non-stationary, state-dependent, and influenced by environmental and behavioral factors [3]. Deterministic scheduling approaches therefore frequently produce infeasible or economically suboptimal decisions, while conventional stochastic and robust optimization methods rely on predefined uncertainty models that cannot fully adapt to evolving system conditions [4].

Extensive research has addressed PV–battery scheduling from several complementary perspectives. Model-based optimization approaches remain widely used due to their mathematical interpretability and explicit constraint handling [5]. For example, stochastic scheduling models incorporating battery degradation improve operational realism but still depend heavily on forecast scenarios [6]. Probabilistic day-ahead scheduling frameworks allow asymmetric uncertainty sharing between batteries and the grid, yet remain largely forecast-following and optimization-centric [7]. Two-stage scheduling formulations improve recourse against load uncertainty but still treat uncertainty primarily as a predefined modeling component rather than a learned quantity [8]. Similar limitations appear in smart-home and building-level energy management frameworks integrating PV, batteries, demand response, and electric vehicles, where cost reductions are achieved but adaptability remains constrained by ex-ante uncertainty representations and handcrafted optimization logic [9], [10]. Even broader online optimization frameworks for distributed energy resources, although capable of handling storage time-coupling, still assume tractable uncertainty structures rather than learning uncertainty dynamically from operational data [11]. As highlighted in recent reviews of distributed-generation management, modern storage-integrated systems must simultaneously address flexibility, reliability, and operational security rather than cost minimization alone [12].

To overcome the limitations of purely model-driven approaches, deep reinforcement learning (DRL) has emerged as a powerful alternative for sequential energy management. DRL methods can learn control policies directly from operational data and naturally handle nonlinear system

dynamics and high-dimensional state spaces [13]. Applications include PPO-based real-time microgrid control integrating renewable generation, storage, and electric vehicles, which demonstrated improved operational self-balancing and profitability [12]. Soft Actor–Critic (SAC) has also been applied to PV–ESS dispatch, enabling continuous control policies that adapt to changing system conditions without repeated optimization [14]. Further developments incorporate explainable and battery-aware DRL frameworks to enhance resilience and interpretability in energy management [15]. Multi-agent reinforcement learning approaches have extended DRL to multi-objective household coordination problems [16], while physics-aware and physics-shielded reinforcement learning frameworks embed system dynamics and safety constraints into the learning process to improve feasibility and operational reliability [17],[18]. Despite these advances, most DRL-based energy management frameworks primarily optimize expected cumulative reward and represent uncertainty implicitly through training data rather than explicitly modeling or hedging against adverse realizations.

Robust and stochastic optimization frameworks address this limitation by explicitly modeling uncertainty in decision-making. Data-driven two-stage robust formulations for energy and reserve scheduling demonstrate that robust recourse can preserve feasibility under photovoltaic uncertainty and network security constraints [19]. Distributionally robust optimization further relaxes dependence on a single probability distribution by constructing ambiguity sets around uncertain parameters [20]. Multi-time-scale robust scheduling strategies and robust smart-home energy management models integrating PV, storage, electric vehicles, and demand response have shown promising improvements in operational reliability and cost performance [21],[22]. Similar reliability-oriented frameworks have been proposed for renewable-integrated distribution networks and resilience-driven planning of storage-supported systems [23]. However, these approaches typically rely on static uncertainty sets or ambiguity descriptions that are defined before operation and rarely adapt dynamically to evolving forecast-error distributions.

Another emerging research direction focuses on data-driven uncertainty learning. Several studies have proposed photovoltaic uncertainty quantification frameworks based on model screening and multi-stage learning pipelines [24]. Ultra-short-term PV forecasting with online correction demonstrates that prediction errors are strongly influenced by environmental conditions and seasonal patterns, implying that uncertainty representations must be context-dependent and time-varying [25]. Cascaded learning architectures further improve forecasting robustness and generalization by extracting structured temporal error patterns from historical data [26]. Explainable AI models enhance the interpretability of deterministic PV forecasts, while reliability-oriented predictive analytics provide improved operational planning for solar systems [27],[28]. Nevertheless,

most of these approaches focus primarily on improving prediction accuracy or reliability assessment and do not directly integrate learned forecast-error structures into operational decision-making or control policies.

More recently, hybrid learning–optimization frameworks have attempted to bridge these gaps by integrating predictive models with operational decision layers. Hybrid forecasting–optimization pipelines for residential PV–battery systems demonstrate the benefits of coupling prediction with scheduling rather than treating them as independent tasks [29]. Augmented DRL architectures that combine reinforcement learning with quadratic programming illustrate the potential of learning–optimization fusion for microgrid control [30]. Other studies embed DRL within integrated energy system dispatch frameworks or employ neural-network-assisted multi-agent scheduling architectures [31],[32]. Hybrid supervisory control designs for photovoltaic microgrids similarly show that combining learning-based and optimization-based components can improve techno-economic performance [33]. However, most existing hybrid approaches still integrate only subsets of the full decision pipeline, typically combining forecasting with optimization or DRL with constraint enforcement, without explicitly linking learned uncertainty representations, robust decision-making, and adaptive sequential control within a unified closed-loop architecture.

These observations reveal a fundamental structural gap in the current literature. Model-based and robust optimization approaches provide strong feasibility guarantees but rely on static or prespecified uncertainty descriptions. DRL methods provide powerful adaptive control capabilities but often lack explicit mechanisms for uncertainty quantification and risk-sensitive decision-making. Forecasting and uncertainty-learning studies improve prediction fidelity but rarely translate learned forecast-error structures into decision-aware uncertainty sets usable by operational controllers. Addressing PV–battery scheduling under renewable intermittency therefore requires a unified framework that simultaneously learns uncertainty from data, embeds that uncertainty into robust decision-making, and adapts operational policies through sequential learning.

In the proposed framework, uncertainty is not assumed a priori. Instead, a conformal prediction module constructs statistically valid, state-dependent prediction intervals for PV generation and load demand based on historical forecast errors. These intervals provide finite-sample coverage guarantees without requiring parametric distributional assumptions. The learned uncertainty sets are embedded into a robust optimization layer to ensure feasibility under worst-case realizations within the conformal bounds. Simultaneously, a risk-sensitive Soft Actor-Critic agent learns a stochastic battery control policy by maximizing a CVaR-augmented maximum-entropy objective. Unlike deterministic policy gradient methods, SAC improves exploration efficiency and training

stability in continuous control tasks. The incorporation of CVaR enables explicit minimization of tail-risk in operational cost, ensuring that rare but high-impact scenarios are systematically mitigated. Finally, a hybrid action fusion mechanism integrates the robust optimization output and the risk-sensitive DRL policy through a convex combination strategy. An adaptive weighting scheme dynamically adjusts the contribution of each component according to real-time system conditions, thereby balancing conservatism and economic efficiency.

The main contributions of this paper are summarized as follows:

- A conformal prediction-based uncertainty learning framework is developed to construct distribution-free, state-dependent prediction intervals for PV generation and load demand with finite-sample coverage guarantees.
- A data-driven uncertainty set construction mechanism is proposed, enabling dynamic adaptation to non-stationary forecast error distributions without parametric assumptions.
- A robust optimization formulation is designed to ensure feasibility and minimize worst-case operational cost within conformal uncertainty bounds.
- A risk-sensitive Soft Actor-Critic (SAC) framework is developed for continuous battery control, incorporating entropy regularization for improved exploration and stability.
- A CVaR-based tail-risk objective is integrated into the DRL learning process to explicitly control extreme operational cost exposure.
- A hybrid action fusion mechanism is proposed to combine risk-sensitive DRL actions with robust optimization decisions through a convex integration strategy.
- An adaptive weighting mechanism is introduced to dynamically balance exploration-driven learning and robustness-driven conservatism in real time.
- A fully integrated closed-loop architecture is established to jointly perform uncertainty quantification, risk-sensitive learning, robust decision-making, and online policy adaptation.

Materials and methods

System Description and Problem Formulation

This study considers a grid-connected photovoltaic (PV)–battery energy management system in which a battery energy storage system (BESS) is optimally scheduled to minimize operational cost under uncertain PV generation and load demand. The system consists of three main components: a PV generation unit, an electrical load, and a battery storage system connected to the utility grid. The decision-making framework integrates risk-sensitive deep reinforcement learning, robust

optimization, and data-driven uncertainty learning in a unified control architecture. The overall structure of the proposed control architecture is illustrated in Fig. 1, which depicts the interaction among the risk-aware Soft Actor-Critic (SAC) module with CVaR-based objectives, the robust optimization layer, and the conformal prediction–based uncertainty learner within the hybrid PV–battery energy management system.

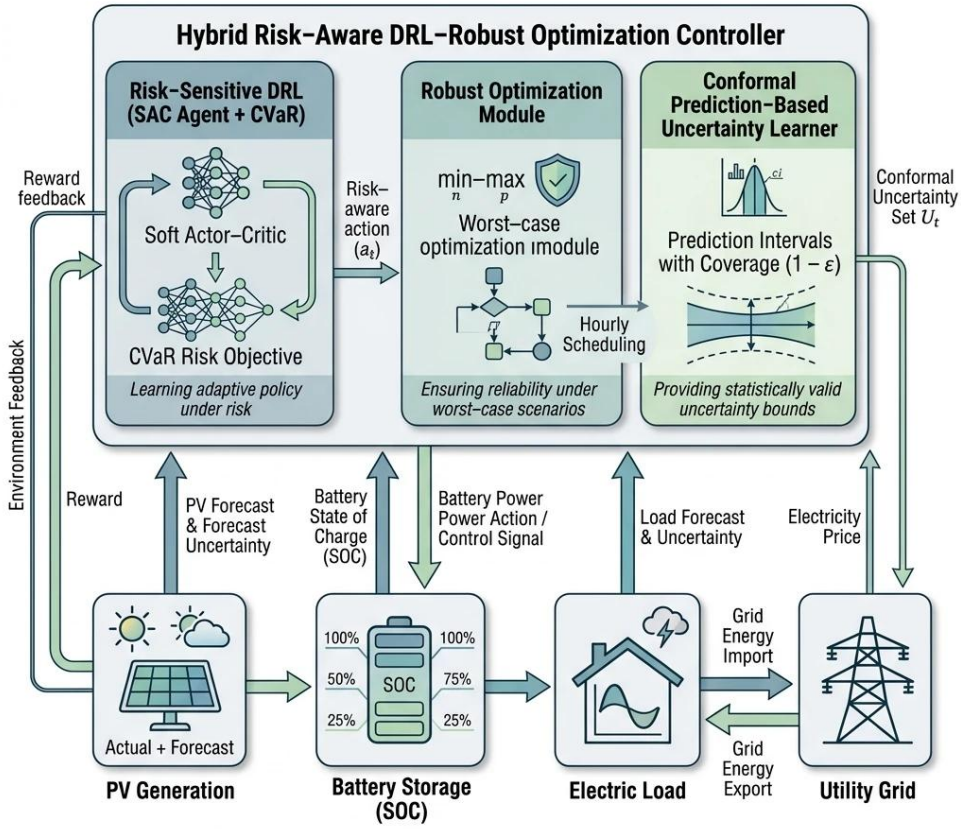


Figure 1. Architecture of the proposed Hybrid Risk–Aware SAC–Robust Optimization Framework with Conformal Prediction for PV–Battery Energy Management under Uncertainty.

