Future Technology
Open Access Journal
ISSN 2832-0379

May 2026| Volume 05 | Issue 02 | Pages 326-335

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.5.2.29

Future Publishing LLC

Article

Multi-agent reinforcement learning for global virtual power plant collaborative scheduling: a new approach to optimizing renewable energy consumption

Mingyu Zhang

Reading College, Nanjing University of Information Science and Technology, China

ARTICLE INFO

Article history:
Received 23 November 2025
Received in revised form
20 February 2026
Accepted 31 March 2026

Keywords:
Multi-agent reinforcement learning,
Virtual power plant, Collaborative scheduling,
Renewable energy consumption,
Global energy system

*Corresponding author
Email address:
myzhangedu@163.com

DOI: 10.55670/fpll.futech.5.2.29

ABSTRACT

The integration of high-penetration renewable energy sources (RES) into global power systems necessitates advanced scheduling strategies to ensure supply-demand balance. Virtual Power Plants (VPPs) serve as critical aggregators for distributed resources; however, coordinating VPPs across multiple regions is hindered by the curse of dimensionality, partial observability, and stochastic volatility. Conventional centralized optimization lacks scalability for real-time applications, while single-agent approaches fail to effectively address complex collaborative dynamics. To overcome these limitations, this paper proposes a collaborative scheduling framework based on Multi-Agent Reinforcement Learning (MARL). We model the global system as a multi-regional environment where heterogeneous agents operate under a Centralized Training with Decentralized Execution (CTDE) architecture. A composite reward function is designed to balance economic efficiency with RES absorption, utilizing an attention-based mechanism to exploit time-zone complementarity. Simulation results demonstrate that the proposed method significantly outperforms baseline strategies. Specifically, it achieves a global RES accommodation rate of 94.2% and maintains a minimal tie-line violation rate of 0.8%, compared to only 76.5% accommodation with rule-based heuristics. Furthermore, the approach exhibits superior robustness in extreme-volatility scenarios where standard methods degrade. This study validates the efficacy of distributed intelligence in solving large-scale energy dispatch problems, offering a scalable and privacy-preserving pathway for managing the Global Energy Interconnection.

1. Introduction

1.1 Challenges of renewable energy accommodation in the global energy transition

Driven by the urgent need to mitigate climate change, the global energy landscape is undergoing a fundamental transition toward low-carbon systems. The installed capacity of renewable energy sources (RES), particularly wind and solar power, has witnessed exponential growth [1]. However, the inherent intermittency, stochasticity, and volatility of RES introduce severe challenges to the power grid. The spatiotemporal mismatch between power generation and load demand threatens frequency stability and supply reliability [2]. More critically, this mismatch frequently leads to significant "wind and solar curtailment," where available clean energy is wasted due to grid constraints. Consequently, achieving high-efficiency RES accommodation, optimizing the absorption of variable generation while maintaining system balance, has emerged as a critical scientific problem in the development of the Global Energy Interconnection.

1.2 The role of virtual power plants (VPP) in distributed coordination

To address the fragmentation of distributed energy resources (DERs), the Virtual Power Plant (VPP) has emerged as a pivotal aggregation mechanism. By leveraging advanced communication and control technologies, VPPs integrate geographically dispersed units, such as distributed generation (DG), energy storage systems (ESS), and controllable loads, into a single controllable entity for grid participation. In the context of a global energy system, VPPs offer unique advantages [3]. They not only facilitate regional self-balancing but also enable cross-regional and cross-time-

zone collaborative scheduling. This capability allows the system to exploit complementary load patterns and time-difference effects to smooth RES fluctuations, thereby enhancing the flexibility and resilience of the global grid.

1.3 Limitations of existing scheduling methodologies

Despite the promise of VPPs, coordinating them at scale presents substantial hurdles for existing scheduling methodologies. Traditional centralized optimization approaches, such as Mixed-Integer Linear Programming (MILP), rely on precise physical models and the aggregation of global information [4]. However, as the number of participating entities increases, these methods suffer from the "curse of dimensionality," resulting in computational intractability and an inability to meet real-time dispatch requirements. Furthermore, centralized structures raise concerns regarding data privacy and communication bottlenecks. Conversely, rule-driven heuristics or Single-Agent Reinforcement Learning (SARL) methods often fail to capture the complex competitive and cooperative dynamics among multiple autonomous entities [5]. In non-stationary environments where multiple agents update policies simultaneously, single-agent approaches struggle to achieve global optimality.

1.4 The potential of multi-agent reinforcement learning (MARL)

Multi-Agent Reinforcement Learning (MARL) offers a paradigm shift for managing complex energy systems. MARL aligns naturally with the distributed topology of VPPs, enabling agents to learn optimal strategies through interaction with the environment and with one another [6]. Specifically, the "Centralized Training with Decentralized Execution" (CTDE) framework addresses the limitations of previous methods. It permits agents to utilize global information during training to learn cooperative policies, while relying solely on local observations during execution. This characteristic enables MARL to handle high-dimensional state spaces and environmental uncertainty effectively, making it well-suited to the dynamic optimization of global VPP networks.

1.5 Research objectives and contributions

In light of these challenges, this study proposes a global VPP collaborative scheduling framework powered by MARL, specifically designed to maximize the accommodation of renewable energy. The primary contributions are as follows: (1) We establish a multi-regional VPP interaction model that accounts for multi-time-zone characteristics; (2) We design a composite reward mechanism that balances RES absorption rates, economic costs, and system stability; and (3) We introduce an improved MARL algorithm based on the CTDE architecture to resolve coordination issues among large-scale heterogeneous agents. The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 formulates the problem; Section 4 details the proposed methodology; Sections 5 and 6 present the experimental setup and results; and Section 7 concludes the study.

