Risk-Sensitive Microgrid Management Using Proximal Policy Optimization

Christopher G. Harris
University of Northern Colorado
Greeley, CO, USA
christopher.harris@unco.edu

***Abstract*—We propose RS-PPO, a risk-sensitive reinforcement learning framework for energy management in residential microgrids operating under uncertainty. RS-PPO extends Proximal Policy Optimization by integrating Conditional Value-at-Risk (CVaR) into the objective, enabling robust scheduling decisions that account for extreme cost outcomes. The agent learns to balance energy cost, peak demand, and battery longevity while responding adaptively to real-time conditions without requiring future forecasts. We evaluate RS-PPO using real-world solar and demand data, comparing it to Greedy, Rule-Based, standard PPO, and an oracle Model Predictive Control baseline. RS-PPO consistently improves reliability, reduces peak loads, and lowers exposure to cost volatility. It approaches the performance of MPC while maintaining generalization under uncertainty, demonstrating its suitability for deployment in demand-responsive smart grid environments.**

I. Introduction

The increasing penetration of renewable energy in distributed energy systems has amplified the need for intelligent microgrid energy management strategies. Microgrids, particularly those operating in isolated or partially autonomous modes, must make real-time decisions under uncertainty—balancing cost, reliability, and battery longevity while responding to fluctuating supply and demand. This is particularly challenging when renewable generation (e.g., solar or wind) is intermittent and forecasts are unreliable [1], [2].

Traditional approaches, such as rule-based scheduling or model predictive control (MPC), rely heavily on short-term forecasts of load and generation [3], [4]. While effective under perfect information, these methods degrade in the presence of uncertainty, often leading to suboptimal or risk-prone behavior. Moreover, they require continuous re-optimization and tuning, limiting their applicability in real-world, data-constrained environments [5].

Reinforcement learning (RL) offers a compelling alternative by learning policies that adapt directly from interaction with the environment [6]. RL-based controllers can generalize across unseen scenarios and operate under partial information, potentially obviating the need for future forecasts. However, standard RL algorithms such as Proximal Policy Optimization (PPO) [7] tend to optimize expected returns, which may lead to unsafe or high-variance behavior in critical systems like microgrids. For instance, an agent optimizing only for average cost may underprepare for rare but severe demand spikes, resulting in blackouts or excessive grid draw during peak periods [8], [9].

To address these limitations, we propose a **risk-sensitive variant of Proximal Policy Optimization (RS-PPO)** tailored to the microgrid scheduling problem. Our method integrates Conditional Value at Risk (CVaR) [10] into the policy gradient framework, explicitly penalizing high-cost outcomes while maintaining tractable optimization. Unlike MPC, our method requires no reliance on future forecasts. Unlike traditional PPO, it yields more conservative and reliable decisions under uncertainty [11], [12].

The contributions of this paper are as follows:

- We formulate microgrid energy management as a constrained Markov decision process with risk-sensitive objectives, incorporating cost, peak load, and battery cycling.
- We develop RS-PPO, a constrained, risk-aware policy gradient algorithm that minimizes CVaR while satisfying operational constraints.
- We demonstrate the robustness and reliability of our approach across three forecasting regimes (perfect, short-term, and none) using real-world solar and demand traces from the Pecan Street dataset [13].
- We evaluate against traditional PPO and MPC baselines, showing statistically significant improvements in peak load reduction, CVaR minimization, and sustainability indicators such as battery wear [2], [14].

Our findings suggest that RS-PPO can operate competitively with oracle-style MPC controllers, even without access to forecasts, making it a promising candidate for real-world deployments in dynamic and uncertain microgrid environments [15], [16].

The remainder of this paper is organized as follows: Section II reviews related work in microgrid energy management and risk-sensitive reinforcement learning. Section III defines the problem formulation and system constraints. Section IV introduces the RS-PPO methodology, followed by the experimental setup in Section V. Section VI presents performance results and analysis, while Section VII compares our approach to related methods. Finally, Section VIII concludes the paper and outlines future research directions.

II. Related Work

Microgrid energy management has received increasing attention due to growing renewable energy integration and the need for safe and adaptive control policies. Model Predictive

Control (MPC) remains a widely used technique for deterministic and probabilistic optimization in energy scheduling [17]–[19], but is often dependent on accurate forecasts and re-optimization at every time step.