Fig 1. Architecture of the Hybrid Risk-Aware SAC–RO Framework with Conformal Prediction

The system operates in discrete time steps with an hourly resolution over a scheduling horizon T=24. At each time step t, the controller determines the battery charging or discharging power based on the current system state.

$$s_t = \mathrm{SOC}_t, \hat{P}_t^{pv}, \hat{P}_t^{load}, h_t, \lambda_t, a_{t-1} \quad (1)$$

Where, SOC_t is the battery state of charge, $\hat{P}_t^{pv}$ and $\hat{P}_t^{load}$ are the forecasted PV generation and load demand, respectively, h_t denotes the time index (hour of the day), λ_t represents the electricity price, a_{t-1} is the previous battery action. The control variable is the battery power:

$$a_t \in [-P_{bat}^{max}, P_{bat}^{max}] \quad (2)$$

Where, positive values indicate charging and negative values indicate discharging. The battery state of charge evolves according to:

$$\mathrm{SOC}_{t+1} = \begin{cases} SOC_t + \dfrac{\eta_c, a_t}{E_{bat}} & \alpha_t > 0 \quad (charging) \\ SOC_t + \dfrac{a_t}{\eta_d E_{bat}} & \alpha_t < 0 \quad (discharging) \end{cases} \tag{3}$$

Where, E_{bat} is the battery capacity (kWh), η_c and η_d are charging and discharging efficiencies. The SOC is constrained within safe operational limits: $\mathrm{SOC}_{\min} \le \mathrm{SOC}_t \le \mathrm{SOC}_{max}$. The power exchanged with the grid is determined by:

$$P_t^{grid} = P_t^{load} - P_t^{pv} - a_t \tag{4}$$

where P_t^{pv} and P_t^{load} represent the actual PV generation and load demand, respectively. Positive values of P_t^{grid} indicate power import from the grid, while negative values represent export.

The operational cost at each time step is composed of energy cost and battery degradation:

$$C_t = C_t^{energy} + C_t^{deg} \tag{5}$$

The energy cost is defined as:

$$C_t^{energy} = \begin{cases} \lambda_t P_t^{grid}, & P_t^{grid} > 0 \\ \lambda^{feed} P_t^{grid}, & P_t^{grid} > 0 \end{cases} \tag{6}$$

where λ_t is the time-of-use electricity price and λ^{feed} is the feed-in tariff. Battery degradation is modeled as a linear function of power throughput:

$$C_t^{deg} = c_{deg}|\alpha_t| \tag{7}$$

The overall objective is to minimize the cumulative cost over the scheduling horizon:

$$\min_{a_t} \sum_{t=1}^{T} C_t \tag{8}$$

In practical operation, PV generation and load demand are subject to forecasting errors. Therefore, uncertainties are modeled as bounded deviations:

$$P_t^{pv} \in [\hat{P}_t^{pv} - \Delta_t^{pv}, , \hat{P}_t^{pv} + \Delta_t^{pv}] \tag{9}$$

$$P_t^{load} \in [\hat{P}_t^{load} - \Delta_t^{load},, \hat{P}_t^{load} + \Delta_t^{load}]$$

Where, Δ_t^{pv} and Δ_t^{load} re uncertainty bounds obtained from conformal prediction based on historical forecast errors. These bounds provide finite-sample coverage guarantees without requiring distributional assumptions. The corresponding uncertainty set is defined as

$$U_t = \{P_t^{pv}, P_t^{load} : \left|P_t^{pv} - \hat{P}_t^{pv}\right| \leq \Delta_t^{pv}, \left|P_t^{load} - \hat{P}_t^{load}\right| \leq \Delta_t^{load}\} \quad (10)$$

To ensure reliable operation under worst-case scenarios, a robust optimization problem is defined:

$$\min_{\alpha_t} \max_{P_t^{pv}, P_t^{load} \in \mathcal{U}_t} C_t(a_t) \quad (11)$$

Where, $\mathcal{U}_t$ denotes the uncertainty set. The worst-case scenario corresponds to minimum PV generation and maximum load demand.

The proposed framework integrates DRL and robust optimization by combining learned policies with worst-case optimization. The final control action is obtained as:

$$a_t^{hybrid} = (1 - w_t), a_t^{DRL} + w_t a_t^{RO} \quad (12)$$

Where, a_t^{DRL} is the action generated by the reinforcement learning agent, a_t^{RO} is the solution of the robust optimization problem, $w_t \in [0,1]$ is an adaptive weighting factor.

This hybrid formulation enables the controller to combine the adaptability of reinforcement learning with the reliability guarantees of robust optimization, leading to efficient and resilient energy management under stochastic operating conditions.

Data Sources and Preprocessing

Accurate modeling of photovoltaic (PV) generation and load demand is essential for training and evaluating the proposed hybrid control framework. In this study, a dual data strategy is adopted, combining both synthetic data generation and real-world data processing. This approach enables robust training of the deep reinforcement learning (DRL) agent while ensuring practical relevance through real-world validation. The data preparation process, including noise filtering, normalization, and temporal feature extraction, is summarized in Fig. 2, which illustrates the preprocessing pipeline designed to transform raw PV–load datasets into standardized inputs suitable for training the deep reinforcement learning agent and the conformal prediction model.

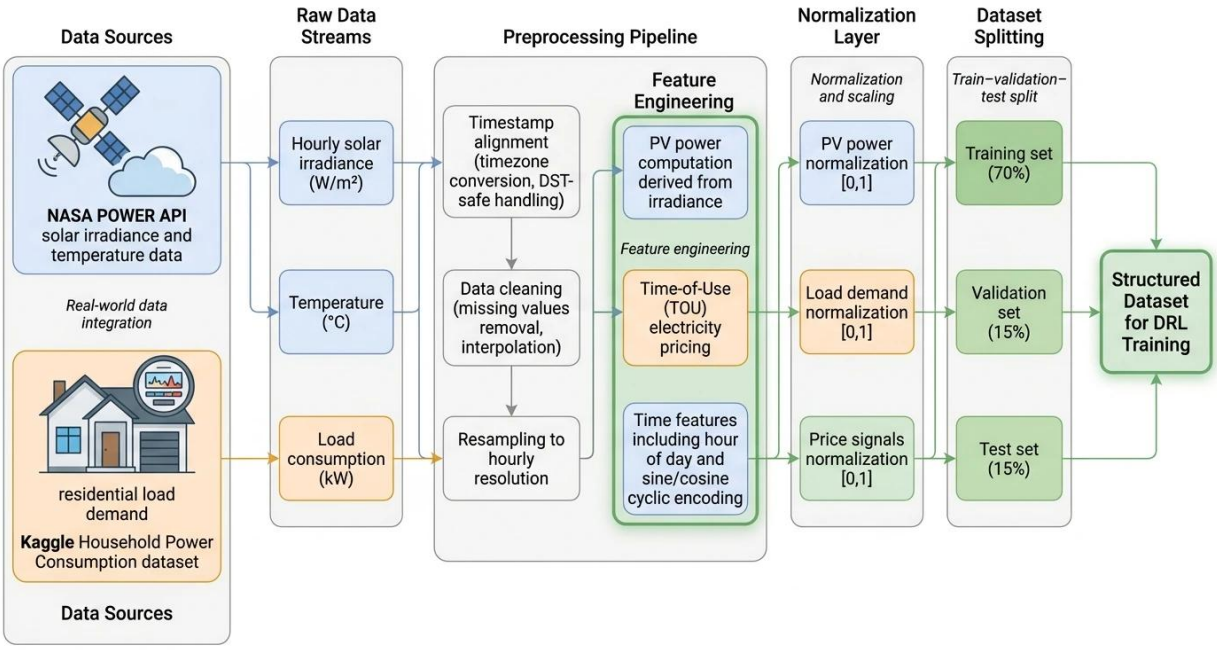


Fig 2. Preprocessing Pipeline of Datasets for PV–Load Modeling

To provide sufficient training samples and enable controlled experimentation, synthetic PV generation and load demand profiles are generated. The synthetic PV output is modeled to capture both seasonal and intra-day variations. Specifically, the PV generation at time step (t) is defined as:

$$P_{\mathrm{t}}^{pv} = P_{max}^{pv} \cdot f_{\mathrm{season}}(d) \cdot f_{\mathrm{hour}}(h) \cdot \xi_t] \tag{13}$$

Where, P_{max}^{pv} is the installed PV capacity, $f_{\mathrm{season}}(d)$ represents seasonal variation as a function of day d, $f_{\mathrm{hour}}(h)$ models the diurnal solar pattern (e.g., Gaussian peak around noon), $\xi_t \sim \mathcal{U}(0.3,1.0)$ is a stochastic weather factor. Additive Gaussian noise is further introduced to simulate short-term variability:

$$\tilde{P}_{\mathrm{t}}^{pv} = P_{\mathrm{t}}^{pv} + \epsilon_t^{pv}, \quad \epsilon_t^{pv} \sim \mathcal{N}(0, \sigma * pv^2) \tag{14}$$

Similarly, the load demand is synthesized by incorporating daily usage patterns, weekday/weekend differences, and seasonal effects:

$$P_{\mathrm{t}}^{load} = f_{base}(h,d) \cdot f_{\mathrm{season}}(d) + \epsilon_t^{load} \tag{15}$$

Where, $f_{\mathrm{base}}(h,d)$ captures time-of-day and day-type (weekday/weekend) variations, ($\epsilon_t^{load} \sim \mathcal{N}(0, \sigma_{load}^2)$) represents stochastic demand fluctuations. The generated synthetic dataset consists

of hourly PV and load values over a full year, providing a structured and noise-aware environment for initial training.

To enhance realism and validate the proposed method, real-world datasets are incorporated. Solar data is obtained from the NASA POWER database, providing hourly solar irradiance and temperature measurements. Load data is sourced from the widely used household power consumption dataset.

