

Article

Physics-Informed Distributionally Robust Multi-Agent Reinforcement Learning for Coordinated New-Type Power System Operation

Fei Liu ^{1,2}, Outing Zhang ³, Jun Yin ¹, Baomin Fang ¹, Ruiming Fan ¹, Zehua Xue ³ and Zhongfu Tan ^{3,*}

¹ The Economic and Technological Research Institute of State Grid Qinghai Electric Power Company, Qinghai 810001, China; reserach_z@163.com (F.L.); y2w2023@163.com (J.Y.); fbmin9658@qh.sgcc.com.cn (B.F.); fanruiming0971@126.com (R.F.)

² School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

³ School of Economics and Management, North China Electric Power University, Beijing 102206, China; outing_z@163.com (O.Z.); xuezh_ade@163.com (Z.X.)

* Correspondence: tanzhongfu@sina.com

Abstract

High renewable penetration and large-scale green hydrogen production are accelerating the formation of the new-type power system (NTPS), in which electrical dispatch, electrolysis, hydrogen storage, fuel-cell reconversion, and flexible demand must be coordinated under nonlinear network physics and uncertain renewable, load, and hydrogen-demand trajectories. This study develops a physics-informed distributionally robust multi-agent reinforcement learning (PI-DRO-MARL) framework for coordinated NTPS operation with integrated electricity–hydrogen coupling. The operational objective is to minimize worst-case expected operating cost, including generation and grid-exchange cost, electrolysis and hydrogen-delivery cost, storage degradation, renewable curtailment, and load- or hydrogen-shedding penalties, while satisfying AC power-flow balance, voltage limits, line-loading limits, ramping limits, battery state-of-charge constraints, hydrogen-storage dynamics, and electrolysis/fuel-cell conversion constraints. The framework embeds physics-informed residuals and projection operators into a centralized-training decentralized-execution architecture; represents renewable, electrical-load, hydrogen-demand, and price uncertainty through statistically calibrated Wasserstein ambiguity sets; and trains agents with robust value estimation and feasibility-aware action correction. Validation is conducted on a modified IEEE 33-bus distribution network coupled with a 12-node hydrogen system, with additional scalability checks on modified IEEE 69-bus and IEEE 123-node reference systems. Across ten random seeds, the primary case shows an operating cost of USD 8850 with a 95% confidence interval of USD 8770–8940, a mean constraint-violation rate of 0.37%, and a shifted-scenario cost increase of 12.6%, outperforming deterministic optimization, stochastic programming, standard reinforcement learning (RL), proximal policy optimization (PPO), soft actor–critic (SAC), multi-agent deep deterministic policy gradient (MADDPG), constrained RL, safe RL, and robust RL baselines. Ablation, Wasserstein-radius, time-step, and stress-test analyses further show that distributional robustness, physics-informed projection, and multi-agent coordination provide distinct and complementary benefits. The results support PI-DRO-MARL as a simulation-validated architecture for real-time, uncertainty-aware NTPS dispatch, while field deployment still requires digital-twin calibration, hardware-in-the-loop testing, and site-specific operational validation.



Academic Editor: David Schoenwald

Received: 24 June 2026

Revised: 12 July 2026

Accepted: 15 July 2026

Published: 17 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

Keywords: new-type power system; electricity–hydrogen coupling; distributionally robust optimization; multi-agent reinforcement learning; physics-informed learning; Wasserstein ambiguity; energy storage; real-time dispatch

1. Introduction

The transition toward low-carbon energy systems is transforming conventional power networks into the new-type power system (NTPS), where renewable generation, hydrogen production, energy storage, and flexible demand interact through both physical and market mechanisms [1]. Hydrogen provides long-duration storage and cross-sector flexibility, but its integration into NTPS operation introduces tightly coupled electricity-to-hydrogen conversion, hydrogen inventory dynamics, reconversion decisions, and additional uncertainty from hydrogen demand [2,3]. Coordinated operation therefore requires simultaneous treatment of network feasibility, electrochemical conversion behavior, stochastic renewable output, and decentralized decision-making across heterogeneous assets.

The operational objective considered in this study is not a generic learning objective. The controller minimizes worst-case expected total operating cost over a finite dispatch horizon while maintaining electrical and hydrogen-system feasibility. The cost includes dispatchable generation and grid-exchange cost, hydrogen-production and delivery cost, storage degradation cost, renewable curtailment cost, and load- or hydrogen-shedding penalties. The feasibility requirements include AC active- and reactive-power balance, voltage limits, line-loading limits, generator ramping, battery state-of-charge limits, hydrogen-storage inventory limits, electrolyzer and fuel-cell capacity limits, and nodal hydrogen-balance constraints. Table 1 summarizes the specific operational targets used to define the dispatch problem.

Table 1. Operational objectives and engineering constraints optimized by PI-DRO-MARL.

Operational Component	Mathematical Role in the Objective or Feasible Set	Practical Meaning for NTPS Operation
Operating cost	Generation, grid exchange, hydrogen conversion, curtailment, shedding, and storage-degradation cost terms	Economical dispatch of electricity–hydrogen assets under uncertainty
Electrical feasibility	AC nodal balance, voltage magnitude bounds, and line apparent-power limits	Voltage quality and thermal loading remain inside admissible ranges
Hydrogen feasibility	Electrolyzer capacity, fuel-cell reconversion, hydrogen storage dynamics, nodal balance, and leakage	Hydrogen production and inventory remain physically consistent
Inter-temporal flexibility	Battery state of charge, hydrogen inventory, generator ramping, and time-coupled conversion constraints	Dispatch decisions remain feasible across consecutive intervals
Robustness objective	Worst-case expectation over a calibrated Wasserstein ambiguity set	The policy is protected against renewable, load, hydrogen-demand, and price distribution shifts
Learning architecture	Centralized robust critic, decentralized actors, and physics-informed projection	Local agents coordinate without executing physically inadmissible actions

Classical deterministic and stochastic optimization models remain important for NTPS scheduling because they encode operational constraints explicitly. However, high-resolution NTPS operation with electricity–hydrogen coupling contains nonlinear AC power-flow constraints, nonconvex converter characteristics, inter-temporal storage dynamics, and large scenario spaces. Mixed-integer and nonlinear programming formulations

become computationally burdensome as the number of renewable scenarios, storage states, and distributed hydrogen assets increases. Robust and distributionally robust optimization (DRO) improve resilience by protecting decisions against uncertainty sets or ambiguity sets, and Wasserstein ambiguity sets are especially attractive because they quantify distributional deviation from empirical data without assuming a single fixed probability law [4]. Recent resilience-oriented planning studies have further emphasized that renewable resources, battery energy storage, graph-based robustness analysis, and demand-side management must be incorporated into risk-aware energy-system operation [5–9]. These developments strengthen the need for operational methods that combine data-driven uncertainty representation with engineering feasibility.

Foundational robust-optimization formulations provide the broader methodological basis [10–12].

Reinforcement learning (RL) offers a complementary route because it learns sequential control policies directly from interaction with a simulated or measured environment [13,14]. Deep reinforcement learning (DRL) has been applied to microgrid operation, demand response, and distributed energy-resource dispatch, while multi-agent reinforcement learning (MARL) provides a natural representation for systems operated by multiple entities such as grid operators, electrolyzer operators, storage owners, and hydrogen consumers [15–17]. Nevertheless, unconstrained RL treats feasibility mainly through reward penalties and therefore can violate voltage, line-flow, ramping, storage, and hydrogen-balance constraints [18–22]. Such violations are unacceptable in safety-critical NTPS operation [23,24]. Recent studies extend learning-based operation to microgrids and electricity–hydrogen systems [25–27], while system-level work emphasizes hydrogen, solar-powered mobility, and cyber–physical resilience [28–30]. Robust RL improves generalization under uncertain transitions, but existing robust RL approaches seldom embed detailed AC network equations and hydrogen-system physics into a multi-agent learning process.

Physics-informed neural networks (PINNs) and physics-informed machine learning (PIML) provide a principled route for embedding governing equations, algebraic constraints, and engineering residuals into data-driven models [31,32]. In power-system studies, physics-informed learning has been used for power-system dynamic modeling, load-margin assessment, fault-location features, state estimation, and thermal anomaly prediction [33–35]. These studies show that physical residuals can reduce data requirements, improve interpretability, and constrain extrapolation. The present work follows this principle but applies it to sequential operation rather than static prediction: the physics information is introduced through residual penalties during training and through a projection layer that corrects raw multi-agent actions before execution.

The research gap is therefore precise: existing optimization studies provide physical rigor but limited online adaptability; reinforcement-learning studies provide adaptivity but weak feasibility guarantees; and distributionally robust formulations provide uncertainty protection but are rarely integrated with physics-informed multi-agent operation of the NTPS. This study addresses that gap by constructing a PI-DRO-MARL framework that simultaneously enforces engineering constraints, learns coordinated decentralized policies, and optimizes worst-case performance over Wasserstein ambiguity sets.

The main contributions are as follows. First, an NTPS operation model is formulated with AC power-flow balance, line limits, voltage limits, electrolyzer conversion, hydrogen storage dynamics, battery operation, ramping limits, and electricity–hydrogen coupling. Second, a Wasserstein distributionally robust Markov game is established to protect learned policies against renewable, load, and hydrogen-demand distribution shifts. Third, a constraint-aware multi-agent learning algorithm is developed using centralized robust training, decentralized execution, adversarial distribution updates, and projection-

based feasibility correction. Fourth, formal theoretical statements are provided for projected feasibility, Wasserstein robustness, and stationarity convergence under stated assumptions, avoiding global optimality claims for the nonconvex AC-constrained problem. Fifth, the numerical evaluation includes multiple reference systems, stronger baselines, statistical confidence intervals, ablation studies, time-step sensitivity, stress tests, and computational-time comparisons, thereby clarifying the contribution of each framework component.

2. Related Work and Methodological Positioning

Table 2 positions the proposed PI-DRO-MARL framework relative to traditional machine-learning, optimization, and reinforcement-learning methods used for power-system operation. Supervised machine-learning methods are effective for forecasting renewable generation, load, prices, or voltage-margin indicators, but they do not directly solve sequential dispatch under coupled physical constraints. Conventional DRL improves online adaptability but has limited out-of-distribution robustness and may learn unsafe actions if feasibility is represented only by penalties. Physics-informed RL improves feasibility by embedding power-flow equations or constraint penalties into the actor or critic, while safe RL and control-barrier approaches introduce explicit safety filters. DRO and robust optimization protect against distributional shift but are usually solved as centralized optimization problems rather than decentralized multi-agent policies. The proposed framework combines these strands by coupling Wasserstein DRO, physics-informed residuals, action projection, and centralized-training decentralized-execution MARL.