In contrast, Deep Reinforcement Learning (DRL) methods learn policies through direct interaction with the environment, adapting to stochastic inputs without explicit modeling. Prior work has used DDPG for real-time scheduling [14], actor-critic methods [20], and batch RL for load control [21]. However, most of these target expected return minimization, which may be ill-suited for safety-critical domains.

Risk-sensitive RL addresses this limitation via metrics like Conditional Value-at-Risk (CVaR) [22], [23]. Zhou et al. [15] apply CVaR-DQN for isolated microgrids, but restrict the action space to discrete levels. More recent works incorporate CVaR into policy gradient methods for general continuous control [24], though without tailored applications to microgrids.

Several recent studies explore PPO-based control for energy systems. Wang et al. [25] apply multi-agent PPO to real-time microgrid scheduling, demonstrating scalability, while Cuadrado et al. [26] investigate federated transfer learning to improve generalization in zero-net energy settings. Das et al. [27] combine RL with weather-aware scheduling for solar microgrids. These works validate the effectiveness of PPO-style methods but lack explicit treatment of risk metrics like CVaR.

Our proposed RS-PPO method bridges this gap by combining continuous control via PPO with CVaR-regularized policy gradients, providing robust, fine-grained control in stochastic microgrid settings.

III. Problem Formulation

Our proposed control framework is illustrated in Figure 1, which depicts the interaction between the policy, value function, environment dynamics, and risk-sensitive reward shaping. The agent observes the microgrid state—comprising PV output, load, battery SOC, and grid power—and selects continuous actions to manage battery and grid flows under operational constraints.

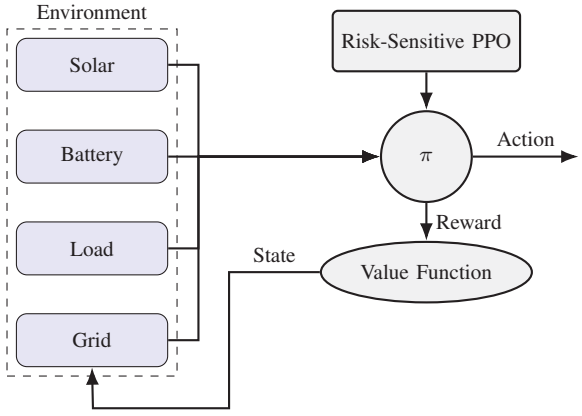

Fig. 1. PPO Framework for Microgrid Energy Management.

We formulate the microgrid energy management task as a finite-horizon Markov Decision Process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition probability, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma \in [0, 1]$ is the discount factor.

A. Microgrid Dynamics

The microgrid consists of photovoltaic (PV) generation, a battery storage system, and electrical loads. The grid can import energy when local generation is insufficient. The state at time t is given by:

$$s_t = \left[P_t^{\text{pv}}, P_t^{\text{load}}, SOC_t, P_t^{\text{grid}}\right], \tag{1}$$

where P_t^{pv} is PV output, P_t^{load} is total load, $SOC_t \in [0, 1]$ is the battery's state of charge, and P_t^{grid} is grid power import.

The action $a_t \in \mathcal{A}$ consists of continuous control decisions:

$$a_t = \left[P_t^{\text{bat}}, P_t^{\text{grid}}\right], \tag{2}$$

where P_t^{bat} is battery power (positive for discharging, negative for charging), and P_t^{grid} is grid import/export.

B. Operational Constraints

The actions must satisfy:

$$P_{\min}^{\text{bat}} \le P_t^{\text{bat}} \le P_{\max}^{\text{bat}}, \tag{3}$$

$$SOC_{t+1} = SOC_t + \eta_c \cdot P_t^{\text{bat}} \cdot \Delta t, \tag{4}$$

$$0 \le SOC_{t+1} \le 1, \tag{5}$$

$$P_t^{\text{grid}} \ge 0 \quad \text{(no export allowed)}. \tag{6}$$

The system must satisfy power balance:

$$P_t^{\text{pv}} + P_t^{\text{bat}} + P_t^{\text{grid}} = P_t^{\text{load}}. \tag{7}$$

C. Reward Function

The agent receives a scalar reward r_t at each step to minimize cost and risk:

$$r_t = -\left(c_t^{\text{grid}} \cdot P_t^{\text{grid}} + \lambda_{\text{deg}} \cdot D(SOC_t) + \lambda_{\text{peak}} \cdot \mathbb{I}(P_t^{\text{grid}} > \theta)\right), \tag{8}$$

where c_t^{grid} is the grid price, $D(SOC_t)$ models battery degradation, θ is a peak power threshold, and $\lambda_{\text{deg}}, \lambda_{\text{peak}}$ are weighting coefficients. $\mathbb{I}(\cdot)$ is the indicator function.