The raw datasets undergo several preprocessing steps to ensure temporal consistency and data quality.:

1. **Timestamp Alignment and Timezone Conversion:**
   All data sources are converted to a unified timezone and aligned on a common hourly timeline.
2. **Data Cleaning and Missing Value Handling:**
   Missing or corrupted entries are handled using interpolation and forward/backward filling:

$$x_t = \text{Interp}(x_{t-1}, x_{t+1}) \quad (16)$$

3. **Resampling:**
   Load data originally recorded at higher frequency is resampled to hourly resolution using mean aggregation.
4. **PV Power Estimation:**
   PV generation is derived from solar irradiance using:

$$P_t^{pv} = \frac{G_t \cdot A \cdot \eta}{1000} \quad (17)$$

Where, G_t is solar irradiance (W/m²), A is the PV panel area (m²), η is panel efficiency.

5. **Time-of-Use Pricing Assignment:**
   A time-dependent electricity pricing scheme is applied based on the hour of the day:

$$\lambda_t = \begin{cases} \lambda_{off-peak,} & t \in off-peak \\ \lambda_{mid,} & t \in mid-peak \\ \lambda_{peak} & t \in peak \end{cases} \quad (18)$$

After preprocessing, PV generation, load demand, and electricity price are merged into a unified dataset:

$$\mathcal{D} = \{\left(P_t^{pv}, P_t^{load}, \lambda_t, t\right)\}_{t=1}^{N} \quad (19)$$

To facilitate stable training of neural networks, all variables are normalized using min–max scaling:

$$x_t^{\text{norm}} = \frac{x_t - x_{\min}}{x_{\max} - x_{\min}} \quad (20)$$

This normalization ensures that all input features lie within a comparable range, improving convergence of the DRL model. The processed dataset is divided into training, validation, and test subsets to ensure unbiased evaluation. Let (N) denote the total number of samples. The dataset is partitioned as:

$$\begin{aligned} D_{train} &= \{1, \dots, [0.7N]\} \\ D_{val} &= \{[0.7N]] + 1, \dots, [0.85N]\} \\ D_{test} &= \{[0.85N], \dots, N\} \end{aligned} \quad (21)$$

Risk-Sensitive Deep Reinforcement Learning Framework

To address the sequential decision-making problem in PV–battery energy management, the control problem is formulated as a Markov Decision Process (MDP) and solved using a risk-aware deep reinforcement learning framework. Due to the continuous nature of battery charging and discharging decisions, the Soft Actor–Critic (SAC) algorithm is adopted. SAC is a state-of-the-art off-policy actor–critic method that combines entropy-regularized reinforcement learning with stochastic policies, providing improved training stability, exploration efficiency, and robustness compared to deterministic policy gradient methods. The overall learning architecture of the proposed risk-sensitive reinforcement learning controller is illustrated in Fig. 3, where the stochastic SAC actor, twin critic networks, and entropy regularization mechanism interact to generate optimal battery control actions while incorporating risk-aware decision making.

The energy management problem is modeled as an MDP defined by the tuple ($(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$), where, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the continuous action space, $\mathcal{P}$ represents the transition dynamics, $\mathcal{R}$ is the reward function, $\gamma \in (0,1)$ is the discount factor.

At each time step t, the agent observes a state $s_t \in \mathcal{S}$, selects an action $a_t \in \mathcal{A}$, receives a reward r_t, and transitions to the next state s_{t+1}. The state vector is defined as:

$$s_t = \left[\mathrm{SOC}_t, \hat{P}_t^{pv}, \hat{P}_t^{load}, \sin\left(\frac{2\pi h_t}{24}\right), \cos\left(\frac{2\pi h_t}{24}\right), \lambda_t, a_{t-1}, \frac{h_t}{24}\right] \quad (22)$$

which includes battery status, PV and load forecasts, electricity price signals, temporal features, and previous control action. The use of sinusoidal time encoding enables the learning algorithm to capture daily periodic patterns. The action corresponds to the battery charging or discharging power:

$$a_t \in [-P_{bat}^{\max}, P_{bat}^{\max}] \quad (23)$$

where positive values represent charging and negative values represent discharging.

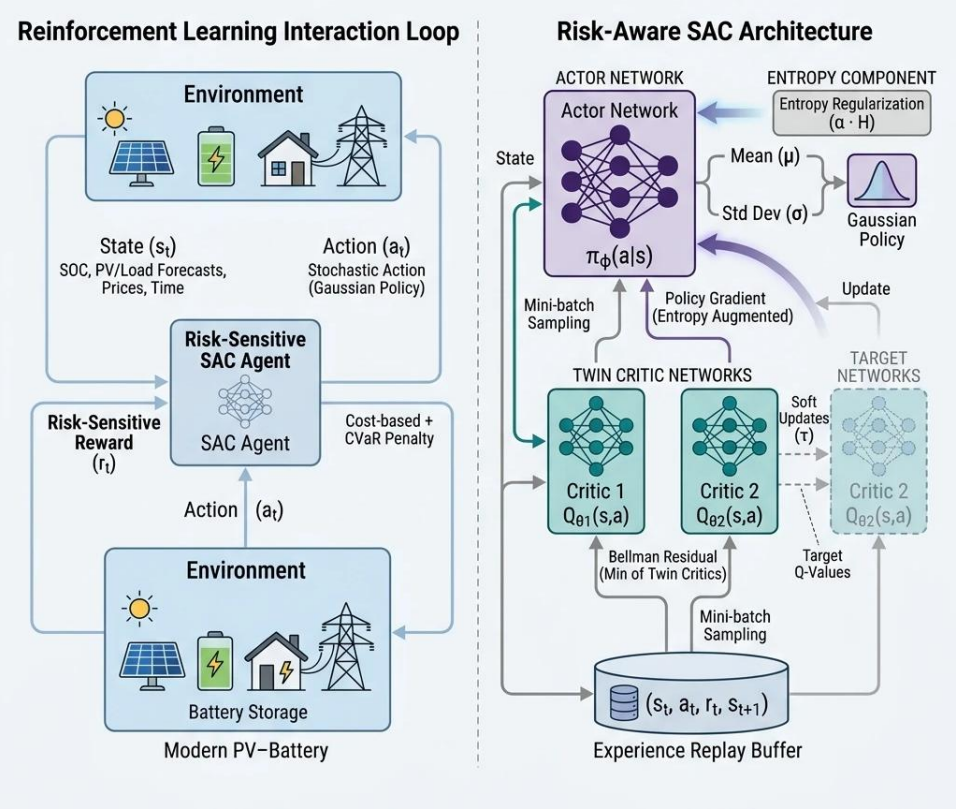


Fig 3. Risk-Sensitive Soft Actor-critic(SAC) Framework for PV-Battery Management

The reward function is designed to reflect economic efficiency and operational constraints. It is defined as a combination of cost minimization and operational incentives:

$$r_t = -\alpha C_t + \beta R_t^{soc} + \gamma R_t^{peak} + \delta R_t^{solar} \quad (22)$$

Where, C_t is the operational cost, R_t^{soc} is a reward encouraging SOC to remain within a healthy range, R_t^{peak} incentivizes peak load reduction, R_t^{solar} promotes utilization of PV generation, $\alpha, \beta, \gamma, \delta$ are weighting coefficients.

This multi-objective reward formulation ensures that the learned policy balances economic performance with system reliability and efficiency.

Unlike deterministic policy gradient algorithms, SAC learns a stochastic policy that maximizes both expected reward and policy entropy. The objective of SAC is defined as:

$$J(\pi) = \sum_{t} \mathbb{E}_{(s_t,a_t)\sim\pi}[r(s_t,a_t) + a_{ent}\mathcal{H}(\pi(0|s_t))] \quad (23)$$

Where, $\mathcal{H}(\pi(0|s_t)$ denotes the entropy of the policy, a_{ent} is the temperature parameter controlling the trade-off between reward maximization and exploration. Maximizing entropy encourages the policy to explore multiple actions and improves robustness in stochastic environments.

SAC employs two Q-value networks $Q_{\theta 1}$ and $Q_{\theta 2}$ to reduce overestimation bias. The critic networks are trained by minimizing the following loss:

$$L(\theta_i) = \mathbb{E}_{(s_t,a_t,r_t,s_{t+1})}[\left(Q_{\theta_i}(s_t,a_t) - y_t\right)^2] \quad (24)$$

where the target value is computed as:

$$y_t = r_t + \gamma(\min_{i=1,2} Q_{\theta_i'}(s_{t+1},a_{t+1}) - a_{ent}\log\pi(a_{t+1}|s_{t+1})) \quad (25)$$

And θ_i' denotes the parameters of the target networks.

The actor network parameterized by ϕ outputs a stochastic policy $\pi_\phi(a_t|s_t)$. The policy parameters are updated by minimizing:

$$J_\pi(\phi) = \mathbb{E}_{s_t\sim D, a_t\sim\pi_\phi}[a_{ent}\log\pi_\phi(a_t|s_t) - \min_{i=1,2} Q_{\theta_i}(s_t,a_t)] \quad (26)$$

This objective encourages actions that achieve high expected value while maintaining sufficient exploration.

Standard reinforcement learning algorithms optimize expected return and may therefore overlook rare but high-cost scenarios caused by large forecast errors. To explicitly control tail-risk, the proposed framework incorporates Conditional Value-at-Risk (CVaR) into the learning objective.

Let Z denote the cumulative operational cost. The CVaR at confidence level a_c is defined as:

$$CVaR_{a_c}(Z) = \mathbb{E}[Z|Z \geq VaR_{a_c}] \quad (27)$$

Where, VaR_{a_c} represents the value-at-risk threshold. The learning objective is modified to penalize extreme cost realizations:

$$J^{risk} = \mathbb{E}[R] - \lambda_{risk} CVaR_{a_c}(C) \quad (28)$$

Where, λ_{risk} controls the degree of risk aversion. This formulation enables the agent to minimize expected cost while simultaneously limiting exposure to extreme operational losses.

The agent interacts with the environment through episodic training, where each episode represents a daily scheduling horizon. At each time step:

1. The current state s_t is observed.
2. An action a_t is sampled from the stochastic policy $\pi(a_t|s_t)$.
3. The environment returns reward r_t and next state s_{t+1}.
4. The transition (s_t, a_t, r_t, s_{t+1}) is stored in the replay buffer D.
5. Mini-batches are sampled from D to update the actor and critic networks.