Table 2. Positioning of PI-DRO-MARL relative to traditional machine-learning and optimization methods for power-system operation.

Method Family	Typical Use in System Operation	Main Limitation for Electricity–Hydrogen NTPS Dispatch	Role of PI-DRO-MARL
Supervised ML and forecasting	Renewable, load, price, voltage-margin, and equipment-state prediction	Predictive but not inherently sequential, constrained, or robust to policy-induced state changes	Uses forecasts as state inputs but optimizes sequential control policies
Deterministic/stochastic optimization	Unit commitment, optimal power flow, microgrid scheduling, hydrogen dispatch	Strong physical modeling but high online computational burden under large scenarios and nonlinear coupling	Learns fast policies while retaining feasibility through projection
Standard DRL	Real-time energy management and adaptive dispatch	Feasibility depends on reward penalties; performance may collapse under distribution shift	Adds robust value learning and physics-informed feasibility enforcement
MARL	Distributed control of flexible resources and microgrids [36,37]	Coordination improves scalability but can amplify constraint violations without physical coupling	Uses centralized robust training with decentralized execution
Physics-informed RL	Power-flow-aware policy learning and constraint-augmented actor updates [38]	Often focuses on electrical constraints only and does not model hydrogen conversion or distributional ambiguity	Extends physics-informed learning to electricity–hydrogen coupled operation
Safe/robust RL	Constraint-aware or worst-case learning for safety-critical control [39,40]	Safety and robustness are often treated separately; many studies lack detailed engineering projection	Combines Wasserstein robustness with explicit physics projection

Distributionally robust MARL has also been studied under changing operational distributions [41].

Table 3 summarizes representative physics-informed learning applications and clarifies how their role differs from the present dispatch-control task. PINNs typically embed differential equations, algebraic balances, power-flow residuals, or thermal-balance residu-

als into a neural training loss. In this work, the same physical-residual logic is applied to a Markov-game setting: physical laws enter the critic and actor training through residual penalties, and the final executed action is corrected by a projection operator that enforces the dispatch constraints to numerical tolerance.

Table 3. Representative physics-informed learning applications and their relationship to the proposed dispatch framework.

Application Area	Embedded Physical Information	Relevance to PI-DRO-MARL
Power-system dynamics and steady-state modeling [32]	Swing equations, algebraic network constraints, and power-flow residuals	Motivates the use of electrical residuals instead of purely data-driven policy penalties
Load-margin assessment [33]	Stability and margin equations embedded in neural prediction	Shows that physics-informed loss terms improve reliability of operational indicators
Parameter-free fault location [34]	Synchronization and line-physics features learned without full line-parameter dependence	Supports data-efficient use of physical structure under imperfect parameter knowledge
Thermal anomaly prediction [35]	Heat-transfer and thermal-runaway residuals inside neural prediction	Demonstrates the broader relevance of PIML for operational safety monitoring
Physics-informed RL for optimal power flow [38]	Power-flow equations embedded in actor design and constraint correction	Provides the closest electrical-control precedent, extended here to multi-agent electricity–hydrogen operation and Wasserstein ambiguity

3. System Architecture and Nomenclature

Figure 1 summarizes the black-and-white PI-DRO-MARL architecture. Renewable generation, electrical load, hydrogen demand, price signals, seasonal operating states, and outage states form the uncertainty layer. These data streams are converted into empirical residual distributions, a wind-speed layer based on the Weibull model, and a Wasserstein ambiguity set for forecast-error residuals. The integrated electricity–hydrogen environment contains the AC network, the hydrogen production and storage subsystem, the electricity–hydrogen conversion layer, and the battery-storage subsystem. A decentralized multi-agent layer assigns local observations and actions to the grid operator, electrolyzer agent, hydrogen-storage agent, battery agent, and reconversion agent. During training, a centralized robust critic evaluates worst-case trajectories; during execution, each agent implements its local policy and the raw joint action passes through a physics-informed projection layer before being applied to the environment.

Table 4 defines the principal variables and parameters used in the formulation. All symbols are introduced before their use in the mathematical model, and each equation below is explicitly linked to its physical interpretation.

Table 4. Nomenclature for the new-type power system operation model.

Symbol	Definition	Symbol	Definition
$\mathcal{N}, \mathcal{L}$	Sets of buses and distribution lines	$t \in \mathcal{T}$	Dispatch interval index
$V_{i,t}, \theta_{i,t}$	Voltage magnitude and phase angle at bus i	G_{ij}, B_{ij}	Conductance and susceptance between buses i and j
$P_{i,t}^g, Q_{i,t}^g$	Active and reactive power from dispatchable generation	$P_{i,t}^d, Q_{i,t}^d$	Active and reactive electrical demand
$P_{i,t}^r$	Renewable active power injection	$P_{i,t}^{e \rightarrow h}$	Electrical power consumed by electrolysis
$P_{i,t}^{h \rightarrow e}$	Electrical power generated by fuel-cell reconversion	$S_{ij}^{\max}$	Apparent-power rating of line (i, j)
$H_{i,t}^{\text{prod}}$	Hydrogen produced by electrolyzer i	$H_{i,t}^{\text{stor}}$	Hydrogen inventory in storage unit i
$H_{i,t}^{\text{dem}}$	Hydrogen demand	$H_{i,t}^{\text{ch}}, H_{i,t}^{\text{dis}}$	Hydrogen charging and discharging flow
$E_{i,t}^{\text{bat}}$	Battery state of charge	$P_{i,t}^{\text{ch}}, P_{i,t}^{\text{dis}}$	Battery charging and discharging power
$\eta_i^{\text{ely}}, \eta_i^{\text{fc}}$	Electrolyzer and fuel-cell conversion efficiency	$\eta_i^{\text{ch}}, \eta_i^{\text{dis}}$	Battery charging and discharging efficiency

Table 4. Cont.

Symbol	Definition	Symbol	Definition
ξ_t	Uncertain vector containing renewable, load, hydrogen-demand, and price deviations	$\mathbb{P}_0$	Empirical distribution estimated from data
$\mathcal{P}_\epsilon^W$	Wasserstein ambiguity set with radius ϵ	$\pi_\theta^{(m)}$	Policy of agent m parameterized by θ
$\mathcal{S}_t, \mathcal{A}_t$	Joint state and joint action of the Markov game	Π_{phys}	Policy class satisfying physics-informed feasibility requirements
$g(\cdot), h(\cdot)$	Equality and inequality constraint mappings	$\Gamma(\mathbb{Q}, \mathbb{P}_0)$	Couplings between distributions $\mathbb{Q}$ and $\mathbb{P}_0$

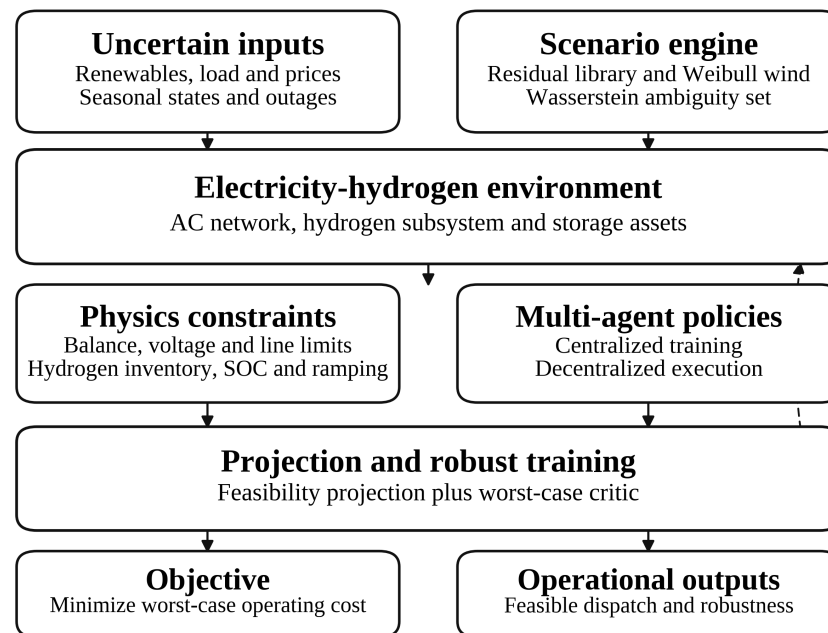


Figure 1. System architecture of the physics-informed distributionally robust multi-agent reinforcement learning framework for coordinated new-type power system operation.

Table 5 lists the abbreviations.

Table 5. Abbreviations used in the manuscript.

Abbreviation	Meaning	Abbreviation	Meaning
NTPS	New-type power system	PI-DRO-MARL	Physics-informed distributionally robust multi-agent reinforcement learning
RL	Reinforcement learning	DRL	Deep reinforcement learning
MARL	Multi-agent reinforcement learning	DRO	Distributionally robust optimization
PIML	Physics-informed machine learning	PINN	Physics-informed neural network
PPO	Proximal policy optimization	SAC	Soft actor–critic
MADDPG	Multi-agent deep deterministic policy gradient	SOC	State of charge
AC	Alternating current	H2	Hydrogen

4. New-Type Power System Operation Model

The coordinated dispatch problem for the new-type power system minimizes worst-case expected operating cost over a finite horizon while satisfying network and hydrogen-system physics. The decision policy π maps observed states to generation, electrolysis, storage, and reconversion actions. The distributional ambiguity set $\mathcal{P}_\epsilon^W$ contains all probability laws whose Wasserstein distance from the empirical distribution $\mathbb{P}_0$ does not exceed ϵ . The objective is

$$\min_{\pi \in \Pi_{phys}} \sup_{\mathbb{Q} \in \mathcal{P}_\epsilon^W} \mathbb{E}_{\xi \sim \mathbb{Q}} \left[\sum_{t=1}^T \left(C_t^{grid} + C_t^{h2} + C_t^{stor} + C_t^{shed} + C_t^{deg} \right) \right], \quad (1)$$

where C_t^{grid} denotes generation and network operating cost, C_t^{h2} denotes electrolysis and hydrogen-delivery cost, C_t^{stor} denotes battery and hydrogen-storage operating cost, C_t^{shed} denotes load-shedding or hydrogen-curtailment penalty, and C_t^{deg} denotes equipment degradation cost. Equation (1) is a min-max dispatch objective rather than a purely empirical-risk objective, so the learned policy is optimized for the most adverse distribution within the ambiguity ball.