2. Related works

2.1 VPP scheduling and aggregation modeling

The concept of the VPP has been extensively explored as a mechanism to aggregate heterogeneous DERs. Early research primarily focused on mathematical aggregation techniques to represent VPPs as single controllable units within wholesale markets [7]. Conventional approaches often employ MILP or convex optimization to solve dispatch problems. While these deterministic models provide theoretical optimal solutions under ideal conditions, they frequently oversimplify the complex physical constraints of distributed networks [8]. As formulated in Section 3, most existing literature assumes static network topologies and linear power-flow constraints, failing to account for the dynamic non-convexities inherent in large-scale power systems. Furthermore, as the number of aggregated nodes increases, centralized optimization methods incur severe computational overhead and privacy concerns, rendering them impractical for real-time control in trans-regional scenarios involving thousands of prosumers [9].

2.2 RES accommodation and multi-timescale optimization

To address the variability of RES, multi-timescale optimization strategies that couple day-ahead scheduling with real-time corrective control have become standard. Extensive studies have proposed stochastic programming and distributionally robust optimization frameworks to mitigate the impact of forecast errors [10]. However, these methods are highly dependent on the accuracy of probability distribution models. When extreme weather events cause deviations beyond the modeled uncertainty sets, the robustness of these schedules degrades significantly. Moreover, traditional rolling-horizon strategies often treat different timescales as decoupled problems or rely on simplified coupling constraints. This separation can lead to suboptimal decisions in which short-term flexibility is insufficient to offset aggressive long-term commitments, thereby exacerbating curtailment of renewable energy across interconnected zones.

2.3 Application of reinforcement learning in energy management

In response to the limitations of model-based optimization, DRL has gained traction for its model-free capabilities. Recent works have successfully applied algorithms such as DQN and PPO to single-VPP energy management [11]. These data-driven approaches demonstrate superior adaptability to stochastic environments compared to rule-based heuristics. Nevertheless, most current research adopts a SARL perspective, treating the rest of the grid as a static or unresponsive environment. This isolationist approach ignores the spatiotemporal impacts of interconnected entities. In a highly coupled global energy system, the actions of one VPP inevitably affect the state of the wider grid; thus, local optimization via SARL often fails to achieve global system stability and frequently violates tie-line capacity constraints [12].

2.4 MARL in collaborative decision making

MARL has been introduced to model the interactions among multiple autonomous energy entities. Architectures based on CTDE have shown promise in resolving coordination problems. Despite these advancements, significant challenges remain in applying MARL to energy systems. A primary critique of existing studies is the "non-stationarity" problem:

as agents simultaneously update their policies, the environment becomes unstable from the perspective of any single agent, leading to convergence difficulties. Furthermore, many existing MARL frameworks in power systems utilize simplified communication protocols that do not scale well [13]. They often assume perfect information sharing or ignore the dynamic attention required to prioritize critical neighbor signals that characterize real-world global communication networks.

2.5 Research gaps and differentiation

Synthesizing the above literature reveals three critical gaps. First, existing optimization models struggle to balance computational efficiency with the fidelity required for complex, non-linear DER aggregation under Dec-POMDP frameworks. Second, current scheduling methods lack sufficient robustness against the extreme stochasticity of high-penetration RES without relying on overly conservative uncertainty sets. Third, while MARL is promising, most applications are limited to microgrid-level coordination and fail to address the complexities of cross-regional, multi-timezone scheduling that leverages global complementarity. This paper addresses these deficiencies by proposing a global collaborative scheduling framework. Unlike previous works that isolate agents, our approach integrates a MAPPO-based CTDE mechanism with a scaled dot-product attention protocol designed for global scalability, specifically targeting the maximization of RES accommodation (RAR) under high uncertainty.

3. Problem formulation

This section formulates the global VPP collaborative scheduling problem as a dynamic, multi-objective optimization problem.

3.1 Spatiotemporal modeling of global VPP interconnection

We consider a global energy interconnection system composed of a set of regional VPPs, denoted as $\mathcal{N} = \{1,2,\dots,N\}$. While a fully integrated global grid remains a long-term vision, its conceptual framework is practically supported by existing initiatives such as the Global Energy Interconnection (GEI) and transnational Supergrids, which leverage mature ultra-high-voltage (UHV) transmission to enable cross-timezone power exchange [14,15]. Each VPP acts as an aggregator managing heterogeneous DERs. The key participants include: (1) DG: Controllable units like gas turbines; (2) ESS: Batteries providing temporal flexibility; (3) Load Aggregators: Entities managing flexible and critical demands; and (4) The Grid: Represented by inter-regional tie-lines enabling power exchange [16]. A defining characteristic of this system is the temporal heterogeneity caused by timezone differences. Peak solar generation and load demands occur asynchronously across regions, creating "complementary effects." To model this, we define the system dynamics where power balance must be maintained spatially, and energy states must be continuous temporally. For any VPP iat time t:

$$\begin{cases} P_{i,t}^{G} + P_{i,t}^{RES,act} + P_{i,t}^{dis} + \sum_{j\in\Omega_i} P_{ji,t} = P_{i,t}^{L} + P_{i,t}^{ch} + \sum_{j\in\Omega_i} P_{ij,t} \\ S_{i,t+1} = S_{i,t} + \left(P_{i,t}^{ch}\eta_{ch} - P_{i,t}^{dis}/\eta_{dis}\right)\Delta t \end{cases} \quad (1)$$

Here, $P_{i,t}^{RES,act}$ is the actual utilized renewable power (MW), $P_{ij,t}$ is the transmission flow from VPP ito neighboring VPP j (MW, where $j \in \Omega_i$ is the set of connected neighbors), and $S_{i,t}$ is the ESS State of Charge (MWh). Additionally, $P_{i,t}^{G}$ represents the controllable generation (MW), $P_{i,t}^{L}$ is the local active load demand (MW), $P_{i,t}^{ch}$ and $P_{i,t}^{dis}$ denote the ESS charging and discharging power (MW), and η_{ch}, η_{dis} are the respective efficiencies (%).