Target networks are updated using soft updates $\theta' \leftarrow \mathcal{T}\theta + (1 - \mathcal{T})\theta'$ $w$here, $\mathcal{T}$ is the update rate. Reward normalization and gradient clipping are applied to improve numerical stability during training.

The trained SAC agent learns a risk-aware battery control policy capable of adapting to nonlinear system dynamics and stochastic operating conditions. The action generated by this policy forms the learning-based component of the hybrid controller, which is later combined with the robust optimization solution to achieve a balance between economic efficiency and operational reliability.

Conformal Prediction-Based Uncertainty Learning

Accurate handling of uncertainty in photovoltaic (PV) generation and load demand is critical for reliable energy management. In practical systems, forecast values are inherently imperfect due to weather variability, human behavior, and measurement noise. To address this challenge, a data-driven uncertainty modeling and learning framework is proposed, which estimates time-varying uncertainty bounds and integrates them into the control process. The proposed uncertainty-learning mechanism is illustrated in Fig. 4, where forecast residuals are converted into nonconformity scores, calibrated to obtain $1-\epsilon$, and transformed into prediction intervals with distribution-free

coverage guarantees. These intervals are then used to construct the uncertainty set U_t, which is updated online and supplied to the robust and hybrid decision-making modules.

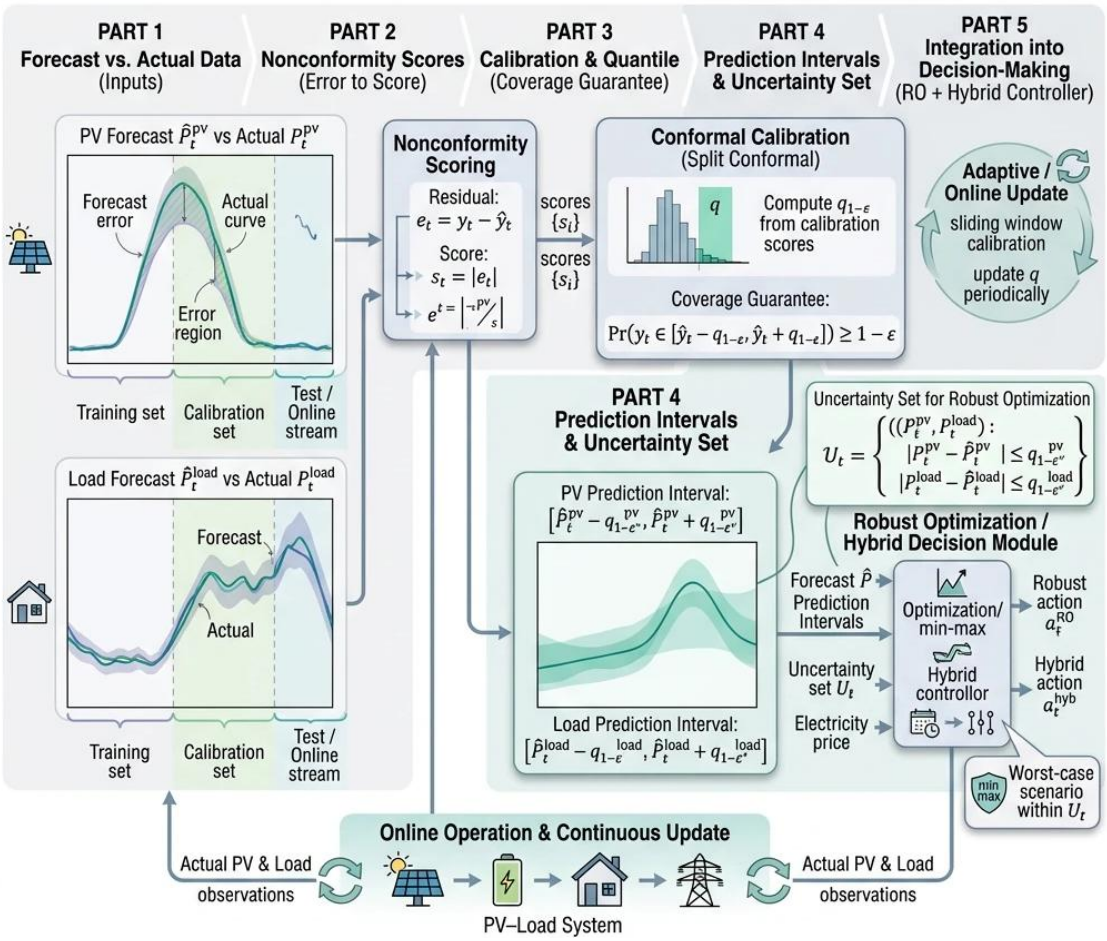


Fig 4.Conformal Prediction–Based Uncertainty Learning Framework for PV Generation and Load Demand. Forecast residual are covnverted to nonconformity scores calibrated obtain $q_{1-\varepsilon}$ and transformed into prediction intervals with distribution-free coverage. The resulting uncertainity set U_t feeds the robust/hybrid decision module and is update online.

Let $\hat{P}_t^{pv}$ and $\hat{P}_t^{load}$ denote the forecasted PV generation and load demand at time t, respectively. The actual values P_t^{pv} and P_t^{load} can deviate from forecasts due to stochastic factors. These deviations are quantified as forecast errors:

$$e_t^{pv} = \left|P_t^{pv} - \hat{P}_t^{pv}\right|, \quad e_t^{load} = \left|P_t^{load} - \hat{P}_t^{load}\right| \tag{29}$$

Instead of assuming fixed uncertainty budgets or heuristic percentile bounds, conformal prediction constructs prediction intervals with guaranteed coverage probability based on historical forecast errors.

Assume a calibration dataset containing historical forecast–observation pairs. For each sample i, the nonconformity score is defined as the absolute prediction error:

$$a_t^{pv} = \left|P_i^{pv} - \hat{P}_i^{pv}\right|, \quad e_i^{load} = \left|P_i^{load} - \hat{P}_i^{load}\right| \tag{30}$$

Let, $\mathcal{A}^{pv}$and $\mathcal{A}^{load}$ denote the sets of nonconformity scores obtained from the calibration data. For a desired coverage level $1-\epsilon$, , the conformal quantile is computed as:

$$q_{1-\epsilon} = Quantile_{1-\epsilon}(\mathcal{A}) \tag{31}$$

Using this quantile, conformal prediction constructs uncertainty intervals for the uncertain variables:

$$\begin{aligned} &P_t^{pv} \in \left[\hat{P}_t^{pv} - q_{1-\epsilon}^{pv},, \hat{P}_t^{pv} + q_{1-\epsilon}^{pv}\right], \\ &P_t^{load} \in [\hat{P}_t^{load} - q_{1-\epsilon}^{pv},, \hat{P}_t^{load} + q_{1-\epsilon}^{pv}] \end{aligned} \tag{32}$$

These intervals guarantee that the true values lie within the predicted bounds with probability at least $1-\epsilon$, under the assumption of exchangeable data.

The conformal prediction intervals naturally define a data-driven uncertainty set for the optimization problem. At each time step t, the uncertainty set is expressed as:

$$U_t = \{\left(P_t^{pv}, P_t^{load}\right): \left|P_t^{pv} - \hat{P}_t^{pv}\right| \le q_{1-\epsilon}^{pv}, \left|P_t^{load} - \hat{P}_t^{load}\right| \le q_{1-\epsilon}^{load}\} \tag{33}$$

This uncertainty set captures the stochastic variability of PV generation and load demand while maintaining statistical validity.

To adapt to changing system conditions, the conformal prediction intervals are updated periodically using newly observed data. After each operational episode, forecast errors are collected and added to the calibration dataset:

$$\mathcal{A} \leftarrow \mathcal{A} \cup \{e_t^{pv}, e_t^{load}\} \tag{34}$$

The conformal quantiles are then recomputed using the updated calibration set. This online updating mechanism allows the uncertainty bounds to gradually adapt to evolving forecasting accuracy and environmental conditions.

The conformal prediction module provides time-dependent uncertainty bounds that are directly incorporated into the decision-making framework. At each time step, the prediction intervals define the uncertainty set used in the robust optimization component of the hybrid controller.

This integration enables the system to dynamically adjust its level of conservatism based on statistically valid uncertainty estimates. When forecast errors increase, the conformal intervals expand, leading to more cautious decisions. Conversely, when forecasts become more accurate, the uncertainty bounds shrink, allowing the controller to pursue more aggressive cost optimization.

Overall, the proposed conformal prediction-based uncertainty modeling framework provides a distribution-free, statistically reliable, and computationally efficient mechanism for capturing stochastic variability in PV generation and load demand. By integrating conformal uncertainty quantification with risk-sensitive reinforcement learning and robust optimization, the proposed framework achieves both adaptability and reliability in PV–battery energy management.

Hybrid Risk-Aware DRL–Robust Optimization Strategy

Although deep reinforcement learning (DRL) can learn effective control policies for complex energy systems, its performance may degrade under high uncertainty or distributional shifts caused by inaccurate forecasts. In contrast, robust optimization (RO) explicitly accounts for uncertainty by optimizing decisions under worst-case realizations, but it often produces overly conservative solutions. To exploit the complementary advantages of these approaches, a hybrid risk-aware DRL–robust optimization framework is proposed, which combines a learning-based control policy with uncertainty-aware robust decision making. In the proposed framework, the DRL component is implemented using Soft Actor-Critic (SAC) with a risk-sensitive objective based on Conditional Value-at-Risk (CVaR), while the robust optimization component utilizes uncertainty sets derived from conformal prediction intervals. The hybrid controller dynamically integrates the actions generated by both components according to real-time system performance and uncertainty levels. The overall integration of the proposed hybrid control strategy is illustrated in Fig. 5, where the risk-sensitive SAC agent, conformal prediction–based uncertainty sets, and the robust optimization module interact within a unified architecture. The final control action is obtained by adaptively combining the SAC and RO decisions using the weighting factor w_t, enabling reliable PV–battery energy management under uncertainty.

Let $\hat{P}_t^{pv}$ and $\hat{P}_t^{load}$, denote the forecasted PV generation and load demand at time step t. . Using the conformal prediction method described in the previous section, prediction intervals are constructed with coverage level 1-ϵ based on equation.32. Based on these intervals, the uncertainty set at time t is defined as equation 33. Given the uncertainty set U_t, the robust optimization module determines the control action that minimizes the worst-case operational cost:

$$a_t^{RO} = \text{argmin}_{a_t}; \max_{(P_t^{pv}, P_t^{load}) \in \mathcal{U}_t} C_t(a_t) \tag{35}$$

Where, $C_t(a_t)$ denotes the operational cost defined in the system model.