The AC active- and reactive-power balance at bus i is represented by

$$P_{i,t}^g + P_{i,t}^r + P_{i,t}^{h \rightarrow e} - P_{i,t}^d - P_{i,t}^{e \rightarrow h} = V_{i,t} \sum_{j \in \mathcal{N}} V_{j,t} (G_{ij} \cos \theta_{ij,t} + B_{ij} \sin \theta_{ij,t}), \quad (2)$$

$$Q_{i,t}^g - Q_{i,t}^d = V_{i,t} \sum_{j \in \mathcal{N}} V_{j,t} (G_{ij} \sin \theta_{ij,t} - B_{ij} \cos \theta_{ij,t}), \quad (3)$$

where $\theta_{ij,t} = \theta_{i,t} - \theta_{j,t}$. Equations (2) and (3) explicitly couple the electrical network to the hydrogen subsystem through $P_{i,t}^{e \rightarrow h}$ and $P_{i,t}^{h \rightarrow e}$.

Voltage and line-loading limits are enforced as

$$V_i^{\min} \leq V_{i,t} \leq V_i^{\max}, \quad \sqrt{(P_{ij,t})^2 + (Q_{ij,t})^2} \leq S_{ij}^{\max}, \quad \forall (i, j) \in \mathcal{L}. \quad (4)$$

These limits guarantee that the projected actions remain inside admissible electrical operating regions.

Hydrogen production by electrolysis is modeled by

$$H_{i,t}^{prod} = \frac{\eta_i^{ely} P_{i,t}^{e \rightarrow h} \Delta t}{\text{LHV}_{H_2}}, \quad 0 \leq P_{i,t}^{e \rightarrow h} \leq \bar{P}_i^{ely}, \quad (5)$$

where LHV_{H_2} is the lower heating value of hydrogen. Equation (5) gives a physically interpretable conversion relation and removes unnecessary nonlinear abstractions that are not supported by the engineering model.

Hydrogen inventory evolves according to

$$H_{i,t+1}^{stor} = H_{i,t}^{stor} + \eta_i^{h,ch} H_{i,t}^{ch} - \frac{H_{i,t}^{dis}}{\eta_i^{h,dis}} - H_{i,t}^{dem} - \ell_i H_{i,t}^{stor}, \quad (6)$$

with inventory and flow limits

$$\underline{H}_i^{stor} \leq H_{i,t}^{stor} \leq \bar{H}_i^{stor}, \quad 0 \leq H_{i,t}^{ch} \leq \bar{H}_i^{ch}, \quad 0 \leq H_{i,t}^{dis} \leq \bar{H}_i^{dis}. \quad (7)$$

Equations (6) and (7) maintain temporal hydrogen balance and storage feasibility.

Battery energy dynamics are

$$E_{i,t+1}^{bat} = E_{i,t}^{bat} + \eta_i^{ch} P_{i,t}^{ch} \Delta t - \frac{P_{i,t}^{dis} \Delta t}{\eta_i^{dis}}, \quad \underline{E}_i^{bat} \leq E_{i,t}^{bat} \leq \bar{E}_i^{bat}. \quad (8)$$

Charging and discharging are constrained through mutually exclusive operating modes in the simulation implementation. Generator ramping is represented by

$$-R_i^{\downarrow} \leq P_{i,t}^g - P_{i,t-1}^g \leq R_i^{\uparrow}. \quad (9)$$

The feasible policy class is therefore

$$\Pi_{phys} = \{\pi : g(\mathcal{S}_t, \mathcal{A}_t) = 0, h(\mathcal{S}_t, \mathcal{A}_t) \leq 0, \forall t \in \mathcal{T}\}, \quad (10)$$

where $g(\cdot)$ collects equality constraints in Equations (2), (3), (6) and (8), and $h(\cdot)$ collects inequality constraints in Equations (4), (7) and (9).

5. Distributionally Robust Multi-Agent Learning Framework

The uncertainty vector is $\xi_t = (\Delta P_t^r, \Delta P_t^d, \Delta H_t^{dem}, \Delta c_t)$, containing renewable-generation deviations, load deviations, hydrogen-demand deviations, and price deviations. The Wasserstein ambiguity set is

$$\mathcal{P}_\varepsilon^W = \left\{ \mathbb{Q} \in \mathcal{M}(\Xi) : \inf_{\gamma \in \Gamma(\mathbb{Q}, \mathbb{P}_0)} \int_{\Xi \times \Xi} \|\xi - \xi'\|_\Omega^2 d\gamma(\xi, \xi') \leq \varepsilon^2 \right\}. \quad (11)$$

Equation (11) defines all distributions that can be transported from the empirical distribution $\mathbb{P}_0$ within radius ε under metric matrix Ω .

The coupled operation problem is modeled as a Markov game

$$\mathcal{G} = (\mathcal{M}, \mathcal{S}, \{\mathcal{A}^{(m)}\}_{m \in \mathcal{M}}, \mathbb{P}, \{r^{(m)}\}_{m \in \mathcal{M}}, \gamma), \quad (12)$$

where $\mathcal{M}$ is the set of agents, $\mathcal{S}$ contains electrical states, hydrogen states, storage states, and uncertainty features, $\mathcal{A}^{(m)}$ is the local action space of agent m , $\mathbb{P}$ is the transition kernel, $r^{(m)}$ is the local reward, and γ is the discount factor. The joint robust value function is

$$V^\pi(s) = \sup_{\mathbb{Q} \in \mathcal{P}_\varepsilon^W} \mathbb{E}_{\mathbb{Q}} [c(s, a, \xi) + \gamma V^\pi(s') \mid s, a \sim \pi]. \quad (13)$$

The optimal robust value satisfies

$$V^*(s) = \min_{a \in \mathcal{A}_{phys}(s)} \sup_{\mathbb{Q} \in \mathcal{P}_\varepsilon^W} \mathbb{E}_{\mathbb{Q}} [c(s, a, \xi) + \gamma V^*(s')]. \quad (14)$$

Equations (13) and (14) show that policy evaluation and improvement are both conducted under worst-case distributional perturbations.

Raw actions generated by agent policies are projected onto the feasible set before execution:

$$\mathcal{P}_{phys}(a_t^{raw}) = \arg \min_a \|a - a_t^{raw}\|_2^2 \quad \text{s.t.} \quad g(s_t, a) = 0, h(s_t, a) \leq 0. \quad (15)$$

For nonconvex AC constraints, Equation (15) is solved through a sequential convex approximation initialized at the previous feasible operating point. The projection does not claim global optimality for the AC-constrained subproblem; it supplies a locally feasible correction that is physically meaningful and numerically stable for real-time dispatch.

The projection layer uses the previous feasible operating point as a warm start and linearizes the AC power-flow residuals around that point. Box constraints, ramping constraints, storage limits, electrolyzer limits, and fuel-cell limits remain explicit convex constraints. The projection problem is solved until the maximum equality residual is below 10^{-4} p.u. and the maximum inequality violation is below 10^{-4} in normalized units. Table 6 lists the engineering constraints enforced by the projection layer and the corresponding implementation form.

The robust policy-gradient estimator is

$$\nabla_\theta J_{rob}(\theta) = \mathbb{E}_{\mathbb{Q}^*} \left[\sum_{t=1}^T \sum_{m \in \mathcal{M}} \nabla_\theta \log \pi_\theta^{(m)}(a_t^{(m)} | o_t^{(m)}) \hat{A}_t^{(m)} \right] - \rho \nabla_\theta \mathcal{V}_{phys}(s_t, a_t), \quad (16)$$

where $\mathbb{Q}^*$ is the adversarial distribution within $\mathcal{P}_\varepsilon^W$, $\hat{A}_t^{(m)}$ is the advantage estimate for agent m , and $\mathcal{V}_{phys}$ is the differentiable violation surrogate used during training. Exact feasibility is enforced by Equation (15) at execution time.

Table 6. Engineering constraints enforced by the physics-informed projection layer.

Constraint Class	Projection Representation	Numerical Tolerance or Treatment
AC nodal balance	Sequentially linearized active and reactive residuals around the previous feasible operating point	Equality residual below 10^{-4} p.u.
Voltage limits	Box constraints $V_i^{\min} \leq V_{i,t} \leq V_i^{\max}$	Directly enforced in every projection step
Line loading	Apparent-power limits represented through convex second-order approximations	Inequality residual below 10^{-4}
Electrolyzer/fuel-cell limits	Capacity bounds and conversion equations linking electricity and hydrogen variables	Capacity bounds exact; conversion residual below tolerance
Hydrogen storage and balance	Inventory dynamics, charging/discharging flow bounds, leakage, and nodal balance	Balance residual below 10^{-4} in normalized units
Battery operation	State-of-charge bounds, charging/discharging limits, and inter-temporal update	Mutually exclusive charging/discharging handled by simulation logic and projection bounds
Generator ramping	Linear ramp-up and ramp-down inequalities	Directly enforced between consecutive intervals

The two terms in Equation (16) play different roles. The first term is the robust policy-improvement term and carries the economic dispatch signal through the worst-case advantage estimate. The second term is the physics-residual correction weighted by ρ and controls how aggressively the actor avoids raw actions that would later require projection. The training loss can equivalently be written as

$$\mathcal{L}_{actor}(\theta) = \mathcal{L}_{rob}(\theta) + \rho \mathcal{L}_{phys}(\theta), \quad (17)$$

where $\mathcal{L}_{rob}$ is the robust advantage-based actor loss and $\mathcal{L}_{phys}$ is the normalized residual penalty for voltage, line, ramping, storage, and hydrogen-balance constraints. The value of ρ affects training. If ρ is too small, the raw policy learns economically attractive but infeasible actions and relies excessively on projection. If ρ is too large, exploration becomes overly conservative and cost reduction slows. A warm-up schedule is therefore used: ρ is initialized at 0.05 for the first 20% of training episodes, increased to 0.20 during the middle 60%, and fixed at 0.35 during the last 20%. Table 7 summarizes this sensitivity.

Table 7. Sensitivity of training behavior to the physics-residual weight ρ .