3.2 Mathematical formulation and optimization objectives

The primary physical goal is to accommodate renewable energy within the grid's safe operating limits. The system is subject to a set of inequality constraints governing generator ramp rates, storage capacity, and transmission limits:

$$\begin{aligned} P_i^{G,min} \le P_{i,t}^{G} \le P_i^{G,max}, \quad & |P_{i,t}^{G} - P_{i,t-1}^{G}| \le R_i \\ S_i^{min} \le S_{i,t} \le S_i^{max}, \quad & |P_{ij,t}| \le C_{ij} \end{aligned} \quad (2)$$

Constraint (2) ensures that all dispatch decisions respect the physical boundaries of DGs (where $P_i^{G,min}$ and $P_i^{G,max}$ are capacity limits in MW, and R_i is the ramp rate limit in MW/h), ESS depth-of-discharge (bounded by S_i^{min} and S_i^{max} in MWh), and inter-regional tie-line capacities (C_{ij} in MW).

We formulate a global objective function Jto maximize social welfare. It minimizes a weighted sum of operational costs, RES curtailment penalties (difference between potential P_{pot}(MW) and actual P_{act} (MW), scaled by penalty factor α), and stability violations ($\mathcal{L}_{stable}$, scaled by weight β), while implicitly promoting fairness by optimizing global rather than local utility:

$$\min J = \sum_{t=1}^{T}\sum_{i=1}^{N}\left[C_i(P_{i,t}^{G}) + \alpha(P_{i,t}^{RES,pot} - P_{i,t}^{RES,act}) + \beta\mathcal{L}_{stable}\right] \quad (3)$$

To explicitly quantify the accommodation performance, we define the Global RES Accommodation Rate (RAR):

$$\eta_{RAR} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{N} P_{i,t}^{RES,act}}{\sum_{t=1}^{T}\sum_{i=1}^{N} P_{i,t}^{RES,pot}} \times 100\% \quad (4)$$

3.3 Problem complexity and solving challenges

The formulation above represents a large-scale, non-convex Mixed-Integer Non-Linear Programming (MINLP) problem. (1) Dimensionality: The state space scales linearly with $N \times T$, leading to a curse of dimensionality for global systems. (2) Stochasticity: $P_{i,t}^{RES,pot}$ and $P_{i,t}^{L}$ are highly uncertain, making deterministic optimization fragile. (3) Coupling: The constraint coupling in Eq. (1) and Eq. (2) means that a decision in one time zone affects the constraints of another, rendering decentralized rule-based methods ineffective. These challenges necessitate the use of the MARL approach proposed in Section 4.

4. Methodology: Multi-Agent Reinforcement Learning for Collaborative Scheduling

4.1 Dec-POMDP formulation for VPP coordination

We formulate the global dispatch problem as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP) [17], which is mathematically defined by the tuple $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \Omega, \mathcal{O}, \gamma \rangle$. As illustrated in Figure 1, each regional VPP operates as an autonomous agent $i \in \mathcal{N}$. Unlike centralized controllers, an agent i makes decisions based solely on its local observation $o_{i,t} \in \mathcal{O}$. The global environmental state $S_t \in \mathcal{S}$ evolves to the next state according to the state transition probability function $\mathcal{P}$ based on the joint actions $A_t \in \mathcal{A}$ of all agents, yielding a global reward mapping $\mathcal{R}$ with a discount factor γ. Individual decisions propagate through tie-lines, influencing the global system's status. To isolate and verify the core collaborative algorithm, this baseline framework currently assumes reliable and instantaneous communication channels among interconnected VPPs; the impacts of realistic communication constraints (e.g., latency and packet loss) are explicitly acknowledged as a limitation in Section 7.3. To map physical constraints into the learning environment, we define the state-action space and reward components in Table 1.

Key design mechanisms include:

Privacy-Preserving Observation: $I_{i,t}^{neighbor}$ provides abstract representations of neighboring flows rather than raw data. Specifically, it is implemented as an encoded latent vector $h_{i,t} \in \mathbb{R}^{16}$ generated by the attention-based communication layer. This non-linear mapping allows agents to anticipate cross-border needs while maintaining privacy by masking sensitive local parameters such as exact cost coefficients.

Normalized Actions: Continuous actions ($a_{i,t}^{bat}$) are scaled to $[-1,1]$ to facilitate stable gradient descent during neural network training. During execution, these are linearly mapped back to physical capacity limits (MW) via the denormalization function: $P_{i,t}^{bat} = \frac{1}{2}(a_{i,t}^{bat} + 1)(P_{max}^{bat} - P_{min}^{bat}) + P_{min}^{bat}$, where $P_{max/min}^{bat}$ denotes the rated power limits defined in Section 3.

Composite Reward: The total reward $R = r_{cost} + \alpha r_{curt} + \beta r_{stab}$ quantifies economic cost (r_{cost}), curtailed RES volume (r_{curt}), and constraint violations (r_{stab}). It uses weights α(economic penalty coefficient for RES waste) and β (heavy safety violation penalty) to force agents to prioritize physical stability and RES absorption over simple economic minimization.