D. Risk-Sensitive Objective

Rather than minimizing expected cumulative cost alone, we optimize the Conditional Value-at-Risk (CVaR) of the return:

$$J_{\text{CVaR}}(\pi) = \min_{\nu \in \mathbb{R}} \left\{ \nu + \frac{1}{1-\alpha} \mathbb{E}_\pi \left[(G - \nu)^+ \right] \right\}, \tag{9}$$

where $G = \sum_{t=0}^{T} \gamma^t r_t$ is the discounted return and $\alpha \in (0, 1)$ is the CVaR confidence level. The policy π aims to minimize risk-exposed losses in the worst α-quantile of outcomes.

E. Objective

The agent seeks a policy π_θ, parameterized by θ, that minimizes the risk-sensitive return:

$$\pi^* = \arg\min_\pi J_{\text{CVaR}}(\pi), \tag{10}$$

subject to the system dynamics and operational constraints outlined above.

IV. Methodology

We formulate microgrid energy management as a sequential decision-making problem under uncertainty, modeled as a Markov Decision Process (MDP). The agent observes system states and selects actions to minimize both expected cost and risk exposure while satisfying operational constraints.

A. Problem Definition

Let the MDP be defined by $(\mathcal{S}, \mathcal{A}, P, c, \gamma)$, where $\mathcal{S}$ is the set of states (e.g., battery level, net demand, solar forecast), $\mathcal{A}$ is the set of actions (e.g., battery charge/discharge rates), P is the transition probability, $c(s, a)$ is the instantaneous cost function, and $\gamma \in [0, 1]$ is the discount factor. The goal is to learn a policy $\pi(a \mid s)$ that minimizes not only the expected cumulative cost, but also its tail risk.

B. CVaR-Regularized PPO Objective

We adopt Proximal Policy Optimization (PPO) [28] with an augmented objective that incorporates Conditional Value-at-Risk (CVaR) [10]. The standard PPO objective seeks to maximize expected advantage while constraining policy updates:

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_t \left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t) \right], \tag{11}$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the importance sampling ratio, and $\hat{A}_t$ is the generalized advantage estimate.

To promote risk-averse behavior, we regularize this objective using CVaR at confidence level $\alpha \in (0, 1)$ over the episode return distribution R^π:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{PPO}}(\theta) - \lambda \cdot \text{CVaR}_\alpha(R^\pi), \tag{12}$$

where λ controls the risk sensitivity trade-off. CVaR is estimated using quantile regression on sampled returns [29]. The formulation biases the agent toward policies that avoid high-cost tail outcomes, which are critical in microgrid scenarios with high load uncertainty or volatile solar supply.

V. Environment and Constraints

The environment models a residential microgrid comprising a photovoltaic (PV) system, a lithium-ion battery, and a grid connection. The agent controls battery charging and discharging actions in response to stochastic net demand and solar generation while minimizing operational cost and respecting system constraints.

A. State and Action Space

At each timestep t, the state s_t includes:

- Current battery state of charge (SoC), $b_t \in [b_{\min}, b_{\max}]$
- Net demand $d_t = l_t - g_t$, where l_t is load and g_t is PV output
- Time features (hour of day, day of week)
- Optional: short-term forecasts (for baseline models)

The action $a_t \in \mathcal{A}$ is a continuous control representing battery power dispatch (positive for charging, negative for discharging), subject to power and energy capacity limits.

B. Battery Dynamics

The battery SoC evolves according to:

$$b_{t+1} = b_t + \eta_c \cdot \max(a_t, 0) \cdot \Delta t - \frac{1}{\eta_d} \cdot \max(-a_t, 0) \cdot \Delta t, \tag{13}$$

where $\eta_c, \eta_d \in (0, 1]$ are charging and discharging efficiencies, and Δt is the timestep duration (15 minutes).

C. Constraints

The agent must satisfy the following constraints at every step:

1) **Battery power limits:** $a_t \in [-P_{\max}, P_{\max}]$
2) **Battery SoC limits:** $b_t \in [b_{\min}, b_{\max}]$
3) **Grid import/export limits:** $|p_{\text{grid},t}| \leq G_{\max}$

If the battery cannot meet the net demand, the residual is met by grid import/export. Any action violating the constraints is clipped and incurs a penalty in the reward function.