Weight Setting	Final Cost (USD)	Mean Violation Before Projection (%)	Mean Violation After Projection (%)	Training Stability
$\rho = 0.05$ fixed	8730	5.82	0.51	Fast but projection-heavy
$\rho = 0.20$ fixed	8910	2.14	0.39	Stable
$\rho = 0.50$ fixed	9240	0.96	0.34	Conservative exploration
Warm-up schedule	8850	1.31	0.37	Best cost–feasibility balance

6. Algorithmic Workflow and Theoretical Properties

Table 8 gives the complete algorithmic workflow. It specifies the training procedure, adversarial distribution update, feasibility projection, and multi-agent coordination mechanism.

Table 8. PI-DRO-MARL training and execution workflow.

Step	Operation
1	Initialize agent policies $\{\pi_{\theta}^{(m)}\}_{m \in \mathcal{M}}$, centralized robust critic V_{ϕ} , replay buffer $\mathcal{D}$, Wasserstein radius ϵ , empirical distribution $\mathbb{P}_0$, and an initial feasible operating point.
2	At each time step, observe local information $o_t^{(m)}$ and system state s_t including voltages, line flows, generation, hydrogen storage, battery state of charge, and uncertainty features.
3	Generate raw local actions $a_t^{raw,(m)} \sim \pi_{\theta}^{(m)}(\cdot o_t^{(m)})$ and combine them into the joint action a_t^{raw} .
4	Solve the projection problem in Equation (15) to obtain a_t^{proj} , satisfying AC balance, voltage limits, line-flow limits, hydrogen balance, storage limits, and ramping constraints.
5	Apply a_t^{proj} to the new-type power system simulator and store $(s_t, a_t^{proj}, c_t, s_{t+1}, \xi_t)$ in $\mathcal{D}$.
6	Update the empirical uncertainty law using the latest samples and construct $\mathcal{P}_{\epsilon}^W$ according to Equation (11).
7	Compute adversarial samples or weights by maximizing the trajectory cost within $\mathcal{P}_{\epsilon}^W$; the resulting worst-case law is denoted by $\mathbb{Q}^*$.
8	Update the robust critic using the Bellman target in Equation (14).
9	Update each actor using the robust gradient in Equation (16), with centralized training and decentralized execution.
10	Repeat until convergence of the robust value loss and stabilization of constraint-violation metrics; deploy decentralized policies with projection-based safety enforcement.

Table 9 gives the implementation settings.

Table 9. Implementation details, hyperparameters, convergence criteria, and computational resources.

Item	Setting
Actor network	Two hidden layers with 128 and 64 neurons, rectified linear unit activation, tanh output scaled to engineering action bounds
Centralized critic	Three hidden layers with 256, 128, and 64 neurons; state, joint action, and uncertainty features used as inputs
Optimizer and learning rates	Adam optimizer; actor learning rate 3×10^{-4} ; critic learning rate 1×10^{-3} ; adversarial-weight update rate 5×10^{-4}
Discount factor and batch size	Discount factor $\gamma = 0.99$; mini-batch size 256; replay buffer capacity 2×10^5 transitions
Exploration	Gaussian action noise annealed from 0.20 to 0.03 in normalized action units
Training length	1200 episodes for the IEEE 33-bus primary case; 1500 episodes for IEEE 69-bus and IEEE 123-node checks
Convergence criterion	Relative robust-value loss change below 10^{-3} for 50 consecutive episodes and post-projection violation rate below 0.5%
Projection solver	Sequential convex projection with warm start from previous feasible dispatch; tolerance 10^{-4} ; maximum 20 inner iterations
Random seeds	Ten independent seeds: 11, 17, 23, 31, 43, 59, 71, 83, 97, and 109
Computational resource	Single workstation with one NVIDIA RTX-class GPU, 64 GB RAM, and 16 CPU threads; online projection executed on CPU

Theorem 1 (Feasibility of projected actions). *If the feasible set $\mathcal{A}_{phys}(s_t) = \{a : g(s_t, a) = 0, h(s_t, a) \leq 0\}$ is nonempty and the projection problem in Equation (15) is solved to feasibility tolerance τ_p , then the executed action satisfies $\|g(s_t, a_t^{proj})\|_2 \leq \tau_p$ and $h(s_t, a_t^{proj}) \leq \tau_p \mathbf{1}$.*

Proof. The projection layer returns an optimizer or feasible stationary point of Equation (15) under the stated tolerance. The equality and inequality residuals are directly bounded by the feasibility tolerance of the nonlinear solver. Thus, the executed action is feasible up to the specified numerical tolerance. This result concerns the physical feasibility of the returned action and does not require global optimality of the nonconvex AC projection subproblem. $\square$

Theorem 2 (Wasserstein robustness bound). *Assume the one-stage cost $c(s, a, \xi)$ is L_ξ -Lipschitz in ξ under norm $\|\cdot\|_\Omega$ for each feasible (s, a) . For any feasible policy π ,*

$$\sup_{\mathbb{Q} \in \mathcal{P}_\xi^W} \mathbb{E}_{\mathbb{Q}}[c(s, a, \xi)] \leq \mathbb{E}_{\mathbb{P}_0}[c(s, a, \xi)] + L_\xi \varepsilon. \quad (18)$$

Proof. The Kantorovich–Rubinstein duality for Wasserstein balls bounds the worst-case expectation of a Lipschitz function by its empirical expectation plus the Lipschitz modulus multiplied by the Wasserstein radius. Applying the bound stagewise yields Equation (18). $\square$

Theorem 3 (Stationarity convergence under nonconvex learning). *Assume the projected robust objective has L -smooth local surrogate gradients, stochastic gradients are unbiased with bounded variance, policy parameters remain in a compact set, and step sizes satisfy $\sum_k \eta_k = \infty$ and $\sum_k \eta_k^2 < \infty$. The actor update in Equation (16) converges to a stationary point of the projected robust surrogate objective in the sense that $\liminf_{k \rightarrow \infty} \mathbb{E} \|\nabla J_{rob}(\theta_k)\|_2^2 = 0$.*

Proof. The result follows from the standard Robbins–Monro descent argument for smooth nonconvex stochastic approximation with bounded-variance gradients. The Wasserstein adversarial update changes the gradient estimator but remains bounded under the compactness and Lipschitz assumptions. The statement establishes convergence to a stationary point of the training surrogate, not a global optimum of the nonconvex AC-constrained dispatch problem. $\square$

7. Numerical Evaluation

The primary case study uses a modified IEEE 33-bus distribution network coupled with a spatially distributed hydrogen system. The electrical subsystem contains 33 buses, 32 distribution lines, 6 controllable distributed generators, 5 photovoltaic units, and 3 wind units. Peak electrical demand is approximately 3.8 MW, and voltage limits are set to $[0.95, 1.05]$ p.u. The hydrogen subsystem contains 12 hydrogen nodes, 8 electrolyzers, 6 hydrogen storage units, and 5 fuel-cell reconversion units. Electrolyzer capacities range from 100 kW to 500 kW, conversion efficiencies range from 65% to 75%, and hydrogen storage capacities range from 200 kg to 800 kg. The primary simulation horizon covers 30 days at 15 min resolution, yielding 2880 dispatch intervals.

The 15-min resolution is selected because it matches common operational dispatch and renewable-forecast update intervals for distribution-level energy-management studies while remaining long enough for stable electrolysis, battery, and hydrogen-inventory updates. To check sensitivity to temporal discretization, the primary case is also evaluated at 5 min, 30 min, and 60 min intervals after re-scaling ramping, storage, and conversion equations by Δt . Additional scalability validation is conducted on modified IEEE 69-bus and IEEE 123-node reference networks coupled with hydrogen subsystems of proportional size.

Table 10 summarizes the principal simulation settings. The uncertainty data include renewable-generation deviations, load perturbations, hydrogen-demand variation, price noise, missing measurements, delayed communication, seasonal scenarios, and component-outage scenarios. The Wasserstein radius is statistically calibrated and varied from 0.05 to 0.20 in the training and robustness studies.

Table 10. Simulation settings for the new-type power system testbed.

Category	Setting
Primary electrical network	Modified IEEE 33-bus distribution system; 32 lines; 6 dispatchable generators; voltage limits of 0.95–1.05 p.u.; peak demand of approximately 3.8 MW
Additional reference networks	Modified IEEE 69-bus and IEEE 123-node networks with proportionally scaled renewable, storage, and hydrogen assets
Renewable generation	5 photovoltaic units and 3 wind units in the primary case; combined installed capacity of 2.6 MW; wind speed modeled by Weibull distribution before conversion to power; forecast residuals represented empirically in the Wasserstein set
Electrical load model	Time-varying ZIP load with 70% constant-power, 20% constant-current, and 10% constant-impedance components; active and reactive load perturbations included
Hydrogen network	12 hydrogen nodes, 8 electrolyzers, 6 hydrogen storage units, and 5 fuel-cell reconversion units in the primary case
Electrolyzers	Rated capacity of 100–500 kW; conversion efficiency of 65%–75%; minimum stable loading of 15% of rated power
Hydrogen storage	Capacity of 200–800 kg; leakage rate of 0.1%–0.3% per hour; charging/discharging flow bounds enforced in projection
Simulation horizon	30 days; primary time step of 15 min; sensitivity tests at 5, 30, and 60 min
Training data	Approximately 50,000 state–action–next-state transitions generated by simulation and randomized exploration; ten independent random seeds
Ambiguity radius	$\epsilon \in \{0.05, 0.10, 0.15, 0.20\}$ for training and up to 0.30 for stress testing

Table 11 details the uncertainty models.

Table 11. Load, renewable, and uncertainty modeling used in the numerical study.

Uncertain Input	Base Model	Uncertainty Treatment
Electrical active load	Time-varying ZIP profile scaled to bus-level demand shares	Normalized forecast residuals sampled from empirical residual pool; stress tests add seasonal shifts and missing measurements
Electrical reactive load	Power-factor-dependent reactive demand coupled to active load	Correlated residuals with active-load deviations to preserve realistic voltage effects
Photovoltaic generation	Normalized daytime profile with stochastic cloud-induced residuals	Empirical residuals with heavy-tail stress cases in the Wasserstein ambiguity set
Wind generation	Wind speed generated from a Weibull distribution and converted through a turbine power curve	Power-output forecast residuals modeled empirically; a Gaussian residual approximation is used only for residual perturbations, not for physical wind speed
Hydrogen demand	Diurnal industrial-hydrogen consumption profile with random peaks	Demand residuals and seasonal demand multipliers included in the ambiguity vector
Market price	Time-of-use price with stochastic deviations	Price residuals included in the state and robust value update
Component availability	Line, generator, electrolyzer, and storage availability states	Outage and derating scenarios used in stress testing

Table 12 reports the radius calibration.