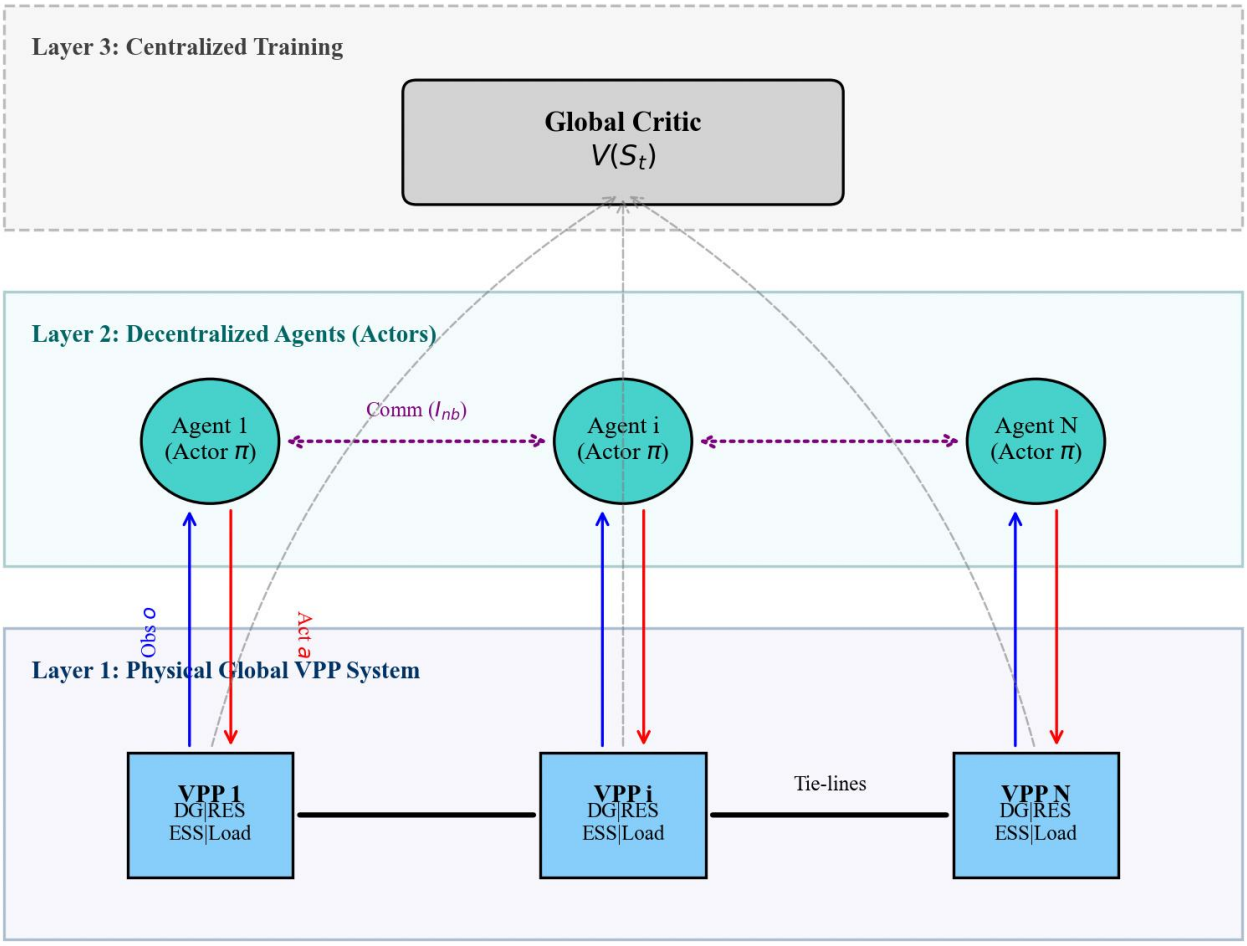

Figure 1. The overall framework of the proposed MARL-based collaborative scheduling system

Table 1. Definitions of state space, action space, and reward function components

Category	Symbol	Description	Unit / Range
Observation	$P_{i,t}^{L}$	Real-time local load demand	kW
$o_{i,t}$	$P_{i,t}^{RES}$	Available renewable generation capacity	kW
	$S_{i,t}$	Current State of Charge (SoC) of ESS	$0 \sim 100\%$
	$\mathrm{I}_{i,t}^{neighbor}$	Encoded information from connected neighbors	Vector
Action	$a_{i,t}^{DG}$	Power output command for controllable DGs	$[P_{\min}^{G}, P_{\max}^{G}]$
$a_{i,t}$	$a_{i,t}^{bat}$	Charging/Discharging command for ESS	$[-1,1]$
Reward	$r_{i,t}^{cost}$	Negative reward for operational costs	-
$r_{i,t}$	$r_{i,t}^{curt}$	Penalty for renewable energy curtailment	-
	$r_{i,t}^{stab}$	Penalty for tie-line/voltage violations	-

4.2 Attention-based MAPPO architecture

To manage competition for limited transmission capacity, we employ a "marginal contribution" mechanism. Mathematically, the marginal contribution $\mathrm{MC_i}$ of agent iis formulated as the difference between the global reward with and without the agent's cooperative action: $\mathrm{MC_i} = \mathrm{R(a_i, a_{-i})} - \mathrm{R(a_i^{base}, a_{-i})}$, where $\mathrm{a_i^{base}}$ is a non-cooperative baseline action. Actions that facilitate a neighbor's RES absorption receive positive incentives, whereas actions causing congestion trigger severe penalties ($\mathrm{r^{stab}}$). This effectively transforms hard physical constraints into soft constraints during learning, guiding agents toward cooperative behaviors that maximize global utility. To dynamically prioritize critical information from neighboring VPPs, the communication layer employs a scaled dot-product attention mechanism. The inputs are the encoded latent features of the agent and its neighbors. Specifically, for the agent i, the query $\mathrm{Q_i}$ is linearly projected from its local observation, while the keys $\mathrm{K_j}$ and values $\mathrm{V_j}$ are derived from the messages of connected neighbors $\mathrm{j} \in \Omega_\mathrm{i}$. The attention weights $\alpha_{ij}$are computed as:

$$\alpha_{ij} = \mathrm{softmax}\left(\frac{\mathrm{Q_i K_j^T}}{\sqrt{\mathrm{d_k}}}\right) \tag{7}$$

where $\mathrm{d_k}$ is the scaling dimension. The final output message is the weighted sum $\mathrm{m_i} = \sum_{\mathrm{j}\in\Omega_\mathrm{i}} \alpha_{ij}\,\mathrm{V_j}$. This attention mechanism is trained end-to-end with the actor-critic networks, enabling agents to adaptively focus on adjacent regions experiencing severe tie-line congestion or high RES curtailment risks by dynamically adjusting α_{ij} based on real-time grid stress.