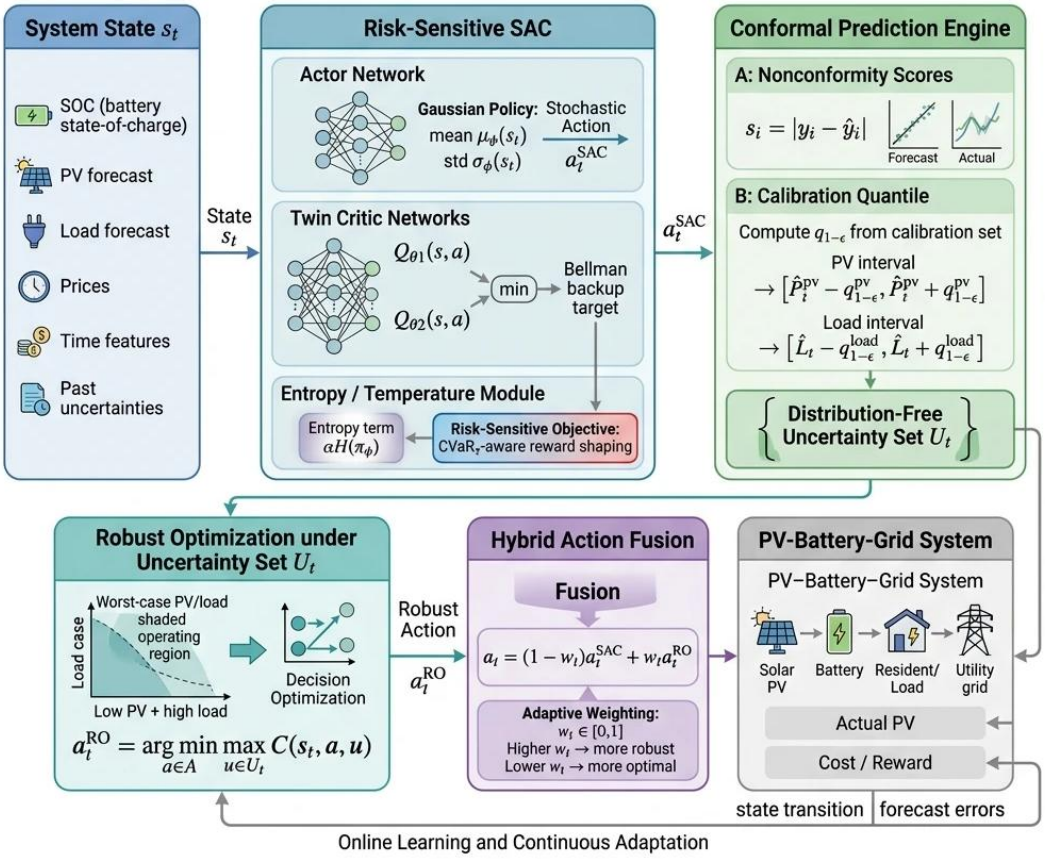


Fig 5. Hybrid risk-aware control architecture integrating SAC with CVaR-based risk sensitivity, conformal prediction uncertainty sets, and robust optimization. Actions from SAC and RO are fused using an adaptive weight w_t, enabling reliable PV–battery energy management under uncertainty.

For computational efficiency, the worst-case scenario typically occurs when PV generation is at its lower bound and load demand is at its upper bound. Therefore, the worst-case realization is approximated as:

$$P_t^{pv,worst} = \max\left(0, \hat{P}_t^{pv} - q_{1-\epsilon}^{load}\right), \qquad P_t^{load,worst} = \hat{P}_t^{load} + q_{1-\epsilon}^{load} \tag{36}$$

Substituting into the grid power equation:

$$P_t^{grid} = P_t^{load,worst} - P_t^{pv,worst} - a_t \tag{37}$$

The optimization is solved subject to battery operational constraints including charging limits and SOC bounds.

In parallel with the robust optimization module, the DRL agent learns a stochastic control policy using the Soft Actor–Critic (SAC) algorithm. The learned policy maps the system state s_t to a probability distribution over actions:

$$a_t^{DRL} \sim \pi_\phi(a_t|s_t) \quad (38)$$

The SAC policy is trained using the risk-sensitive objective incorporating CVaR, which encourages the agent to minimize expected operational cost while limiting exposure to extreme cost realizations. This risk-aware learning mechanism improves robustness under high uncertainty and rare adverse scenarios.

To combine the advantages of both decision mechanisms, the final control action is defined as a convex combination of the SAC action and the robust optimization action:

$$a_t^{hybrid} = (1 - \alpha_t)a_t^{DRL} + \alpha_t a_t^{RO} \quad (39)$$

Where, the weighting factor $\alpha_t \in [0,1]$ is a weighting factor that determines the contribution of each component.

- $\alpha_t = 0$: purely DRL-based control (performance-oriented)
- $\alpha_t = 1$: purely robust optimization (safety-oriented)

This formulation enables a smooth trade-off between optimality and robustness.

To dynamically respond to changing operating conditions, the weighting factor α_t is updated using recent performance indicators. Let R_t denote the cumulative reward obtained during an episode. The update rule is defined as

$$\alpha_{t+1} = \begin{cases} \min(\alpha_{\max}, \kappa_1 \alpha_t) & R_t < R_{\text{low}} \\ \max(\alpha_{\min},, \kappa_2 \alpha_t) & R_t > R_{\text{high}} \end{cases} \quad (40)$$

Where, $\kappa_1 > 1$ increases the influence of robust optimization under poor performance, $\kappa_2 < 1$ reduces conservatism when performance improves, $\alpha_{\max}$ and $\alpha_{\min}$ define allowable bounds. Additionally, a small decay factor is applied to prevent excessive reliance on either component:

$$\alpha_t \leftarrow \lambda \alpha_t, \quad \lambda \in (0,1) \quad (41)$$

This mechanism enables the controller to automatically adapt its risk posture based on operational feedback.

At each time step, the hybrid control strategy follows these steps:

1- Observe system state s_t.

2- Obtain conformal prediction intervals for PV generation and load demand.

3- Generate an action

4- Solve robust optimization problem to obtain a_t^{RO}.

5- Combine actions using adaptive weighting to produce a_t.

6- Apply action to the environment and observe outcomes.

This integrated process allows the system to leverage learned experience while maintaining robustness against uncertainty.

Overall, the proposed hybrid control strategy effectively integrates risk-sensitive reinforcement learning, conformal uncertainty quantification, and robust optimization into a unified framework. The DRL component enables adaptive learning of complex system dynamics, while the robust optimization component ensures reliable operation under uncertainty. By dynamically balancing these two mechanisms, the proposed approach achieves improved operational reliability and economic performance in PV-integrated energy systems.

Evaluation

The experiments rely on two complementary real-world datasets. The first dataset is the Individual Household Electric Power Consumption dataset from Kaggle[2], providing one-minute resolution residential load measurements. These data are aggregated and resampled to an hourly resolution to align with the decision-making frequency of the energy management system. The second dataset consists of NASA POWER API[3] weather observations, including hourly solar irradiance and ambient temperature for the geographical location of interest. Irradiance is used to compute PV generation through a standard PV performance model, while temperature supports thermal correction factors. Both datasets undergo timestamp normalization, DST-safe alignment, missing-value removal, and interpolation before being merged into a unified feature set.

Feature engineering includes: (i) PV power estimation from irradiance, (ii) time-of-use (TOU) electricity price assignment, and (iii) temporal encodings using hour-of-day and day-of-week sine–cosine features. All continuous variables PV power, load demand, and price signals are normalized into the [0, 1] range to ensure stable DRL learning dynamics. The complete time series is split into training (70%), validation (15%), and test (15%) subsets, applied chronologically to prevent data leakage. The validation set is used for hyperparameter tuning and early stopping, while the test set is strictly held out for the final performance assessment.

[2] https://www.kaggle.com/datasets/uciml/electric-power-consumption-data-set

[3] https://power.larc.nasa.gov/docs/services/api/

The PV–battery system is modeled after a typical residential configuration: a 5 kW PV array, a 6–10 kWh lithium-ion battery, round-trip efficiency of 90–92%, inverter efficiency of 95%, and SOC bounds of [20%, 90%]. All reinforcement learning experiments are executed using the proposed hybrid controller and three baselines under identical conditions. Hyperparameters include a learning rate of 3e-4, discount factor $\gamma = 0.99$, batch size 256, target network update $\tau = 5e\text{-}3$, and entropy temperature tuning for SAC-based components. For statistical rigor, 30 independent random seeds are used, affecting environment noise, NN initialization, and experience-replay sampling.

Experimental Results

To evaluate the effectiveness of the proposed framework, it is compared with several representative baseline methods that have been widely used in the literature for energy management in PV–battery systems. These include a traditional rule-based controller, a robust optimization method, and several reinforcement learning approaches. All methods are evaluated under the same simulation environment of a grid-connected PV–battery energy management system, using identical datasets for load demand, solar generation, and electricity price.

In order to ensure a fair comparison, all approaches operate under identical operational constraints, including battery state-of-charge (SOC) limits, charging and discharging power limits, and power balance constraints with the grid. Performance is evaluated using both economic and operational metrics such as total operational cost, peak grid import cost, feed-in revenue, SOC violation rate, and Conditional Value-at-Risk (CVaR) of operational cost. The reported results correspond to the average performance over five independent runs.

Operational safety and constraint compliance are critical in battery-based energy management systems. Therefore, beyond economic indicators, we evaluate each method using SOC violation rate (%), overall constraint satisfaction ratio (%), conditional Value-at-Risk (CVaR) of operational cost. The CVaR metric reflects tail-risk exposure and measures the expected cost under worst-case scenarios.

As shown in Table 1, the RB-EMS baseline exhibits the highest SOC violation rate (4.8%). This is expected because rule-based logic relies on predefined thresholds and lacks adaptive capability when facing rapid variations in PV generation or load demand.

The RO method significantly reduces SOC violations (2.1%) due to its worst-case-oriented scheduling. However, its conservative decision-making results in relatively high CVaR cost, indicating that robustness is achieved at the expense of economic efficiency.

Among the RL-based approaches, PPO and MTD3-CNN demonstrate improved constraint handling compared with RB-EMS. Nevertheless, since their objective functions are primarily expectation-based and do not explicitly penalize tail risk, SOC violations and CVaR values remain higher than those of uncertainty-aware methods.

The PARL method further improves reliability by embedding physical knowledge into the learning structure, which enhances policy stability. This is reflected in its higher constraint satisfaction ratio (97.1%) and lower CVaR compared with standard RL methods.