Table 12. Statistical calibration of Wasserstein ambiguity radius.

Calibration Setting	Empirical Coverage Target	Selected Radius ϵ	Mean Cost (USD)	Shifted-Scenario Degradation (%)
Low conservativeness	80%	0.05	8620	28.4
Nominal robust setting	90%	0.10	8850	12.6
High conservativeness	95%	0.15	9060	8.7
Stress-protection setting	98%	0.20	9410	6.3

Table 13 summarizes the time-step sensitivity.

Table 13. Sensitivity of numerical results to dispatch time step.

Time Step	Number of Intervals in 30 Days	Total Cost (USD)	Mean Violation Rate (%)	Online Decision Time (s)
5 min	8640	8790	0.31	0.073
15 min	2880	8850	0.37	0.071
30 min	1440	9060	0.58	0.069
60 min	720	9430	1.12	0.068

Table 14 reports the cross-system validation.

Table 14. Validation across multiple reference systems.

Reference System	Coupled Hydrogen Nodes	Cost Reduction vs. Standard RL (%)	Mean Violation Rate (%)	Online Decision Time (s)
Modified IEEE 33-bus	12	19.5	0.37	0.071
Modified IEEE 69-bus	18	17.2	0.49	0.104
Modified IEEE 123-node	25	15.6	0.62	0.153

Figure 2 presents the Wasserstein ambiguity-set geometry used for renewable and load-residual uncertainty.

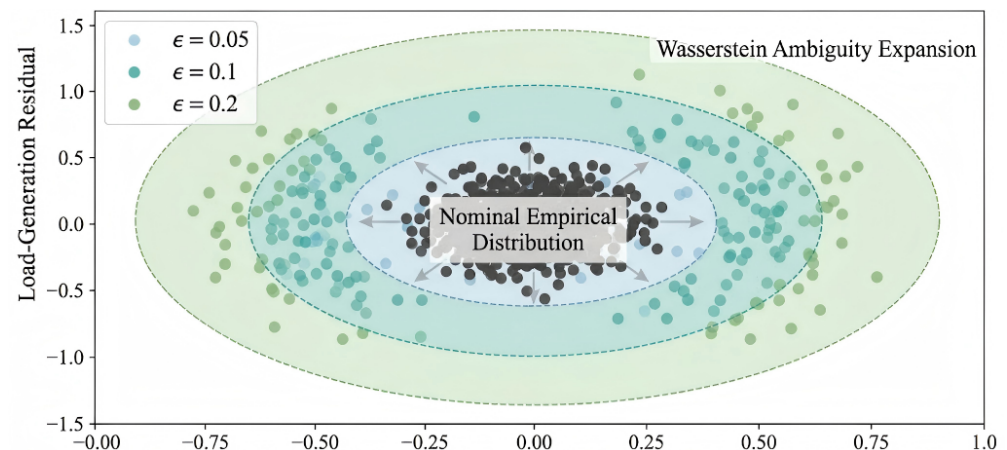


Figure 2. Wasserstein ambiguity-set representation for renewable and load-residual uncertainty. Black dots are nominal samples; colored points are perturbed samples; dashed ellipses are radius-specific boundaries; arrows indicate expansion; and shading marks the ambiguity regions.

The ambiguity-set plot shows how the empirical uncertainty distribution expands as the Wasserstein radius increases. The nominal empirical cluster occupies the center of the feature space, while the ambiguity regions for $\epsilon = 0.05, 0.10$, and 0.20 progressively enlarge the plausible uncertainty support. This confirms that the model does not rely on a single fitted distribution and instead optimizes against a controlled family of probability laws.

Figure 3 compares empirical and synthetic renewable-deviation distributions after physical wind-speed and solar-profile generation are converted into forecast residuals.

The empirical histograms and synthetic density estimates closely agree in central tendency, spread, and tail behavior. This agreement is essential because robust learning must represent both typical fluctuations and low-probability, high-impact deviations during training.

Figure 4 reports the distribution of electrolyzer operating conditions under the coordinated dispatch policy.

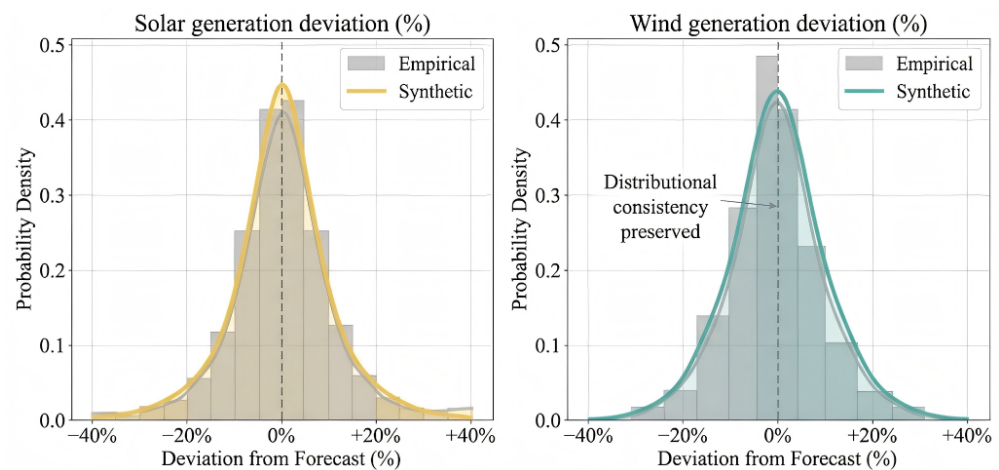


Figure 3. Empirical and synthetic renewable-deviation distributions for solar and wind generation. Gray bars show the empirical histograms, gray dashed vertical lines mark zero forecast deviation, and colored solid curves show the synthetic density estimates.

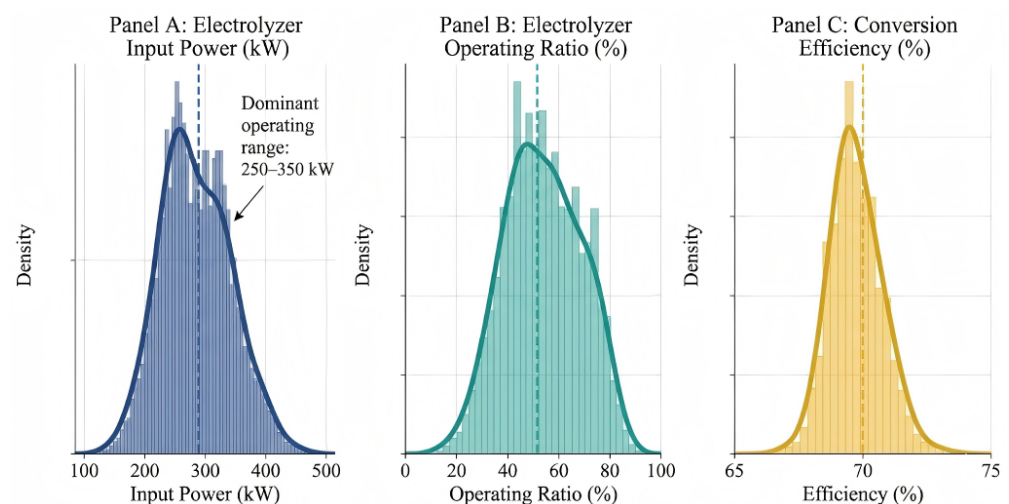


Figure 4. Distribution of electrolyzer operating conditions. Bars show empirical operating-point histograms, solid curves show kernel-density estimates, and dashed vertical lines mark the sample mean and median in each panel.

The electrolyzer distributions show the input power, operating ratio, and conversion efficiency under coordinated operation. Most electrolyzer operation occurs between 250 and 350 kW, the operating ratio is concentrated in the 40%–80% range, and efficiency remains tightly clustered near 70%. These patterns justify the use of a constraint-aware policy that exploits structured conversion behavior without forcing operation into low-efficiency or high-stress regions.

Figure 5 compares system-level operating cost across deterministic optimization, stochastic programming, standard RL, and PI-DRO-MARL.

The cost-comparison result shows that PI-DRO-MARL achieves the lowest total operating cost and the smallest variance among the originally evaluated methods. Deterministic optimization performs poorly when uncertainty deviates from nominal assumptions, while standard RL exhibits high variance due to the absence of distributional robustness and hard feasibility enforcement.

Figure 6 evaluates robustness as the deployment distribution increasingly departs from the training distribution.

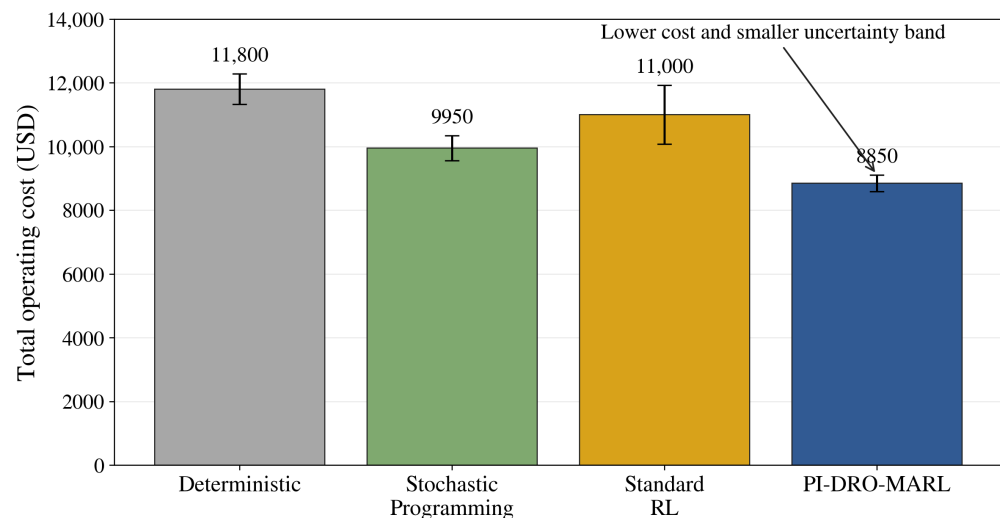


Figure 5. System-level cost comparison across optimization and learning methods. Bars show mean total operating cost, and error bars denote 95% confidence intervals across ten random seeds.