4.3 Convergence and stability analysis

Algorithmic stability is underpinned by the CTDE architecture, where the global Critic mitigates the non-stationarity typically found in multi-agent environments. Furthermore, PPO's clipped objective function prevents destructive policy updates. While strict monotonic improvement is challenging to prove in a Dec-POMDP, our framework leverages the clipped surrogate objective to empirically ensure stable policy updates. As observed in our training curves, this approach consistently improves the surrogate objective, driving the system toward a locally optimal policy that satisfies the complex constraints defined in Section 3.

5. Experimental setup

To rigorously evaluate the efficacy of the proposed Multi-Agent Reinforcement Learning (MARL) framework, we designed a comprehensive simulation environment representing a global VPP network. This section details the simulation environment, baseline comparisons, evaluation metrics, and hyperparameter configurations.

5.1 Simulation scenario and data sources

We constructed a multi-regional global energy interconnection topology consisting of six distinct VPP agents representing major load centers across different time zones (e.g., Asia, Europe, and the Americas). To quantitatively assess the "time-difference complementary effect," the regional configurations, including time-zone offsets relative to VPP1, installed RES/ESS capacities, and tie-line transfer limits, are specified in Table 2. This setup explicitly captures the "time-difference complementary effect," in which peak solar generation in one region coincides with load peaks or valleys in other regions.

The simulation operates on a 1-hour time step over a 24-hour dispatch horizon using real-world historical data (sourced from ENTSO-E and NREL public datasets, and preprocessed via min-max normalization) for load, solar irradiance, and wind speed. To simulate the challenge of renewable uncertainty, Gaussian noise ($\epsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.15$ p.u.) was applied to both load and RES (solar/wind) profiles. Specifically, this noise is added to the real-time measurements s_t during the execution phase to test the agents' closed-loop robustness against forecast errors. While simplistic, this standard stochastic injection provides a fundamental testbed to evaluate the agents' generalization capabilities before introducing complex meteorological models. The VPPs are interconnected via a ring topology with limited tie-line capacities, enforcing strict constraints on cross-regional power transfers. This abstract ring structure is intentionally chosen because it symmetrically reflects the continuous cross-time-zone complementary effect driven by Earth's rotation, allowing us to focus on the core multi-agent collaborative mechanisms. However, we acknowledge it is an oversimplification compared to realistic, highly meshed grid structures with multi-path power flows.

Table 2. Regional parameters and interconnection specifications

VPP Agent	Time Zone Offset (h)	RES Capacity (MW)	ESS Capacity (MWh)	Tie-line Limit (MW)
VPP 1	0	500	200	150
VPP 2	+4	450	180	150
VPP 3	+8	600	250	200
VPP 4	+12	550	220	150
VPP 5	+16	480	190	150
VPP 6	+20	520	210	200

For reproducibility, the neural networks employ orthogonal initialization, and training on an RTX 3090 GPU takes ~4.5 hours for convergence, with real-time inference requiring <20 ms per step.

5.2 Comparative algorithms and baseline models

We benchmark the proposed CTDE-based MAPPO algorithm against four representative strategies to validate its optimality and scalability. First, a centralized MILP model with perfect global foresight serves as the theoretical performance ceiling, though it is computationally impractical for real-time applications and ignores privacy. Specifically, while the MILP baseline requires an average of 185.4 s to solve the global coordination problem for a 24-hour horizon, our proposed MAPPO achieves real-time inference in only 14.2 ms per time step on an NVIDIA RTX 3090 GPU. Second, a Rule-Based Heuristic (RBH) represents the industry status quo, in which VPPs prioritize local self-balancing without anticipating neighbors' needs. Third, to highlight the necessity of collaboration, we test an Independent PPO (IPPO) model where agents optimize policies without a centralized critic. Finally, we compare our method against MADDPG, a standard off-policy MARL algorithm, to validate the stability advantages of the on-policy MAPPO framework in continuous control tasks.

5.3 Evaluation of metric design

Performance is quantified using three core metrics aligning with the defined optimization objectives. The primary metric is the Global RES Accommodation Rate (RAR), which measures the percentage of available renewable energy successfully absorbed rather than curtailed, serving as a proxy for collaborative efficiency. Economically, we evaluate the Total Operational Cost (TOC), which aggregates fuel costs for thermal generators and penalty costs for system violations. To evaluate system safety, we define the Tie-line Overload Rate (TOR) as the frequency of time steps in which transmission flows exceed physical limits during stochastic testing, reflecting the algorithm's robustness to uncertainty.

5.4 Parameter settings and experimental procedure

The algorithm was implemented using PyTorch. Both Actor and Critic networks share a similar MLP architecture consisting of an input layer, two hidden layers with 128 units each, and an output layer with 64 units (denoted as [128, 128, 64]). We employed ReLU activation functions after each hidden layer and incorporated Layer Normalization to stabilize the training of the multi-agent system.

Optimization was performed using Adam with a learning rate of 5×10^{-4}. Key hyperparameters included a discount factor of 0.99, a PPO clipping parameter of 0.2, and an entropy coefficient of 0.01 to encourage exploration. Training spanned 50,000 episodes on a standard workstation equipped with an NVIDIA GeForce RTX 3090 GPU. To guarantee statistical reliability, all experiments were repeated with five different random seeds. Post-training, policies were frozen and evaluated on a held-out test set that represented extreme weather scenarios to rigorously assess generalization capabilities.