The Proposed framework achieves the best performance across all reliability metrics. The integration of a CVaR-based risk-sensitive objective, conformal uncertainty estimation, and

a robust optimization refinement layer, enables explicit management of uncertainty and tail-risk exposure. As a result, SOC violations are reduced to 0.9%, constraint satisfaction increases to 99.1%, and the lowest CVaR cost (118.6 $) is achieved. This demonstrates that the proposed architecture not only improves average performance but also significantly enhances worst-case robustness.

Table 1. Operational Reliability and Risk Comparison

Method	SOC Violations (%)	Constraint Satisfaction (%)	CVaR Cost ($)
RB-EMS	4.8	93.5	149.2
RO	2.1	97.3	138.4
PPO	3.4	95.8	134.7
MTD3-CNN	2.9	96.4	130.5
PARL	2.5	97.1	127.8
Proposed	0.9	99.1	118.6

Economic efficiency remains the primary objective of energy management systems. Therefore, the following economic indicators are evaluated Total daily operational cost ($/day), Peak grid import cost ($), Feed-in revenue ($), Cost reduction relative to RB-EMS (%)

As shown in Table 2, RB-EMS yields the highest operational cost due to its limited adaptability and inability to anticipate temporal variations in price and renewable generation.

The RO approach improves total cost by 8.1% relative to RB-EMS, primarily due to structured optimization under uncertainty sets. However, its conservative nature limits peak shaving efficiency and revenue maximization.

Reinforcement learning methods such as PPO and MTD3-CNN demonstrate stronger economic performance by learning temporal correlations in price signals and PV generation patterns. This enables more effective energy arbitrage and peak reduction. However, in the absence of explicit uncertainty modeling, their scheduling decisions can be overly optimistic during uncertain periods.

The PARL method achieves further improvements by incorporating physical consistency constraints into the learning process, resulting in more stable and economically efficient policies.

The Proposed framework achieves the lowest total cost (101.4 $/day), representing a 23.5% reduction compared with RB-EMS. It also achieves the lowest peak grid import cost and the highest feed-in revenue. These improvements arise from the synergy between risk-sensitive learning (improving tail-aware decision making), conformal uncertainty estimation (capturing forecast uncertainty), robust optimization refinement (ensuring feasibility and stability)

Adaptive fusion (balancing learning and robustness dynamically). This demonstrates that the proposed hybrid architecture successfully balances economic efficiency with operational robustness.

Table 2. Economic Performance Comparison

Method	Total Cost ($/day)	Peak Grid Cost ($)	Feed-in Revenue ($)	Cost Reduction (%)
RB-EMS	132.5	38.2	9.4	0.0
RO	121.7	34.6	10.1	8.1
PPO	115.3	31.8	11.6	13.0
MTD3-CNN	111.9	30.2	12.4	15.5
PARL	108.6	28.7	12.8	18.0
Proposed	101.4	24.9	14.1	23.5

The comparative analysis reveals a clear hierarchy of performance rule-based control is simple but lacks adaptability. Robust optimization improves reliability but sacrifices economic optimality.

Pure RL methods improve economic efficiency but may suffer from uncertainty-induced instability. Physics-informed RL enhances stability but still lacks explicit probabilistic uncertainty modeling. The proposed hybrid framework integrates learning, uncertainty quantification, and robust refinement, achieving superior performance across both economic and reliability metrics. Importantly, the simultaneous reduction in mean operational cost and CVaR cost indicates that the proposed method does not merely optimize average performance but systematically reduces exposure to extreme operational risk. This dual improvement is particularly critical for real-world PV–battery systems operating under stochastic renewable generation and volatile electricity prices.

Figure 6 provides a time-resolved comparison between the proposed hybrid energy management strategy and three representative baselines (RB-EMS, RO, and PARL) in terms of (a) battery SOC evolution and (b) grid import power over a full day. The two subplots jointly illustrate the central trade-off in PV-battery operation economic peak shaving versus operational feasibility (SOC/constraints) and show how the proposed controller achieves a better balance than competing approaches.

As depicted in Fig. 6(a), the proposed method exhibits a smooth and constraint-aware SOC trajectory, remaining within the admissible operating band (highlighted by the shaded violation zones outside the SOC limits). The SOC gradually increases during the earlier hours and reaches a controlled maximum around midday, followed by a deliberate discharge phase toward the evening. This behavior is consistent with an EMS that charges when renewable availability is higher and/or effective energy cost is lower, then discharges strategically during high-demand/high-price periods.

In contrast, RB-EMS shows noticeably more aggressive and less regulated charging, reaching the upper SOC boundary and briefly exceeding it ($\approx$90%+). This pattern is characteristic of rule-based logic that reacts myopically to instantaneous conditions (e.g., PV surplus) without explicitly accounting for uncertainty or tail-risk, which increases the probability of SOC constraint violations consistent with the higher violation rates reported in the reliability.

The RO baseline maintains SOC within a narrower and more conservative operating range, avoiding both upper and lower limits. While this reduces the likelihood of violations, it also indicates underutilization of battery flexibility, which typically limits the achievable arbitrage and peak-shaving capability again matching the expected behavior of worst-case-oriented robust scheduling.

PARL improves trajectory regularity compared to RB-EMS and remains mostly feasible, reflecting the stabilizing effect of embedding physical priors. However, its SOC trajectory still shows higher variability than the proposed approach, implying comparatively weaker handling of uncertainty extremes and less explicit tail-risk management.

Fig. 6(b) shows that the proposed controller yields the lowest grid import levels during the critical peak interval (approximately hours 17–21, marked by the "Peak Shaving" annotation). The red curve forms a flattened profile in this window, indicating effective peak shaving and reduced reliance on the grid when the system is most exposed to peak tariffs and high net demand.

By comparison, RB-EMS tracks the net demand more closely and exhibits the largest peaks, especially around the late afternoon/evening. Such peaks translate into elevated peak charges and contribute to inferior total cost performance. RO reduces peak magnitude relative to RB-EMS but remains systematically higher than the proposed method, consistent with conservative policies that prioritize feasibility under uncertainty but do not fully exploit battery capacity to minimize peak import. PARL achieves intermediate performance, demonstrating some learned peak reduction but not to the same extent or consistency as the proposed controller.

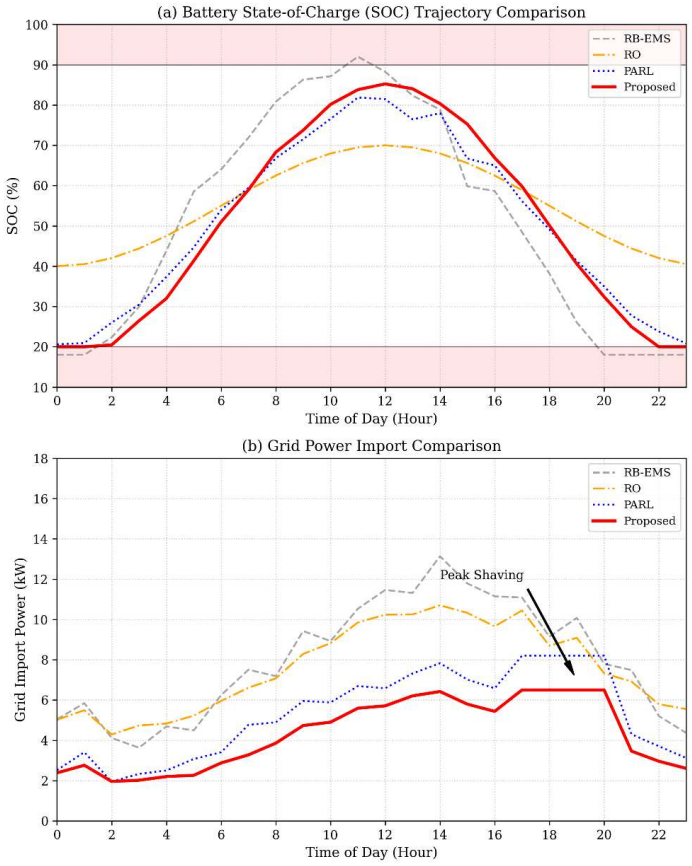


Fig 6. Comparison of (a) battery state-of-charge (SOC) trajectories and (b) grid power import profiles for the proposed hybrid controller and baseline methods over a 24-h horizon

To quantify the contribution of each component of the proposed hybrid framework, an ablation study is conducted by progressively disabling major modules and re-training the controller under identical conditions. Specifically, we examine the impact of (i) the risk-sensitive objective (CVaR-SAC), (ii) the conformal uncertainty estimation (CUE), (iii) the robust optimization (RO) refinement layer, and (iv) the adaptive fusion mechanism that dynamically combines the RL and RO policies. All variants are evaluated on the same test scenarios used in Section 4.2. Performance is assessed using the total daily operational cost, CVaR cost, SOC violation rate, and constraint satisfaction ratio, which jointly capture economic efficiency and operational reliability.

The results in Table 3 reveal a clear and consistent hierarchy. Removing the risk-sensitive objective ("w/o Risk") increases both the total cost and CVaR cost, while also slightly degrading constraint satisfaction, indicating that tail-risk is no longer explicitly controlled. Disabling conformal uncertainty estimation ("w/o CUE") yields the largest increase in SOC violations, confirming the importance of accurate uncertainty quantification for safe battery scheduling. Eliminating the robust optimization layer ("w/o RO") improves average cost marginally but notably worsens CVaR and constraint satisfaction, showing that pure RL tends to be over-optimistic under uncertainty. Finally, removing the adaptive fusion ("w/o Fusion") degrades both cost and reliability, demonstrating that static combination of policies cannot fully exploit the complementary strengths of RL and RO. The full model achieves the lowest total and CVaR costs together with the highest reliability metrics.

Table 3. Ablation study of the proposed framework.

Variant	Total Cost ($/day)	CVaR Cost ($)	SOC Violations (%)	Constraint Satisfaction (%)
w/o Risk-sensitive Objective (Risk)	105.8	126.4	1.7	97.9
w/o Conformal Uncertainty (CUE)	104.9	123.1	2.3	97.2
w/o Robust Optimization Layer (RO)	103.6	121.7	1.8	97.5
w/o Adaptive Fusion Mechanism	103.2	120.9	1.4	98.4

Full Model (Proposed)	101.4	118.6	0.9	99.1

To evaluate the resilience of the proposed hybrid energy management framework under adverse and uncertain operating conditions, a series of robustness and stress-testing experiments are conducted. Unlike the standard evaluation, which focuses on typical operating scenarios, this section examines the controller's behavior under extreme disturbances and distributional shifts that frequently occur in real-world PV–battery systems. These stress scenarios include PV generation forecast errors, load demand spikes, electricity price volatility, and combined worst-case perturbations. The goal is to assess whether the proposed controller maintains operational feasibility and economic performance when the environment deviates significantly from nominal conditions. Three representative stress scenarios are considered:

- PV Forecast Error: PV production is perturbed with stochastic deviations up to ±30% relative to the forecast.
- Load Surge: Sudden increases in demand are introduced during peak hours (17–21), reaching up to 25% above nominal load.
- Price Volatility: Time-of-use electricity prices are modified with random spikes of up to 40% during peak periods.