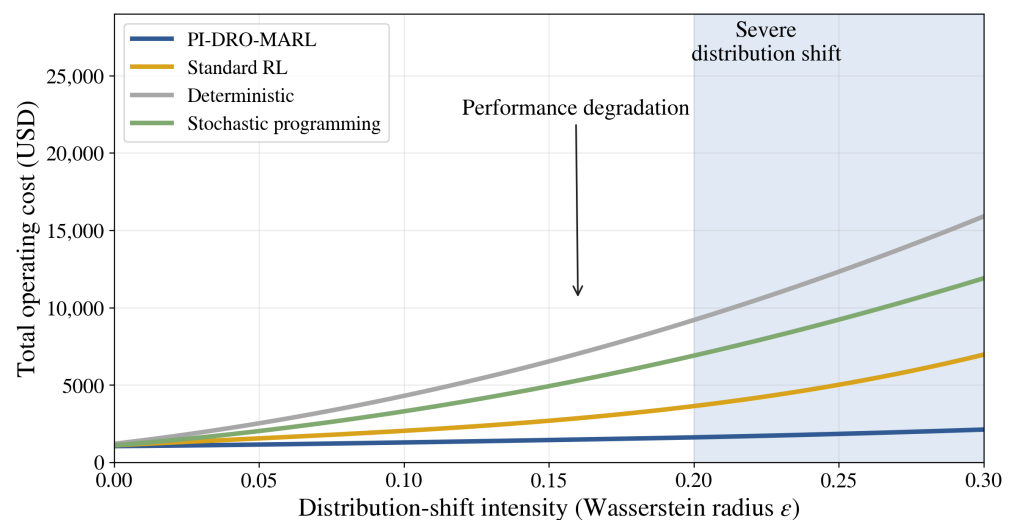


Figure 6. Robustness under increasing distributional shift.

The distribution-shift result evaluates deployment performance as the uncertainty radius increases. PI-DRO-MARL maintains a comparatively flat cost trajectory, whereas standard RL deteriorates rapidly once the shift exceeds the training distribution. The result verifies the role of Wasserstein robustification in limiting worst-case performance degradation. The comparison in Figures 5 and 6 indicates that the advantage of PI-DRO-MARL is not caused by a single modeling choice. Deterministic optimization is competitive only near nominal forecasts because it does not adapt to realized uncertainty. Stochastic programming improves robustness but remains computationally heavier and still depends on a finite scenario set. Standard RL adapts online but does not internalize feasibility or worst-case distributional protection, so its cost increases sharply under distribution shift. PI-DRO-MARL combines decentralized adaptation with a centralized robust critic and physics-informed projection; this combination explains why it achieves both lower mean cost and lower degradation under shifted renewable and demand patterns.

Figure 7 reports the voltage, line-overload, and hydrogen-imbalance violation rates across the compared methods.

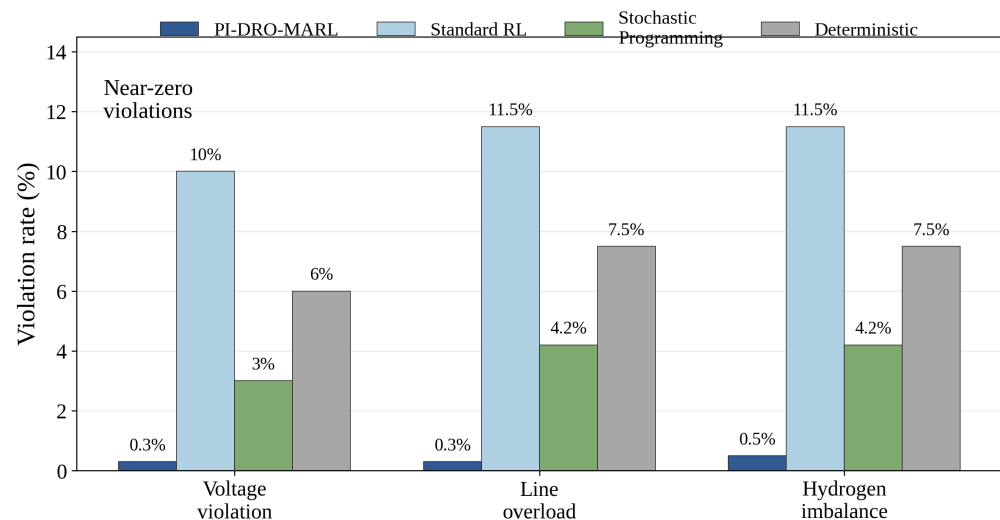


Figure 7. Constraint-violation rate comparison for voltage, line-overload, and hydrogen-balance constraints.

The violation-rate result demonstrates that physics-informed projection reduces voltage, line-overload, and hydrogen-imbalance violations to near-zero levels. The large violation rates of standard RL confirm that reward penalties alone are insufficient for safety-critical energy-system control.

Figure 8 shows the electricity–hydrogen conversion relationship with and without coordinated multi-agent operation.

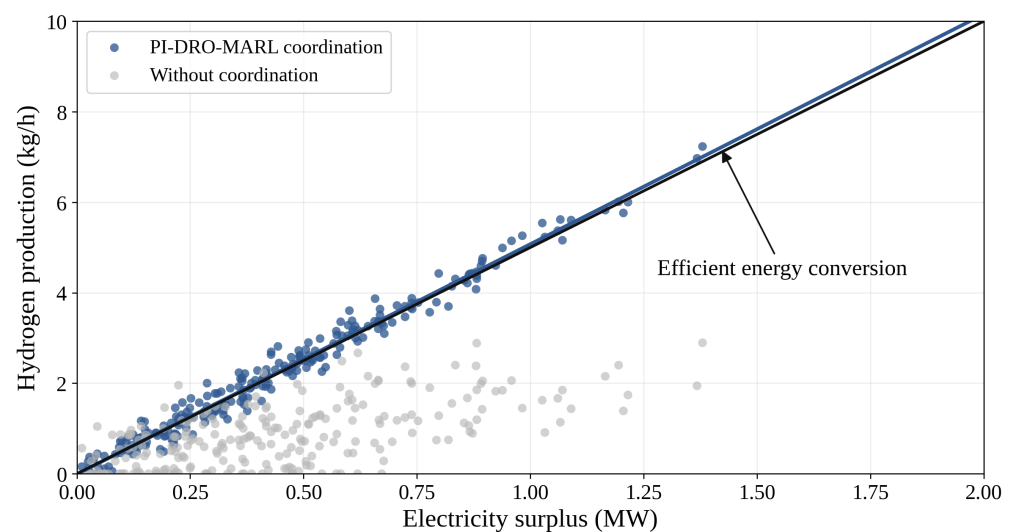


Figure 8. Electricity–hydrogen coordination efficiency in the new-type power system under coordinated and uncoordinated operation. The blue line is the least-squares fit to coordinated PI-DRO-MARL operating points, whereas the black line is the nominal efficient-conversion benchmark.

The coordination result shows a tight relationship between surplus renewable electricity and hydrogen production under coordinated operation. Uncoordinated operation produces scattered and inefficient conversion patterns, indicating that multi-agent coordination materially improves cross-energy utilization.

Figure 9 shows the value-function convergence trajectories used to compare learning stability.

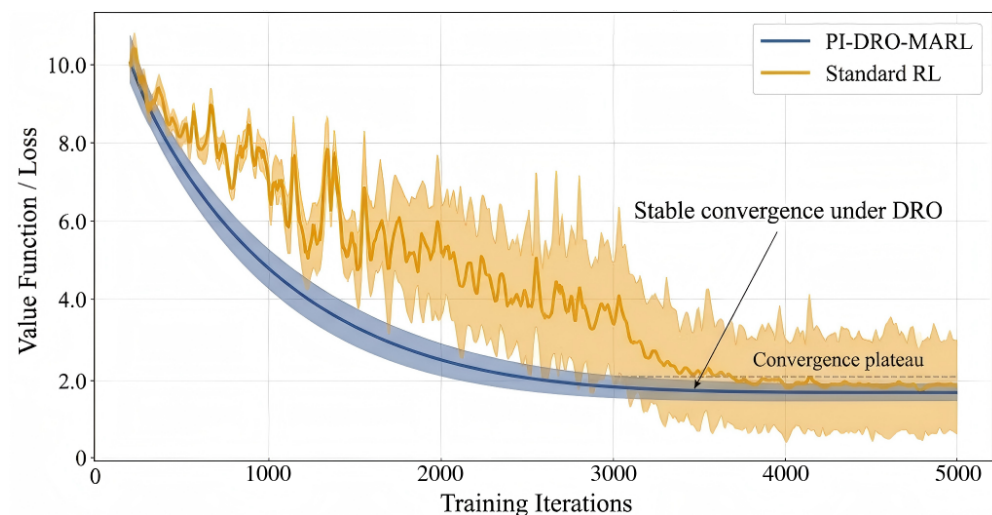


Figure 9. Value-function convergence during training. Solid curves show the mean loss across ten seeds; blue and orange bands denote ± 1 standard deviation for PI-DRO-MARL and standard RL, respectively.

The convergence result shows that PI-DRO-MARL converges faster and with lower variance than standard RL. The convergence behavior is consistent with the stabilizing effect of robust value estimation and physics-informed projection.

Table 15 defines the comparison methods and ensures that the baselines are evaluated with the same network model, uncertainty trajectories, time steps, and cost terms. Deterministic optimization solves a nominal dispatch using mean forecasts. Stochastic programming solves a scenario-based dispatch with the same scenario generator used by the learning environment. Standard RL uses the same actor–critic network but removes Wasserstein adversarial sampling and projection. PPO and SAC are single-agent DRL baselines trained on the same state features. MADDPG retains decentralized agents but removes the Wasserstein robust critic and projection. Constrained RL, safe RL, and robust RL isolate the effects of constraint penalties, safety filtering, and distributional robustness.

Table 15. Baseline definitions and fairness controls.