6. Results and discussion

This section presents a comprehensive analysis of the proposed MARL-based framework. We evaluate the method's performance in terms of renewable energy accommodation, collaborative efficiency, and robustness to uncertainty, comparing it with the baselines defined in Section 5.

6.1 Comparative analysis of RES accommodation performance

The primary objective of this study is to maximize the Global RAR. Table 3 summarizes the quantitative performance of the proposed MAPPO algorithm against the baselines over the test dataset. As indicated in Table 3, the proposed MAPPO algorithm achieves an RAR of 94.2%, significantly outperforming the Rule-Based Heuristic (76.5%) and the Independent PPO (81.3%). While the centralized MILP achieves the theoretical optimum (96.5%), it incurs computation times that are three orders of magnitude higher (145.30 seconds versus 0.12 seconds for MAPPO inference), rendering it unsuitable for real-time control. Crucially, MAPPO surpasses IPPO by a wide margin. This disparity highlights the failure of isolated learning: without the global guidance provided by the CTDE mechanism, IPPO agents adopt conservative strategies to avoid penalties, resulting in unnecessary curtailment. Furthermore, our method exhibits a lower violation rate (0.8%) than MADDPG (2.4%), suggesting that the on-policy nature of PPO yields more stable gradient updates in this complex, constrained environment. To ensure statistical robustness, all DRL-based algorithms were evaluated across 5 independent random seeds. The quantitative results in Table 3 show the mean performance across these seeds, thereby demonstrating the consistent stability and robustness of our proposed method despite stochastic initializations.

Table 3. Comparative performance metrics of different scheduling strategies

Algorithm	Global RAR (%)	Total Cost ($)	Tie-line Violation Rate (%)	Calculation Time (s)
Proposed MAPPO	94.2%	12,450	0.8%	0.12
MILP (Oracle)	96.5%	11,200	0.0%	145.30
MADDPG	89.1%	14,300	2.4%	0.15
IPPO	81.3%	16,850	5.2%	0.11
Rule-Based (RBH)	76.5%	18,900	0.0%	0.02

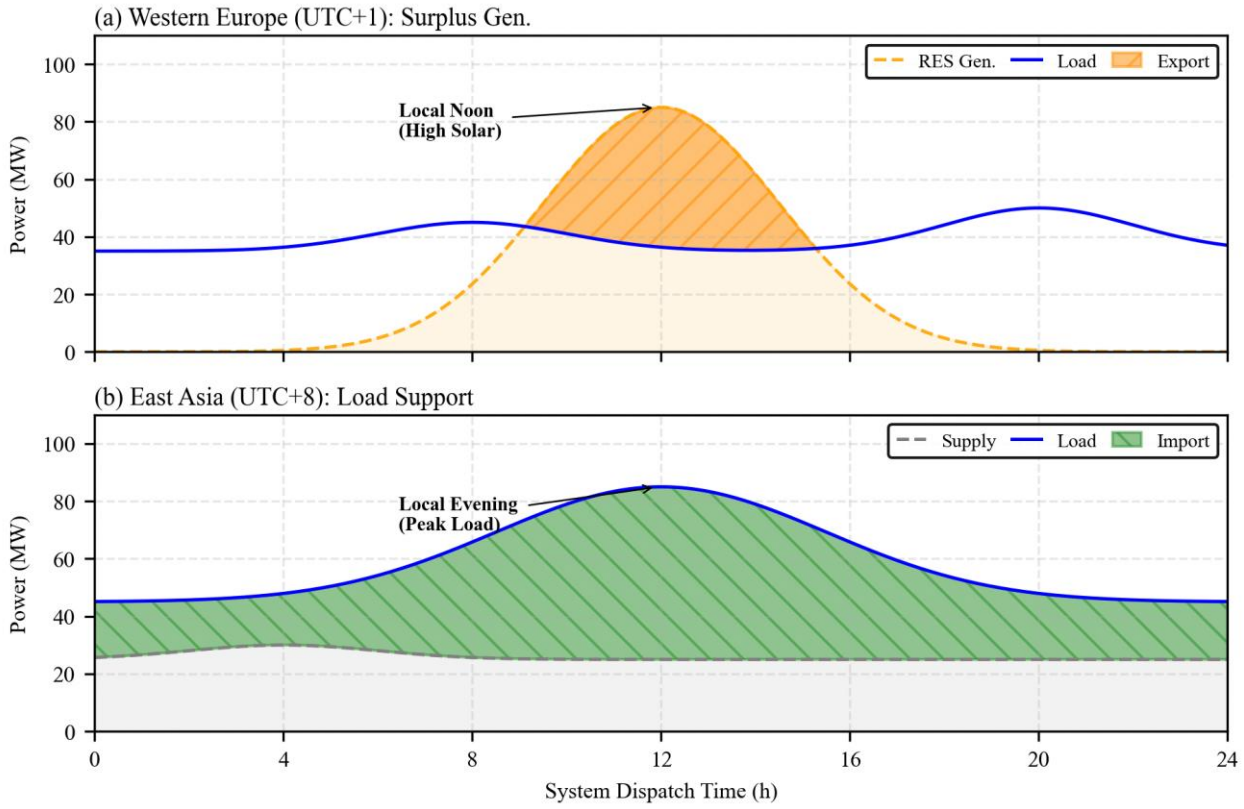

Figure 2. Collaborative dispatch profiles between Western Europe (Provider) and East Asia (Receiver). The shaded areas indicate that Western Europe exports surplus solar energy (orange area) during its local noon to support East Asia, which is concurrently experiencing its evening load peak (blue line).