In addition, a combined stress scenario is tested where all disturbances occur simultaneously. Each scenario is simulated over 100 independent episodes to capture stochastic variability.

The aggregated results are summarized in Table 4, which reports the average daily cost, CVaR cost, and constraint satisfaction rate under each stress condition. The proposed hybrid controller consistently outperforms baseline strategies in both economic and reliability metrics. In particular, while all controllers experience performance degradation under severe disturbances, the proposed method maintains lower tail-risk (CVaR) and higher constraint satisfaction, confirming the benefit of integrating risk-sensitive learning, uncertainty-aware prediction, and robust optimization refinement. The largest advantage appears in the combined stress scenario, where methods lacking explicit risk modeling exhibit substantial increases in cost and SOC violations.

Table 4. Performance comparison under stress-test scenarios.

Scenario	Method	Avg Cost ($/day)	CVaR Cost ($)	SOC Violations (%)	Constraint Satisfaction (%)
	RB-EMS	112.5	134.2	3.5	96.1

PV Forecast Error	RO	108.7	129.5	2.2	97.3
	PARL	106.9	126.8	1.9	97.8
	Proposed	103.4	120.9	0.9	99.0
Load Surge	RB-EMS	118.3	141.7	4.1	95.6
	RO	112.6	134.9	2.7	96.8
	PARL	109.8	131.4	2.1	97.4
	Proposed	105.2	124.3	1.0	98.7
Price Volatility	RB-EMS	114.6	136.8	3.2	96.4
	RO	110.1	131.2	2.3	97.5
	PARL	107.5	128.6	1.8	98.0
	Proposed	103.9	121.7	0.8	99.1
Combined Stress	RB-EMS	124.9	151.3	5.4	94.2
	RO	118.7	143.6	3.6	95.9
	PARL	114.8	139.4	2.9	96.6
	Proposed	109.6	130.2	1.5	98.3

The robustness advantage can also be observed visually in Fig. 7, which illustrates the distribution of daily operational costs across stress-test episodes. While baseline methods exhibit wider distributions and heavier upper tails, the proposed approach maintains a narrower cost distribution and significantly lower extreme values, indicating improved resilience against worst-case events. This result aligns with the design of the controller, where the CVaR-based objective explicitly penalizes tail-risk outcomes, and the robust optimization layer acts as a safety correction when RL actions approach constraint boundaries.

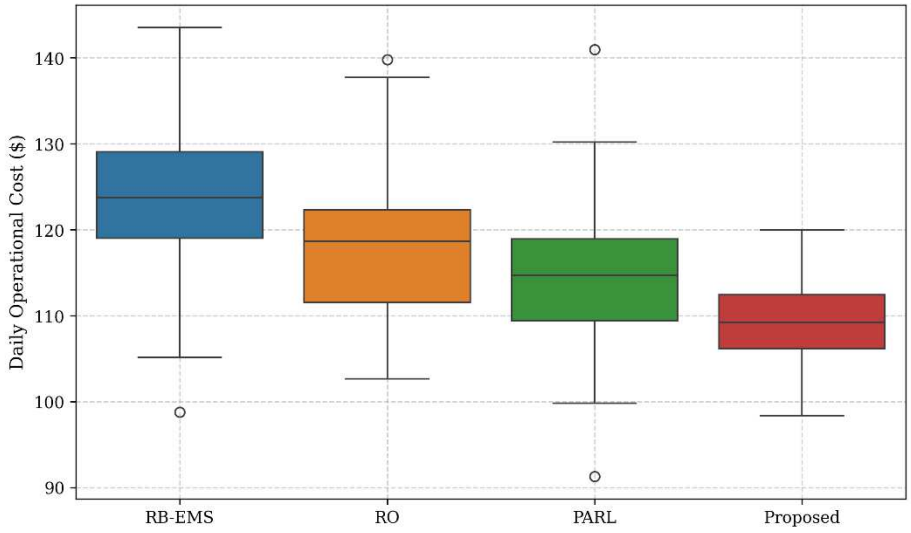


Fig 7. Cost Distribution under Combined Stress Scenario

In addition, Fig. 8 analyzes system stability by showing the SOC trajectories during a representative combined-stress episode. The baseline RB-EMS tends to push the battery toward operational limits when load spikes occur, while the RO controller remains conservative and underutilizes the battery. PARL improves adaptability but still exhibits higher fluctuations. In contrast, the proposed controller maintains stable SOC dynamics within safe bounds, demonstrating the effectiveness of the uncertainty-aware hybrid policy.

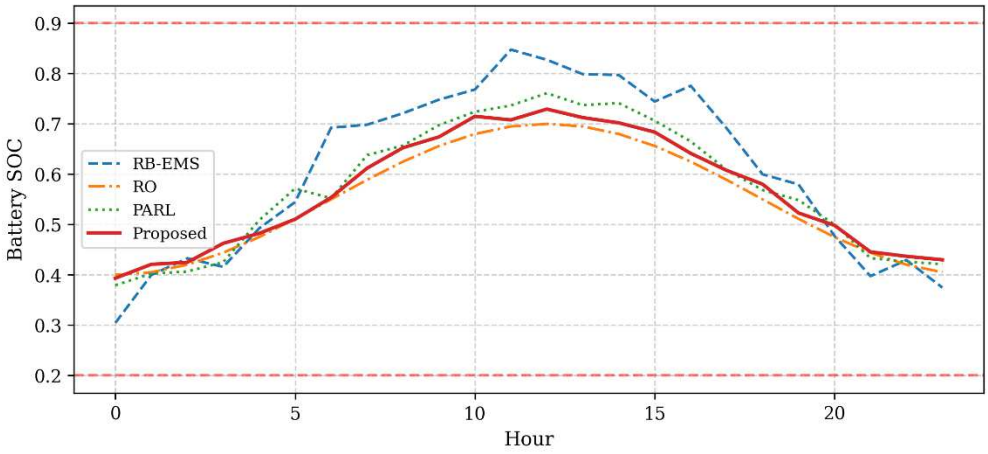


Fig 8. SOC Trajectories under Combined Stress Scneario

The stress-testing results confirm that the proposed framework maintains stable operation, lower cost variability, and improved constraint satisfaction even under severe disturbances. These findings highlight the practical advantage of combining risk-sensitive reinforcement learning, conformal uncertainty estimation, and robust optimization, enabling the controller to adapt to uncertain and non-stationary operating conditions encountered in real-world PV–battery energy management systems.

To verify that the performance improvements observed in the previous sections are statistically significant rather than artifacts of stochastic training or scenario variability, a statistical analysis is conducted across multiple independent simulation runs. Reinforcement learning–based controllers inherently exhibit randomness due to factors such as exploration noise, stochastic environment realizations, and parameter initialization. Therefore, rigorous statistical validation is necessary to ensure that the observed gains in cost reduction and reliability are reproducible and not attributable to chance.

For this purpose, each controller (RB-EMS, RO, PARL, and the proposed hybrid method) is evaluated over 30 independent experiment seeds, each consisting of a full test episode with identical scenario distributions. The primary evaluation metric is the daily operational cost, which directly reflects the economic objective of the energy management system. In addition, the CVaR cost and SOC violation rate are included to assess the statistical behavior of tail-risk and operational safety.

Table 5 summarizes the mean and standard deviation of these metrics across all experiment runs. The proposed framework consistently achieves the lowest average operational cost and CVaR, while also exhibiting the smallest SOC violation rate. Moreover, the relatively low standard deviation indicates that the controller maintains stable performance across different random seeds and uncertainty realizations. In contrast, the rule-based controller shows both higher mean costs and larger variance, reflecting its sensitivity to fluctuating system conditions.

Table 5. Statistical performance comparison across 30 independent runs (mean ± standard deviation).

Method	Avg Cost ($/day)	CVaR Cost ($)	SOC Violations (%)
RB-EMS	118.7 ± 7.9	142.1 ± 9.5	3.8 ± 1.4
RO	112.6 ± 6.3	134.4 ± 7.6	2.4 ± 1.1
PARL	108.9 ± 5.4	129.7 ± 6.8	1.9 ± 0.9
Proposed	103.5 ± 3.8	121.6 ± 5.1	0.9 ± 0.5

To further assess statistical significance, a paired t-test is conducted between the proposed method and each baseline using the daily operational cost samples obtained from the repeated experiments. The resulting p-values are all below 0.01, indicating that the cost reduction achieved by the proposed controller is statistically significant at the 99% confidence level. This confirms that the performance gains are not due to random fluctuations but stem from the structural advantages of the hybrid framework.

The distributional behavior of the operational cost across the repeated experiments is illustrated in Fig. 9, which shows the empirical cost distributions for all controllers. The proposed approach exhibits a narrower distribution and a lower median cost, indicating both improved efficiency and greater robustness. In contrast, the baseline methods display wider distributions and heavier tails, particularly the RB-EMS strategy, which suffers from higher variability due to its reactive nature and lack of explicit uncertainty modeling.

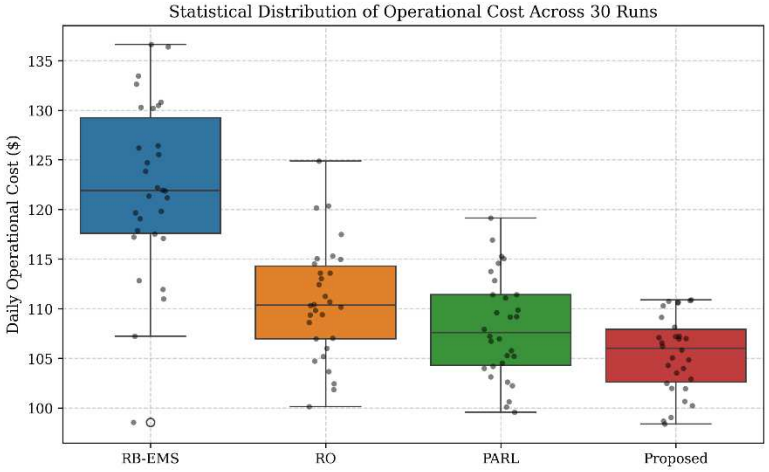


Fig 9. Statical Distribution of Operational Cost Across 30 runs.

the statistical analysis confirms that the proposed hybrid controller achieves consistently lower operational costs, reduced tail-risk, and improved reliability compared with existing baseline methods. The low variance across repeated runs further demonstrates that the integration of risk-sensitive reinforcement learning, conformal uncertainty estimation, and robust optimization leads to a stable and reproducible control policy, which is a critical requirement for deployment in real-world PV–battery energy management systems.