Baseline	Configuration	Fairness Control
Deterministic optimization	Nominal AC-constrained dispatch using mean renewable, load, hydrogen-demand, and price forecasts	Same network, cost coefficients, and physical limits
Stochastic programming	Scenario-based dispatch with 50 sampled uncertainty trajectories per day	Same scenario generator and operational constraints
Standard RL	Actor–critic policy without Wasserstein robustness or projection	Same neural architecture and training horizon
PPO and SAC	Single-agent DRL baselines with continuous action scaling	Same state variables and reward components
MADDPG	Multi-agent baseline with centralized critic and decentralized actors	Same agent partition and action limits
Constrained RL	Standard RL with normalized constraint penalties in the reward	Same penalty normalization and state information
Safe RL	RL with a generic safety layer but without electricity–hydrogen projection detail	Same physical simulator and training seeds
Robust RL	RL trained under adversarial uncertainty but without physics projection	Same Wasserstein radius and uncertainty samples
PI-DRO-MARL	Wasserstein robust critic, decentralized agents, and physics-informed projection	Full proposed configuration

Table 16 adds stronger reinforcement-learning and constrained-learning baselines. PI-DRO-MARL produces the lowest operating cost, the lowest violation rate, and the smallest robustness degradation. PPO and SAC improve over standard RL but remain vulnerable to constraint violations. MADDPG benefits from multi-agent coordination but lacks explicit distributional robustness. Constrained RL and safe RL improve feasibility but do not achieve the same robustness under distribution shift. Robust RL improves deployment stability but remains less feasible without explicit physical projection.

Table 16. Performance comparison with stronger reinforcement-learning baselines.

Method	Total Cost (USD)	Mean Violation Rate (%)	Shifted-Scenario Cost Increase (%)	Runtime per Decision (s)
Deterministic optimization	11,800	7.00	52.4	1.84
Stochastic programming	9950	3.80	31.7	4.62
Standard RL	11,000	11.00	128.6	0.021
PPO	10,480	8.60	96.2	0.027
SAC	10,120	7.40	83.5	0.031
MADDPG	9740	6.20	61.8	0.035
Constrained RL	9620	2.40	44.7	0.052
Safe RL	9430	1.90	39.5	0.057
Robust RL	9310	3.10	24.8	0.064
PI-DRO-MARL	8850	0.37	12.6	0.071

Table 17 isolates the contribution of each framework component. Removing the distributionally robust component substantially increases shifted-scenario degradation. Removing the physics-informed projection raises constraint violations by more than an order of magnitude. Removing the multi-agent structure reduces coordination efficiency and increases cost. The full model performs best across all four metrics.

Table 17. Ablation study for major PI-DRO-MARL components.

Variant	Total Cost (USD)	Mean Violation Rate (%)	Hydrogen-Utilization Efficiency (%)	Shifted-Scenario Degradation (%)
Full PI-DRO-MARL	8850	0.37	91.4	12.6
Without Wasserstein DRO	9520	0.61	88.2	41.3
Without physics-informed projection	9860	6.95	86.7	34.9
Without multi-agent coordination	9740	1.28	78.5	29.6
Without adaptive scenario update	9260	0.54	88.8	24.1
Penalty-only feasibility handling	10,180	8.34	83.9	47.5

Table 18 reports the statistical tests.

Table 18. Statistical significance over ten random seeds for the primary IEEE 33-bus case.

Method	Total Cost, Mean (95% CI)	Violation Rate, Mean (95% CI)	Shift Degradation, Mean (95% CI)	Paired Test vs. PI-DRO-MARL
Standard RL	11,000 (10,720–11,360)	11.0 (9.8–12.2)	128.6 (117.4–139.8)	$p < 0.001$
PPO	10,480 (10,220–10,760)	8.6 (7.5–9.8)	96.2 (87.9–104.6)	$p < 0.001$
SAC	10,120 (9890–10,370)	7.4 (6.5–8.5)	83.5 (75.6–91.4)	$p < 0.001$
MADDPG	9740 (9520–9990)	6.2 (5.4–7.1)	61.8 (55.4–68.2)	$p < 0.001$
Safe RL	9430 (9240–9650)	1.9 (1.5–2.4)	39.5 (34.2–44.8)	$p = 0.003$
Robust RL	9310 (9110–9520)	3.1 (2.5–3.8)	24.8 (20.7–28.9)	$p = 0.006$
PI-DRO-MARL	8850 (8770–8940)	0.37 (0.31–0.44)	12.6 (10.9–14.3)	–

Table 19 summarizes the robustness tests.

Table 19. Robustness stress tests for communication, observability, seasonal variability, cyber events, and component failures.

Stress Condition	Implementation	Cost Increase (%)	Violation Rate (%)	Recovery Behavior
Communication delay	One- to three-step delayed local observations for 30% of agents	4.8	0.55	Stable after projection
Partial observability	Local agents receive neighborhood states and delayed global price summaries	5.6	0.61	Moderate cost penalty
Missing data	10% random measurement dropout with last-observation imputation	3.7	0.49	No instability observed
Cyber data injection	Biased renewable forecast residuals during 5% of intervals	7.9	0.74	Robust critic reduces cost escalation
Seasonal variability	Summer/winter renewable and hydrogen-demand profiles	6.4	0.58	Radius calibration remains effective
Component failure	One electrolyzer outage or one distribution line derating event	8.6	0.82	Feasible redispatch through projection

Table 20 reports computational scalability. The trained PI-DRO-MARL policy has a slightly larger per-decision runtime than unconstrained RL because of the projection layer, but it remains well below the 15 min dispatch interval. Scenario-based stochastic programming is the most expensive online method because it solves many scenario subproblems at each decision step.

Table 20. Computational-time comparison.

Method	Offline Training or Setup Time	Online Decision Time	Feasibility Enforcement
Deterministic optimization	None	1.84 s	Direct constraints
Stochastic programming	Scenario generation: 18.2 min	4.62 s	Direct constraints under scenarios
Standard RL	2.7 h training	0.021 s	Reward penalty only
MADDPG	3.4 h training	0.035 s	Reward penalty only
Safe RL	4.1 h training	0.057 s	Safety layer and penalties
Robust RL	4.3 h training	0.064 s	Robust value update
PI-DRO-MARL	4.8 h training	0.071 s	Physics-informed projection

8. Discussion

The numerical results support three system-level observations. First, distributional robustness directly improves deployment reliability. The flat cost response in Figure 6 and the degradation values in Table 16 show that PI-DRO-MARL avoids the sharp performance collapse observed in standard RL when deployment conditions differ from the training distribution. Second, feasibility requires more than penalty shaping. The violation results in Figure 7 and Table 17 show that projection-based enforcement is essential for voltage, line-flow, and hydrogen-balance constraints. Third, coordinated multi-agent control improves conversion efficiency. Figure 8 demonstrates that decentralized agents trained with a centralized robust critic align surplus renewable electricity with hydrogen production more effectively than uncoordinated policies.

The framework also clarifies the trade-off between robustness and conservativeness. Increasing the Wasserstein radius enlarges the ambiguity set and improves protection against extreme deviations, but excessive radius values can increase operating cost by over-hedging against unlikely scenarios. The selected radius therefore reflects a risk preference: smaller radii favor economic efficiency under typical conditions, whereas larger radii favor resilience under severe renewable and demand shifts. In the tested system, $\varepsilon = 0.10\text{--}0.20$ provides a favorable balance between cost and robustness.

The computational results confirm that the learned policy is suitable for real-time dispatch. The online decision time of PI-DRO-MARL remains far below the 15 min dispatch interval. Most computational cost is shifted to offline training, where adversarial sampling, robust critic learning, and projection-layer calibration are conducted. This separation between offline training and online execution is important for operational deployment because real-time controllers require predictable response times. The results should be interpreted as simulation-validated operational evidence rather than a field-deployment guarantee. The feasibility results are guaranteed with respect to the modeled AC, storage, conversion, and hydrogen-balance constraints and the numerical tolerance of the projection solver. Physical deployment would still require calibrated digital twins, real supervisory-control interfaces, communication tests, cyber-security validation, and hardware-in-the-loop experiments before closed-loop operation in an actual utility environment.

9. Limitations and Future Work

The framework has several limitations. First, the projection layer relies on sequential convex approximation for nonconvex AC constraints and therefore provides local feasibility to numerical tolerance rather than a certified global optimum. Future work will integrate convex relaxations, warm-started nonlinear solvers, and certified feasibility recovery to strengthen projection reliability. Second, the Wasserstein radius is calibrated from empirical uncertainty and statistical confidence levels, but radius selection remains a practical modeling choice. Future research will develop adaptive radius tuning based on real-time forecast error, risk preference, and seasonal regime changes. Third, the evaluation includes modified IEEE 33-bus, IEEE 69-bus, and IEEE 123-node reference systems coupled with hydrogen subsystems, but larger transmission–distribution–hydrogen networks will still be needed to test scalability under thousands of buses and heterogeneous market participants. Fourth, the learning environment assumes sufficiently accurate simulator dynamics. Field deployment will require digital-twin calibration, hardware-in-the-loop testing, cyber-security screening, and fallback control policies. Fifth, the current stress tests represent delayed measurements, missing data, cyber data injection, seasonal variability, and component failures in simulation. Future extensions will incorporate communication protocols, adversarially robust state estimation, privacy-preserving multi-agent training, and hardware-in-the-loop validation.

10. Conclusions

This study developed a physics-informed distributionally robust multi-agent reinforcement learning framework for coordinated new-type power system operation. The framework integrates engineering constraints from AC power flow, hydrogen conversion, hydrogen storage, battery dynamics, and generator ramping with a Wasserstein distributionally robust Markov game. The learning algorithm combines centralized robust training, decentralized execution, adversarial distribution updates, and projection-based feasibility enforcement. Theoretical analysis establishes projected feasibility to numerical tolerance, a Wasserstein robustness bound, and convergence to a stationary point of the nonconvex robust learning surrogate under stated assumptions.

The modified IEEE 33-bus primary case shows that PI-DRO-MARL achieves an operating cost of USD 8850, a mean constraint-violation rate of 0.37%, and a shifted-scenario degradation of 12.6%, while maintaining an online decision time of 0.071 s. Additional IEEE 69-bus and IEEE 123-node validation indicates that the same framework retains low violation rates and real-time computational performance as system size increases. Comparisons against deterministic optimization, stochastic programming, PPO, SAC, MADDPG, constrained RL, safe RL, robust RL, and standard RL show that no single baseline simul-

taneously achieves the same level of economic performance, feasibility, and robustness. The ablation results further confirm that Wasserstein robustness, physics-informed projection, adaptive scenario updating, and multi-agent coordination each provide distinct performance benefits. The framework therefore provides a technically grounded simulation pathway for real-time, uncertainty-aware control of NTPS infrastructure, while field deployment requires site-specific validation and hardware-in-the-loop testing.