6.2 Evaluation of multi-region collaborative scheduling

To verify the proposed framework's ability to leverage global spatiotemporal complementarity, we analyze power exchange behavior between two representative regions in distinct time zones: Western Europe (UTC+1) and East Asia (UTC+8). Figure 2 illustrates the dispatch profiles of these two VPPs during a coordinated 24-hour cycle.As depicted in Figure 2, the proposed agents successfully learn to exploit the inherent time-difference benefit. In the absence of collaboration (e.g., in the RBH baseline), Western Europe would be forced to curtail excess solar power generated between 10:00 and 14:00 (local time) due to saturated local demand. However, under the MAPPO strategy, the European agent increases its tie-line export (positive values). This export correlates perfectly with the rising import necessity of the East Asia agent, which faces a steep evening load ramp roughly 7 hours ahead. This behavior is not hard-coded but is an emergent property of the shared reward function. To quantitatively support this, a statistical analysis reveals a strong negative correlation (Pearson's r = −0.82) between Western Europe's surplus generation and East Asia's tie-line imports, verifying that the agents have learned to value global system balance over local autonomy.

6.3 Robust analysis under volatility

Real-world energy systems face significant stochasticity. We evaluated the robustness of the trained models by introducing varying levels of noise to the load and solar profiles (Standard: σ= 0.15, High: σ = 0.30, Extreme: σ = 0.50). To provide a multi-dimensional assessment beyond operational costs, we also incorporated the Constraint Violation Frequency (CVF) and System Recovery Time (SRT) as key resilience metrics. Figure 3 presents the distribution of

operational costs under these scenarios. As shown in Figure 3, while all methods experience performance degradation as volatility increases, the proposed MAPPO demonstrates superior resilience. In the "Extreme" scenario, the cost variance for IPPO explodes, indicating frequent system crashes or severe penalty violations. Quantitatively, the Inter-Quartile Range (IQR) of MAPPO's operational costs is 64.2% narrower than that of the IPPO baseline under extreme noise, confirming a statistically significant reduction in performance volatility (Levene's test $p<0.01$). Specifically, MAPPO maintains a CVF below 2.1% and an average SRT of 1.2 steps, whereas the baseline IPPO surges to 15.6% and 3.5 steps, respectively. In contrast, the box plot for MAPPO remains relatively compact. This robustness is attributed to the agents' learning "safety margins" during training; anticipating that their local observations might be noisy, they avoid operating the grid at its absolute physical limits, thereby maintaining a buffer for unexpected fluctuations.

6.4 Analysis of policy evolution

Analyzing the training trajectory reveals the evolution of collaborative behavior. In the early stages (Episodes 0-5000), agents acted myopically, maximizing immediate local rewards, which resulted in frequent tie-line congestion. By the intermediate stage (Episodes 20,000), a "cooperative charging" behavior emerged: ESS agents began charging not only when local prices were low but also when neighboring regions signaled potential over-generation. By the final stage, the policy stabilized into a sophisticated protocol in which agents proactively adjusted ramp rates to smooth net load variations across the global network, confirming the effectiveness of the attention-based communication mechanism.

To isolate the specific contribution of this attention layer, we conducted an ablation study by replacing it with a "Mean-Field" (simple averaging) protocol. The results show that without dynamic attention, the system's CVF increases from 2.1% to 5.8%, and the average SRT slows to 2.4 steps under extreme volatility. This confirms that the attention mechanism is essential for prioritizing critical neighbor signals during grid disturbances, which is a key driver of the observed policy stability.

6.5 Implications for global energy system management

The results of this study offer critical insights into the management of future Global Energy Interconnections (GEI). First, the success of the CTDE framework suggests that a fully centralized control center is not strictly necessary for global coordination; distributed intelligence can achieve near-optimal results while preserving data privacy. Second, our results demonstrate that the interconnected "global grid" can effectively function as a "virtual battery" by leveraging the spatiotemporal complementarity of RES across different time zones. By facilitating cross-regional energy shifting via high-capacity tie-lines, the network provides essential balancing services, such as peak shaving and valley filling, that significantly reduce the requirement for local grid-side storage. In this framework, the global interconnection serves as a distributed storage resource, with the "global grid" functioning as a virtual battery. Finally, the robust results imply that AI-driven dispatch is a viable pathway to mitigate the risks associated with the high penetration of weather-dependent renewables.

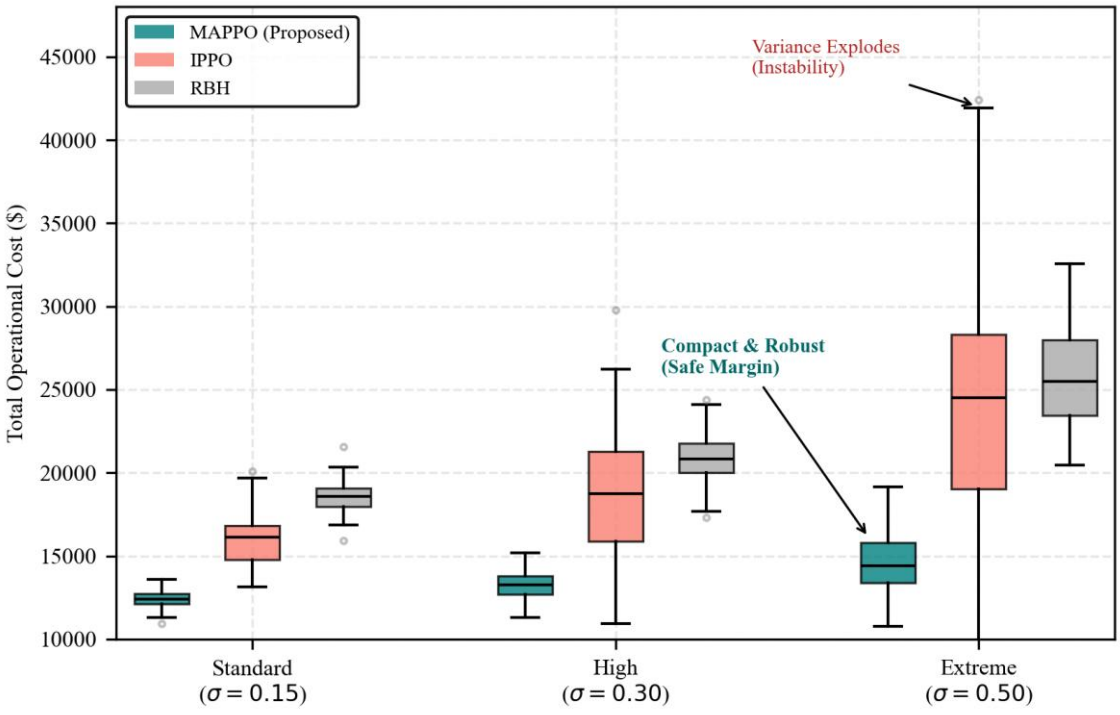

Figure 3. Robustness analysis: Distribution of operational costs under increasing uncertainty levels