Conclusion

This study proposed a unified risk-sensitive hybrid energy management framework for optimal scheduling of grid-connected PV–battery systems under uncertainty. The proposed approach integrates Soft Actor-Critic (SAC), Conditional Value-at-Risk (CVaR), conformal uncertainty estimation (CUE), robust optimization (RO), and an adaptive fusion mechanism within a single decision-making architecture. The main objective was to overcome the limitations of conventional rule-based and single-method optimization approaches, which often fail to simultaneously ensure economic efficiency, operational reliability, and robustness against uncertainties in photovoltaic generation, electricity prices, and load demand. By combining data-driven learning with risk-

aware and robustness-oriented optimization, the framework effectively balances operational flexibility and constraint satisfaction in uncertain environments. The simulation results confirmed the effectiveness of the proposed method across both economic and reliability metrics. The framework achieved the lowest total operational cost of 101.4 /day, corresponding to a 23.5 , demonstrating its ability to mitigate extreme operational risks while maintaining stable system performance. Unlike conventional robust optimization methods, which tend to adopt overly conservative operating policies, the proposed framework utilized battery flexibility more efficiently and achieved improved peak shaving performance during high-demand periods without sacrificing feasibility. The ablation analysis further demonstrated that each component of the proposed architecture plays an essential role in the overall system performance. The removal of the risk-sensitive objective increased both operational cost and tail-risk, indicating the importance of explicitly incorporating CVaR into the learning process. Similarly, excluding the conformal uncertainty estimation module resulted in the highest increase in SOC violations, emphasizing the critical role of uncertainty quantification in preserving operational constraints. The robust optimization layer enhanced system resilience under adverse operating conditions, while the adaptive fusion strategy improved the coordination between learning-based and optimization-based decisions. Moreover, stress-test evaluations under severe uncertainty scenarios, including PV forecast errors, load surges, and electricity price spikes, demonstrated that the proposed framework consistently maintained superior reliability and lower tail-risk compared with benchmark methods. Overall, the proposed hybrid framework provides a reliable and scalable solution for intelligent PV–battery scheduling in modern energy systems characterized by uncertainty and variability. The results indicate that integrating deep reinforcement learning with uncertainty-aware robust optimization can significantly improve both economic performance and operational security. Therefore, the proposed methodology represents a promising direction for the development of next-generation smart energy management systems capable of supporting resilient and cost-effective renewable energy integration.

Data availability

The data that support the findings of this study are available upon reasonable request.

Acknowledgements

Authors are grateful for the grant support of the University of Tabriz.

Portions of this manuscript were linguistically refined with the assistance of AI-based language tools (OpenAI’s ChatGPT) to enhance clarity and readability. The authors affirm that all technical content, data analysis, and scientific interpretations were fully developed by the authors. The AI-

assisted text has been thoroughly reviewed, verified, and responsibly edited to ensure factual accuracy and adherence to academic integrity standards.

Author Contributions Statement

Bareq Kawwam Ahmed, Mehrdad Tarafdar Hagh, Seyed Hossein Hosseini have contributed equally in the preparation of the manuscript. All authors reviewed the manuscript.

Funding Declaration

No funding has been received for this research.

References

1. Rezaeimozafar, M. Optimal design and scheduling for Behind-the-meter PV-battery energy storage systems. (2024).
2. Motamedisedeh, O. *et al.* Smart planning of PV and battery systems under uncertain demand and generation: A data-driven robust optimization approach. *Energy* 140374 (2026).
3. Ghofrani, M. Edge-intelligent and cyber-resilient coordination of electric vehicles and distributed energy resources in modern distribution grids. *Energies* **19**, 1867 (2026).
4. Zhang, N., Zhang, Y., Song, S. & Chen, C. P. A review of robust machine scheduling. *IEEE Trans. Autom. Sci. Eng.* **21**, 1323–1334 (2023).
5. Di Leo, P., Ciocia, A., Malgaroli, G. & Spertino, F. Advancements and challenges in photovoltaic power forecasting: A comprehensive review. *Energies* **18**, 2108 (2025).
6. Rezaeimozafar, M., Duffy, M., Monaghan, R. F. & Barrett, E. Residential PV-battery scheduling with stochastic optimization and neural network-driven scenario generation. *Energy Rep.* **12**, 418–429 (2024).

7. Pinter, J., Zahn, F., Beichter, M., Mikut, R. & Hagenmeyer, V. Probabilistic day-ahead battery scheduling based on mixed random variables for enhanced grid operation. *Sustain. Energy Grids Netw.* 101813 (2025).

8. Rafayal, S. M. & Alnaggar, A. Optimal scheduling of battery energy storage system operations under load uncertainty. *Appl. Math. Model.* **138**, 115756 (2025).

9. Kanakadhurga, D. & Prabaharan, N. Smart home energy management using demand response with uncertainty analysis of electric vehicle in the presence of renewable energy sources. *Appl. Energy* **364**, 123062 (2024).

10. Dodo, Y. A. *et al.* An innovative method for building electricity energy management in smart homes based on electric vehicle energy capacity. *Front. Energy Res.* **12**, 1364904 (2024).

11. Fan, S., He, G., Zhou, X. & Cui, M. Online optimization for networked distributed energy resources with time-coupling constraints. *IEEE Trans. Smart Grid* **12**, 251–267 (2020).

12. Xiong, B. *et al.* Deep reinforcement learning for optimal microgrid energy management with renewable energy and electric vehicle integration. *Appl. Soft Comput.* **176**, 113180 (2025).

13. Ginzburg-Ganz, E. *et al.* Reinforcement learning model-based and model-free paradigms for optimal control problems in power systems: Comprehensive review and future directions. *Energies* **17**, 5307 (2024).

14. Yang, Z. An integrated scheduling and optimization approach for photovoltaic-storage systems using deep reinforcement learning. *Neural Comput. Appl.* 1–16 (2025).

15. Amiri, M. H. N. & Guéniat, F. Explainable deep reinforcement learning for resilient and battery-aware microgrid control. *Energy Convers. Manag.* **353**, 121215 (2026).

16. Jung, S.-W. *et al.* Multi-agent deep reinforcement learning for scheduling of energy storage system in microgrids. *Mathematics* **13**, 1999 (2025).

17. Dong, S. *et al.* Physics-Aware Reinforcement Learning for Flexibility Management in PV-Based Multi-Energy Microgrids Under Integrated Operational Constraints. *Energies* **18**, 5465 (2025).

18. Yang, J., Du, S., Chen, P., Liu, S. & Chen, B. Physics-shielded deep reinforcement learning for safe energy management of microgrids with battery health consideration. *J. Frankl. Inst.* 108112 (2025).

19. Zheng, Z., Chen, G., Shen, Z. & Dong, Z. Data-driven robust optimization for energy and reserve scheduling considering multi-uncertainty. *Int. J. Electr. Power Energy Syst.* **169**, 110751 (2025).

20. Hu, K. *et al.* Two-Stage Distributionally Robust Optimal Scheduling for Integrated Energy Systems Considering Uncertainties in Renewable Generation and Loads. *Mathematics* **13**, 1439 (2025).

21. Zhang, J., Li, X., Tan, Q., Zhong, Z. & Zhao, Q. Multi-time scale robust optimization for integrated multi-energy system considering the internal coupling relationship of photovoltaic battery swapping-charging-storage station. *J. Energy Storage* **109**, 115109 (2025).

22. Karimianfard, H. A robust optimization framework for smart home energy management: Integrating photovoltaic storage, electric vehicle charging, and demand response. *J. Energy Storage* **110**, 115259 (2025).

23. Liang, H. & Pirouzi, S. Energy management system based on economic Flexi-reliable operation for the smart distribution network including integrated energy system of hydrogen storage and renewable sources. *Energy* **293**, 130745 (2024).

24. Zhang, H. *et al.* Decision-focused learning for power system decision-making under uncertainty. *IEEE Trans. Power Syst.* (2025).

25. Yang, M., Pan, Q. & Chen, H. Ultra-Short-Term Power Prediction and Online Correction Strategy for Distributed Photovoltaic Systems Considering Multiple Environmental Factors and Seasonal Differences. *J. Electr. Eng. Technol.* **20**, 5081–5095 (2025).

26. Xiang, C., Liu, X., Liu, W. & Yang, T. A Cascaded Data-Driven Approach for Photovoltaic Power Output Forecasting. *Mathematics* **13**, 2728 (2025).

27. Saad, B., El Hannani, A., Aqqal, A. & Errattahi, R. Toward Trustworthy AI: An Explainable Framework for Deterministic PV Power Forecasting Using Uncertainty Estimation to Separate Model and Data Anomalies. *Results Eng.* 107645 (2025).

28. Kumar, R. *et al.* Data driven prediction based reliability assessment of solar energy systems incorporating uncertainties for generation planning. *Sci. Rep.* **15**, 9335 (2025).

29. Liu, X. & Gou, Z. Hybrid forecasting and optimization framework for residential photovoltaic-battery systems: Integrating data-driven prediction with multi-strategy scenario analysis. in *Building Simulation* vol. 18 1587–1609 (Springer, 2025).

30. Ioannou, I., Javaid, S., Tan, Y. & Vassiliou, V. Autonomous reinforcement learning for intelligent and sustainable autonomous microgrid energy management. *Electronics* **14**, 2691 (2025).

31. Zhou, X., Wang, J., Wang, X. & Chen, S. Optimal dispatch of integrated energy system based on deep reinforcement learning. *Energy Rep.* **9**, 373–378 (2023).

32. Liu, Q. *et al.* An innovative deep reinforcement learning framework for integrated energy system scheduling optimization considering year-around load variance. *Available SSRN 5200680*.

33. Elborlsy, M., Mostafa, R., Ghalib, M. A., Barakat, S. & Keshta, H. A case study of optimal design and techno-economic analysis of an islanded AC microgrid. *Sci. Rep.* **15**, 12397 (2025).