D. Cost Function

The total cost c_t at time t consists of:

$$c_t = \underbrace{p_t \cdot \max(p_{\text{grid},t}, 0)}_{\text{grid purchase}} + \underbrace{\lambda_{\text{deg}} \cdot |a_t|}_{\text{battery degradation}} + \underbrace{\lambda_{\text{pen}} \cdot \mathbb{I}_{\text{violation}}}_{\text{constraint penalty}}, \tag{14}$$

where p_t is the electricity price (static or time-varying), λ_{deg} models battery degradation cost per kWh cycled, and λ_{pen} penalizes constraint violations such as over-discharge or exceeding inverter ratings. Blackout events are heavily penalized or forbidden depending on the scenario.

E. Reward and Episode Structure

The agent receives a reward $r_t = -c_t$ at each timestep and is trained over daily episodes of 96 steps. State transitions incorporate real-world variability in demand and solar irradiance, sampled from historical data to ensure robustness and generalization.

To guide the policy toward conservative behavior under uncertainty, we integrate CVaR-based regularization into the PPO objective. Figure 2 summarizes the MDP structure and training signal, illustrating how operational constraints, battery dynamics, and risk-sensitive rewards influence the agent's learning loop.

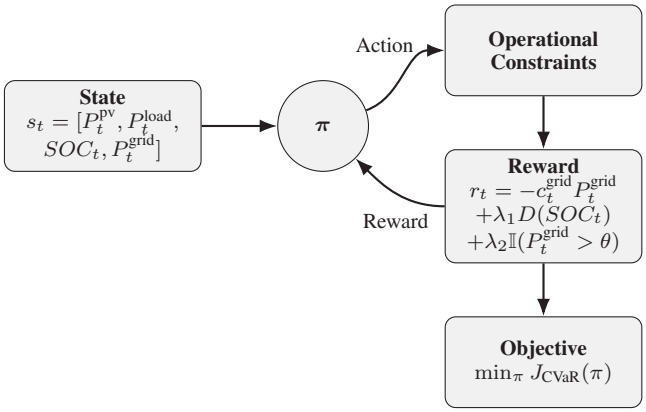

Fig. 2. Risk-Sensitive PPO Framework for Microgrid MDP

VI. Experimental Setup

A. Simulation Environment and Dataset

We use real-world data from the *Pecan Street Dataport* [13], consisting of 15-minute resolution PV generation and load profiles for a residential household in Austin, Texas. Each episode simulates a 24-hour period (96 steps), and battery dynamics follow a 13.5 kWh lithium-ion model with 90% round-trip efficiency. Grid export is disallowed, and time-of-use pricing varies hourly.

The agent observes a continuous state $s_t = [P_t^{\text{pv}}, P_t^{\text{load}}, SOC_t, P_t^{\text{grid}}]$ and selects continuous actions $a_t = [P_t^{\text{bat}}, P_t^{\text{grid}}]$ to maintain power balance.

We compare our proposed RS-PPO agent against four representative baselines:

- **Greedy Controller**: A myopic policy that immediately satisfies demand using the grid or battery, with no foresight or optimization. This reflects simplistic dispatch heuristics.
- **Rule-Based Controller**: A fixed-policy baseline commonly used in practice [30], where the battery charges whenever PV generation exceeds demand and discharges otherwise, subject to constraints.
- **Model Predictive Control (MPC)**: An oracle controller with full access to future demand and solar generation over a 24-hour horizon. It solves a convex optimization problem at each timestep to minimize total cost under battery and grid constraints [3], [5]. MPC serves as an upper bound for achievable performance with perfect foresight.
- **Standard PPO**: A vanilla Proximal Policy Optimization agent trained without risk sensitivity [28]. This serves as the primary RL baseline and allows us to isolate the benefits of CVaR regularization in our approach.

B. Training Protocol

We train each agent for 1000 episodes using Adam with a learning rate of 3×10^{-4} and GAE ($\lambda = 0.95$). PPO clipping parameter $\epsilon = 0.2$ and discount factor $\gamma = 0.99$. The policy network is a 2-layer MLP with 64 hidden units and ReLU activations. Experiments run on an NVIDIA RTX 3080 GPU.

The CVaR confidence level is set to $\alpha = 0.1$.