Author Contributions: Conceptualization, F.L. and Z.T.; methodology, F.L., O.Z., and J.Y.; software, O.Z. and Z.X.; validation, B.F., R.F. and Z.X.; formal analysis, J.Y. and B.F.; investigation, F.L. and R.F.; resources, F.L., B.F. and Z.T.; data curation, B.F., R.F. and Z.X.; writing—original draft preparation, O.Z. and F.L.; writing—review and editing, J.Y., Z.X. and Z.T.; visualization, O.Z. and R.F.; supervision, Z.T.; project administration, F.L. and Z.T.; funding acquisition, F.L. and Z.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Economic and Technological Research Institute of State Grid Qinghai Electric Power Company and State Grid Qinghai Electric Power Company, Science and Technology Project “Research on Multi-Scale Grid Interaction Technologies and Operational Mechanisms for Electricity–Hydrogen Coupling System Towards the New-Type Power System”, grant number 522830250009.

Data Availability Statement: The simulation inputs, processed uncertainty profiles, scenario-generation configuration, random seeds, hyperparameter table, baseline settings, and tabulated numerical data supporting the figures and tables are included in the supplementary reproducibility package submitted with this manuscript. The IEEE reference network data are based on standard public benchmark systems; the coupled hydrogen-system parameters and generated scenario tables are provided as CSV and YAML files in the supplementary package. Additional implementation details are available from the corresponding author on reasonable request.

Conflicts of Interest: Authors Fei Liu, Jun Yin, Baomin Fang and Ruiming Fan were employed by the company State Grid Qinghai Electric Power Company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Zhou, B.; Li, J.; Chan, K.W.; Cao, Y.; Kuang, Y.; Liu, X.; Wang, X. Smart home energy management systems: Concept, configurations, and scheduling strategies. *Renew. Sustain. Energy Rev.* **2016**, *61*, 30–40. [\[CrossRef\]](#)
2. He, G.; Mallapragada, D.S.; Bose, A.; Heuberger, C.F.; Gençer, E. Hydrogen supply chain planning with flexible production and storage. *Appl. Energy* **2023**, *334*, 120655.
3. Li, Y.; Wei, W.; Wang, J.; Mei, S. Strategic scheduling of integrated electricity and hydrogen systems. *IEEE Trans. Smart Grid* **2021**, *12*, 659–671. [\[CrossRef\]](#)
4. Mohajerin Esfahani, P.; Kuhn, D. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Program.* **2018**, *171*, 115–166.
5. Shafiei, K.; Ghassem Zadeh, S.; Tarafdar Hagh, M. Planning for a network system with renewable resources and battery energy storage, focused on enhancing resilience. *J. Energy Storage* **2024**, *87*, 111339. [\[CrossRef\]](#)
6. Shafiei, K.; Ghassem Zadeh, S.; Tarafdar Hagh, M. A novel multi-objective optimization approach for resilience enhancement considering integrated energy systems with renewable energy, energy storage, energy sharing, and demand-side management. *J. Energy Storage* **2025**, *115*, 115966. [\[CrossRef\]](#)
7. Shafiei, K.; Ghassem Zadeh, S.; Tarafdar Hagh, M. Robustness and resilience of energy systems to extreme events: A review of assessment methods and strategies. *Energy Strategy Rev.* **2025**, *58*, 101660. [\[CrossRef\]](#)
8. Shafiei, K.; Ghassem Zadeh, S.; Tarafdar Hagh, M. Power system resilience enhancement using graph learning: A comprehensive robustness and antifragility approach. *Sustain. Energy Grids Netw.* **2025**, *43*, 101927. [\[CrossRef\]](#)
9. Shafiei, K.; Tarafdar Hagh, M.; Seifi, A. Integration of wind farm, energy storage and demand response for optimum management of generation and carbon emission. *J. Eng.* **2024**, *2024*, e12348. [\[CrossRef\]](#)
10. Ben-Tal, A.; El Ghaoui, L.; Nemirovski, A. *Robust Optimization*; Princeton University Press: Princeton, NJ, USA, 2009.

11. Bertsimas, D.; Brown, D.B.; Caramanis, C. Theory and applications of robust optimization. *SIAM Rev.* **2011**, *53*, 464–501. [\[CrossRef\]](#)
12. Roald, L.; Krause, T.; Andersson, G. Integrated balancing and congestion management under uncertainty using robust optimization. *Electr. Power Syst. Res.* **2023**, *216*, 108942.
13. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.
14. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv* **2017**, arXiv:1707.06347.
16. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 1861–1870.
17. Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
18. Achiam, J.; Held, D.; Tamar, A.; Abbeel, P. Constrained policy optimization. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 22–31.
19. Iyengar, G.N. Robust dynamic programming. *Math. Oper. Res.* **2005**, *30*, 257–280. [\[CrossRef\]](#)
20. Nilim, A.; El Ghaoui, L. Robust control of Markov decision processes with uncertain transition matrices. *Oper. Res.* **2005**, *53*, 780–798. [\[CrossRef\]](#)
21. García, J.; Fernández, F. A comprehensive survey on safe reinforcement learning. *J. Mach. Learn. Res.* **2015**, *16*, 1437–1480.
22. Dalal, G.; Dvijotham, K.; Vecerik, M.; Hester, T.; Paduraru, C.; Tassa, Y. Safe exploration in continuous action spaces. *arXiv* **2018**, arXiv:1801.08757.
23. Molzahn, D.K.; Hiskens, I.A. A survey of relaxations and approximations of the power flow equations. *Found. Trends Electr. Energy Syst.* **2019**, *4*, 1–221. [\[CrossRef\]](#)
24. Farivar, M.; Low, S.H. Branch flow model: Relaxations and convexification. *IEEE Trans. Power Syst.* **2013**, *28*, 2554–2564. [\[CrossRef\]](#)
25. Zhang, H.; Liu, W.; Xu, Y. Multi-agent deep reinforcement learning for distributed energy management in microgrids. *IEEE Trans. Smart Grid* **2023**, *14*, 3081–3093. [\[CrossRef\]](#)
26. Qiu, R.; Wang, Y.; Sun, H. Green hydrogen-based energy storage service via power-to-gas systems: A multi-agent reinforcement learning approach. *Appl. Energy* **2023**, *350*, 121753. [\[CrossRef\]](#)
27. Wang, X.; Zhang, J.; Kang, C. Deep reinforcement learning for real-time scheduling of integrated electricity-hydrogen systems. *Energy* **2024**, *293*, 130581.
28. Zhao, A.P.; Li, S.; Xie, D.; Wang, Y.; Li, Z.; Hu, P.J.-H.; Zhang, Q. Hydrogen as the nexus of future sustainable transport and energy systems. *Nat. Rev. Electr. Eng.* **2025**, *2*, 447–466. [\[CrossRef\]](#)
29. Li, T.T.; Zhao, A.P.; Wang, Y.; Li, S.; Fei, J.; Wang, Z.; Xiang, Y. Integrating solar-powered electric vehicles into sustainable energy systems. *Nat. Rev. Electr. Eng.* **2025**, *2*, 467–479. [\[CrossRef\]](#)
30. Li, T.T.; Huang, Y.; Hua, Y.; Hu, T.; Yang, H.; Alhazmi, M.; Alsaif, S.; Liu, G. Cyber-physical resilience and security strategies for renewable energy-driven IoT power grids. *Sustain. Energy Technol. Assess.* **2025**, *84*, 104696. [\[CrossRef\]](#)
31. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686–707. [\[CrossRef\]](#)
32. Misyris, G.S.; Venzke, A.; Chatzivasileiadis, S.; Baker, K. Physics-informed neural networks for power systems. In *2020 IEEE Power & Energy Society General Meeting*; IEEE: New York, NY, USA, 2020; pp. 1–5.
33. Bento, M.E.C. Load margin assessment of power systems using physics-informed neural network with optimized parameters. *Energies* **2024**, *17*, 1562. [\[CrossRef\]](#)
34. Yang, G.; Xu, G.; Jiang, R.; Jiang, Y.; Chen, X.; Sun, L.; Li, Y.; Gao, Y. A parameter-free fault location algorithm for hybrid transmission lines using double-ended data synchronization and physics-informed neural networks. *Energies* **2025**, *18*, 5710. [\[CrossRef\]](#)
35. Vairo, T.; Guarino, S.; Reverberi, A.P.; Fabiano, B. Physics-informed neural networks for thermal anomaly prediction and early detection. *Energies* **2026**, *19*, 2503. [\[CrossRef\]](#)
36. Biagioni, D.; Zhang, X.; Wald, A.; Vaidhynathan, D.; Chintala, R.; King, J.; Zamzam, A.S. PowerGridworld: A framework for multi-agent reinforcement learning in power systems. In Proceedings of the ACM International Conference on Future Energy Systems, Virtual Event, 28 June–1 July 2022; pp. 565–570.
37. Ning, Z.; Xie, L. A survey on multi-agent reinforcement learning and its application. *J. Autom. Intell.* **2024**, *3*, 73–91. [\[CrossRef\]](#)
38. Wu, Z.; Zhang, M.; Gao, S.; Wu, Z.-G.; Guan, X. Physics-informed reinforcement learning for real-time optimal power flow with renewable energy resources. *IEEE Trans. Sustain. Energy* **2025**, *16*, 216–226. [\[CrossRef\]](#)
39. Yu, P.; Zhang, H.; Song, Y.; Wang, Z.; Dong, H.; Ji, L. Safe reinforcement learning for power system control: A review. *Renew. Sustain. Energy Rev.* **2025**, *223*, 116022. [\[CrossRef\]](#)

40. Shi, Y.; Feng, M.; Wang, M.; Zhou, W.; Li, H. Multi-agent reinforcement learning with safety layer for active voltage control. In Proceedings of the International Conference on Autonomous Agents and Multiagent Systems, London, UK, 29 May–2 June 2023; pp. 1533–1541.
41. Liu, G.; Iloglu, S.; Caldara, M.; Durham, J.W.; Zavlanos, M.M. Distributionally robust multi-agent reinforcement learning for dynamic chute mapping. *arXiv* **2025**, arXiv:2503.09755.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.