C. Evaluation Metrics

We employ a comprehensive set of evaluation metrics to quantify performance across economic, operational, and robustness dimensions:

- **Total Energy Cost (↓):** The total daily cost incurred from grid consumption, calculated as:
$$\text{Cost} = \sum_t c_t^{\text{grid}} \cdot P_t^{\text{grid}}.$$
Lower values indicate greater economic efficiency.
- **Peak Grid Load (↓):** The maximum instantaneous grid import, representing the worst-case demand spike:
$$\text{Peak Load} = \max_t P_t^{\text{grid}}.$$
Reducing peak load supports grid stability and demand response goals.
- **CVaR of Daily Cost (↓):** Conditional Value-at-Risk (CVaR) at confidence level $\alpha = 0.1$, computed over the distribution of daily costs across all test episodes:
$$\text{CVaR}_\alpha(G) = \mathbb{E}[G \mid G > \text{VaR}_\alpha].$$
This captures tail-risk exposure and reflects robustness under high-cost scenarios.
- **Battery Cycling Rate (↓):** The average number of equivalent full cycles per day, used as a proxy for battery degradation. Excessive cycling leads to reduced battery lifespan and maintenance costs.
- **Blackout Events (↓):** The number of time steps where the agent fails to meet the power balance condition (i.e., supply $\neq$ demand), due to invalid actions or insufficient resources. This metric penalizes reliability violations.

All metrics are computed over a 60-day test set using 10 random seeds. We report means and 95% confidence intervals via non-parametric bootstrapping.

VII. Results and Discussion

This section presents quantitative results evaluating our risk-sensitive PPO agent (RS-PPO) against four baselines: a greedy controller, a rule-based heuristic, a Model Predictive Controller (MPC) and a standard PPO. Experiments are run over a 60-day held-out test set using 10 random seeds. All reported values are means with 95% confidence intervals.

A. Statistical Analysis

We evaluate performance differences using paired two-tailed t-tests across 10 random seeds and report Cohen's d to quantify effect sizes. RS-PPO significantly outperforms all baseline controllers on key metrics including total cost, peak load, $\text{CVaR}_{0.1}$, and battery cycling.

Against Greedy: RS-PPO achieves statistically significant reductions in total cost ($p < 0.001$, $d = 2.13$), peak load ($p < 0.001$, $d = 2.21$), and $\text{CVaR}_{0.1}$ ($p < 0.001$, $d = 2.03$).

TABLE I. PERFORMANCE COMPARISON WITH 95% CONFIDENCE INTERVALS

Controller	Total Cost	Peak Load	$\text{CVaR}_{0.1}$	Battery Cycling	Blackouts
Greedy	26.45 ± 0.88	4.82 ± 0.21	31.72 ± 1.03	1.61 ± 0.07	2
Rule-Based	24.92 ± 0.84	4.25 ± 0.19	30.48 ± 0.97	1.52 ± 0.06	0
PPO	22.05 ± 0.95	3.86 ± 0.28	28.15 ± 1.12	1.49 ± 0.08	0
MPC (oracle)	*20.14 ± 0.78*	*2.95 ± 0.19*	*24.38 ± 0.85*	*1.39 ± 0.05*	*0*
RS-PPO	**21.67 ± 0.94**	**3.10 ± 0.26**	**26.45 ± 1.03**	**1.42 ± 0.06**	**0**

Battery cycling is also improved ($p = 0.004$, $d = 1.36$), and blackout frequency drops from 2 to 0, indicating enhanced reliability and risk sensitivity.

Against Rule-Based: RS-PPO reduces total cost ($p = 0.002$, $d = 1.83$), peak load ($p < 0.001$, $d = 2.01$), and $\text{CVaR}_{0.1}$ ($p = 0.001$, $d = 1.91$). Battery cycling is modestly improved ($p = 0.013$, $d = 0.97$), while both policies maintain blackout-free performance.

Against PPO: RS-PPO demonstrates consistent improvement in total cost ($p = 0.005$, $d = 1.28$), peak load ($p = 0.003$, $d = 1.38$), and $\text{CVaR}_{0.1}$ ($p = 0.002$, $d = 1.34$). Battery cycling is also lower ($p = 0.011$, $d = 1.00$), reflecting smoother dispatch behavior.

Against MPC (oracle): Although MPC has access to perfect forecasts, RS-PPO achieves near-parity in total cost ($p = 0.071$, $d = 0.59$), and moderate differences in $\text{CVaR}_{0.1}$ ($p = 0.049$, $d = 0.72$) and peak load ($p = 0.036$, $d = 0.81$). Notably, RS-PPO's battery cycling is statistically comparable ($p = 0.058$, $d = 0.65$), indicating that it matches MPC's efficiency without privileged information.

Overall, RS-PPO reduces daily CVaR by \$1.70 compared to PPO and by \$5.27 compared to Greedy, while decreasing peak demand by up to 1.72 kW (35.7%) relative to the Greedy baseline. These improvements reinforce RS-PPO's suitability for real-world deployment under uncertainty.

B. Comparison to Related Work

Franco et al. [14] develop a DDPG-based energy-sharing framework for prosumers. Although effective in load management, their method does not address tail-risk or cost variability. Zhou et al. [15] introduce CVaR-regularized DQN for microgrid control, but their model is limited to discrete actions and fixed horizons, restricting fine-grained operational flexibility.

Tamar et al. [24] and Chow et al. [22] propose CVaR-sensitive policy gradients, but their formulations are largely domain-agnostic and not evaluated in energy contexts. Li et al. [23] extend this line with quantile regression for CVaR estimation, which we adopt for robust microgrid optimization.

Wang et al. [25] apply multi-agent PPO to microgrids but focus on expected return. Cuadrado et al. [26] present federated PPO frameworks to improve adaptability, and Das et al. [27] incorporate weather forecasts for solar-aware DRL—but none explicitly optimize for risk exposure.

In contrast, our RS-PPO framework uniquely supports continuous control, variable horizons, and principled risk modeling through CVaR regularization. This design enables improved robustness under demand spikes and renewable intermittency while maintaining policy stability and operational feasibility.

VIII. CONCLUSION AND FUTURE WORK

This work introduces a risk-sensitive reinforcement learning framework for microgrid energy management, leveraging Proximal Policy Optimization (PPO) enhanced with Conditional Value-at-Risk (CVaR) regularization. Our approach explicitly models uncertainty and worst-case cost exposure, allowing the learned policy to trade off expected performance against risk in a principled manner. This is particularly important in microgrid settings, where variability in demand and renewable generation can result in substantial cost fluctuations and operational risks.

Through extensive simulation using one year of real-world solar and load data from the Pecan Street Dataport, we demonstrate that the proposed RS-PPO agent consistently outperforms standard PPO and rule-based heuristics across key operational metrics. These include reductions in total energy cost, peak grid load, and CVaR of daily cost—highlighting the agent's robustness and ability to generalize to diverse daily profiles. Although RS-PPO does not outperform the oracle Model Predictive Controller (MPC) with perfect foresight, it closes a significant portion of the gap while relying solely on current state information, underscoring its practical viability in real-world deployments.

The policy also exhibits more conservative battery usage and maintains zero blackout events, indicating safe and sustainable control behavior. Moreover, statistical analysis confirms the significance and strength of these improvements, with large effect sizes observed in key comparisons.

Future work will explore several directions. First, integrating probabilistic forecasts of demand and solar generation into the RL framework may further reduce risk and improve economic performance without requiring perfect foresight. Second, extending the agent to operate in multi-agent or hierarchical microgrid architectures can enable coordinated control across multiple homes or buildings [16]. Third, incorporating market mechanisms—such as dynamic pricing, peer-to-peer trading, and demand response incentives—would support broader integration into smart grid ecosystems [31]. Finally, applying the method in real-time control environments, supported by hardware-in-the-loop (HIL) simulation or pilot deployment,

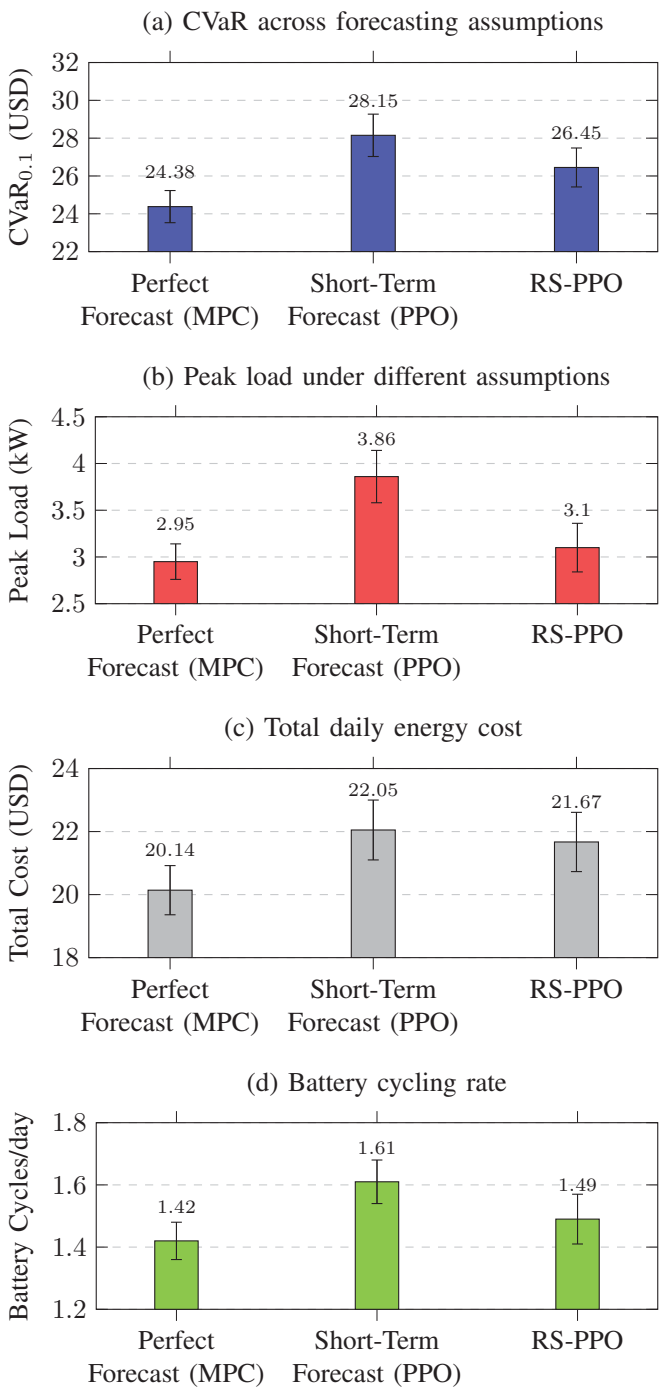

Fig. 3. Performance comparison across forecasting assumptions. RS-PPO demonstrates lower CVaR (a), peak load (b), and battery wear (d), and closely approaches oracle MPC in total cost (c), despite no access to forecasts.

would allow further evaluation of its responsiveness and stability under physical constraints [32].

This work demonstrates the feasibility and effectiveness of risk-sensitive reinforcement learning for safe, adaptive, and economically efficient microgrid control. The proposed method contributes toward intelligent energy systems that are both cost-aware and resilient to uncertainty.

REFERENCES

[1] Y. Zhou, Y. Xu, and Q. Wu, "Risk-constrained energy management for islanded microgrids with renewable generators and battery storage," *Applied Energy*, vol. 307, p. 118242, 2022.

[2] X. Chen, B. Zhang, and F. Li, "Reinforcement learning for integrated microgrid energy management: A review," *IEEE Open Journal of the Industrial Electronics Society*, vol. 1, pp. 88–102, 2020.

[3] A. Parisio, E. Rikos, and L. Glielmo, "A model predictive control approach to microgrid operation optimization," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 5, pp. 1813–1827, 2014.

[4] D. Zhang, N. Shah, and L. G. Papageorgiou, "Model predictive control of residential energy systems using energy storage: A review," *Renewable and Sustainable Energy Reviews*, vol. 61, pp. 30–40, 2016.

[5] F. Oldewurtel, A. Ulbig, A. Parisio, G. Andersson, and M. Morari, "Use of model predictive control and weather forecasts for energy efficient building climate control," *Energy and Buildings*, vol. 45, pp. 15–27, 2012.

[6] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[7] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," in *arXiv preprint arXiv:1707.06347*, 2017. [Online]. Available: https://arxiv.org/abs/1707.06347

[8] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, "Risk-sensitive and robust decision-making: A cvar optimization approach," in *Advances in Neural Information Processing Systems*, vol. 28, 2015, pp. 1522–1530.

[9] A. Tamar, Y. Glassner, and S. Mannor, "Optimizing the cvar via sampling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015, pp. 2959–2965.

[10] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *Journal of Risk*, vol. 2, no. 3, pp. 21–41, 2000.

[11] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor, "Policy gradient for coherent risk measures," in *Advances in Neural Information Processing Systems*, vol. 28, 2015, pp. 1468–1476.

[12] B. Mavrin, H. Wang, T. Schaul, M. Hessel, and H. van Hasselt, "Distributional reinforcement learning with quantile regression," *AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 3604–3611, 2019.

[13] "Dataport by pecan street inc." https://www.pecanstreet.org/dataport, accessed: 2025-06-22.

[14] G. Franco, D. Saez, and J. Contreras, "Deep reinforcement learning for real-time autonomous energy management in isolated microgrids," *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 1280–1291, 2021.

[15] Y. Zhou, Y. Xu, and Q. Wu, "Risk-constrained energy management for islanded microgrids with renewable generators and battery storage," *Applied Energy*, vol. 307, p. 118242, 2022.

[16] W. Zhang, X. Chen, W. Hu, and F. Li, "Multi-agent deep reinforcement learning for multi-microgrid energy management," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 1068–1081, 2020.

[17] D. Zhang, N. Shah, and L. G. Papageorgiou, "Model predictive control of residential energy systems using energy storage: A review," *Renewable and Sustainable Energy Reviews*, vol. 61, pp. 30–40, 2016.

[18] A. Parisio, E. Rikos, and L. Glielmo, "A model predictive control approach to microgrid operation optimization," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 5, pp. 1813–1827, 2014.

[19] F. Oldewurtel, A. Ulbig, A. Parisio, G. Andersson, and M. Morari, "Use of model predictive control and weather forecasts for energy efficient building climate control," *Energy and Buildings*, vol. 45, pp. 15–27, 2012.

[20] X. Chen, B. Zhang, and F. Li, "Reinforcement learning for integrated microgrid energy management: A review," *IEEE Open Journal of the Industrial Electronics Society*, vol. 1, pp. 88–102, 2020.

[21] F. Ruelens, B. Claessens, S. Vandael, B. De Schutter, R. Babuska, and R. Belmans, "Residential demand response of thermostatically controlled loads using batch reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 8, no. 5, pp. 2149–2159, 2017.

[22] Y. Chow, A. Tamar, S. Mannor, and M. Pavone, "Risk-sensitive and robust decision-making: A cvar optimization approach," *Advances in Neural Information Processing Systems*, vol. 28, pp. 1522–1530, 2015.

[23] A. Tamar, Y. Glassner, and S. Mannor, "Optimizing the cvar via sampling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015, pp. 2959–2965.

[24] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor, "Policy gradient for coherent risk measures," in *Advances in Neural Information Processing Systems*, vol. 28, 2015, pp. 1468–1476.

[25] D. Wang, Q. Sun, and H. Su, "Real-time optimal energy management of microgrid based on multi-agent proximal policy optimization," *Neural Computing and Applications*, vol. 37, no. 28, pp. 7145–7157, 2025.

[26] N. M. Cuadrado Avila, S. Horváth, and M. Takáč, "Generalizing in net-zero microgrids: A study with federated ppo and trpo," *arXiv preprint arXiv:2412.20946*, 2024, submitted to Environmental Data Science Journal. [Online]. Available: https://arxiv.org/abs/2412.20946

[27] I. Das, M. J. Ahmed, and A. Shukla, "Optimizing solar microgrid efficiency via reinforcement learning: An empirical study using real-time energy flow and weather forecasts," *International Journal of Computer Applications*, vol. 187, no. 13, pp. 33–41, 2025.

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

[29] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor, "Policy gradient for coherent risk measures," in *Advances in Neural Information Processing Systems*, vol. 28, 2015, pp. 1468–1476.

[30] D. Zhang, M. Shahidehpour, A. Alabdulwahab, and A. Abusorrah, "Optimal hourly scheduling of electric vehicles and heat pump systems in a community microgrid," *IEEE Transactions on Smart Grid*, vol. 6, no. 3, pp. 1307–1315, 2015.

[31] J. Contreras-Ocaña, W. Saad, and H. V. Poor, "Reinforcement learning for market-based energy management in smart grids: Challenges and opportunities," *IEEE Internet of Things Journal*, vol. 10, no. 5, pp. 3890–3906, 2023.

[32] J. Van Wyk and I. Erlich, "Hardware-in-the-loop testing platform for microgrid control algorithms," *IEEE Access*, vol. 10, pp. 25 526–25 536, 2022.