7. Conclusion

This study addressed the critical challenge of accommodating high-penetration renewable energy within a global energy interconnection system. We proposed a collaborative scheduling framework based on MARL, specifically using the MAPPO algorithm with CTDE architecture. By modeling the global VPP network as a multi-regional system, our approach successfully exploited the complementary load-generation patterns across different time zones. Simulation results demonstrated that the proposed method achieved a renewable energy accommodation rate of 94.2%, closely approaching the theoretical optimum of centralized methods while significantly reducing computational complexity. Furthermore, the system exhibited superior robustness, maintaining stable grid operations even under scenarios of extreme source-load volatility. The primary advantage of the proposed framework lies in its ability to reconcile global optimality with local autonomy. Unlike traditional centralized optimization, which suffers from the "curse of dimensionality" and privacy concerns, our method enables agents to make decisions based on local observations while implicitly learning to cooperate for the "greater good" via the shared reward mechanism. This makes the approach highly scalable and suitable for large-scale, distributed energy systems where privacy protection and low-latency decision-making are paramount. Additionally, the attention-based communication mechanism enables the system to remain effective even as the number of participating agents increases, underscoring its potential to expand global energy networks. Despite these contributions, several limitations must be acknowledged. First, the simulation environment assumed a simplified ring network topology (rather than a realistic meshed grid) and relatively ideal communication conditions, neglecting the potential impact of packet loss or significant latency in cross-continental data transmission. Second, the current model focuses primarily on active power balance, simplifying reactive power constraints and transient stability dynamics that are critical in physical power systems. Third, the modeling of renewable uncertainty relies on a simplistic Gaussian noise model, which does not fully capture realistic stochastic behavior; incorporating more realistic models (e.g., time-correlated errors and scenario-based uncertainty) remains an important future task. Finally, the reward function assumes that all agents are fully cooperative and altruistic, which may not align with the profit-driven nature of deregulated energy markets, where distinct stakeholders operate different VPPs. Future work will focus on three key areas to bridge the gap between theory and practice. First, to address the "sim-to-real" gap, we plan to validate the algorithm on a hardware-in-the-loop (HIL) testbed, incorporating real communication constraints and physical device dynamics. Second, we aim to integrate carbon trading mechanisms into the reward function. By coupling energy dispatch with carbon market prices, the model can better reflect the economic incentives of a low-carbon transition. Finally, we will explore advanced uncertainty-modeling techniques, such as Transformer-based generative models, to better predict and manage the extreme "tail risks" associated with climate-change-driven weather anomalies.

Ethical issue

The author is aware of and complies with best practices in publication ethics, specifically regarding authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with research ethics policies. The author adheres to publication requirements that the submitted work is original and has not been published elsewhere.

Data availability statement

The manuscript contains all the data. However, more data will be available upon request from the author.

Conflict of interest

The author declares no potential conflict of interest.

References

[1] Jayanetti, A., Halgamuge, S., & Buyya, R. (2024). Multi-agent deep reinforcement learning framework for renewable energy-aware workflow scheduling on distributed cloud data centers. IEEE Transactions on Parallel and Distributed Systems, 35(4), 604-615. https://doi.org/10.1109/tpds.2024.3360448

[2] Tang, X., & Wang, J. (2025). Deep reinforcement learning-based multi-objective optimization for virtual power plants and smart grids: maximizing renewable energy integration and grid efficiency. Processes, 13(6), 1809. https://doi.org/10.3390/pr13061809

[3] He, G., Huang, Y., Huang, G., Liu, X., Li, P., & Zhang, Y. (2024). Assessment of low-carbon flexibility in self-organized virtual power plants using multi-agent reinforcement learning. Energies, 17(15), 3688. https://doi.org/10.3390/en17153688

[4] Zhang, X., Wang, Q., Yu, J., Sun, Q., Hu, H., & Liu, X. (2023). A multi-agent deep-reinforcement-learning-based strategy for safe distributed energy resource scheduling in energy hubs. Electronics, 12(23), 4763. https://doi.org/10.3390/electronics12234763

[5] Vetter, V., Wohlgenannt, P., Kepplinger, P., & Eder, E. (2025). Deep reinforcement learning approaches the MILP optimum of a multi-energy optimization in energy communities. Energies, 18(17), 4489. https://doi.org/10.20944/preprints202508.0033.v1

[6] Sun, Z., & Lu, T. (2024). Collaborative operation optimization of distribution system and virtual power plants using multi-agent deep reinforcement learning with parameter-sharing mechanism. IET Generation, Transmission & Distribution, 18(1), 39-49. https://doi.org/10.1049/gtd2.13037

[7] Li, Y., Chang, W., & Yang, Q. (2025). Deep reinforcement learning based hierarchical energy management for virtual power plant with aggregated multiple heterogeneous microgrids. Applied Energy, 382, 125333. https://doi.org/10.1016/j.apenergy.2025.125333

[8] Aoun, A., Adda, M., Ilinca, A., Ghandour, M., & Ibrahim, H. (2024). Optimizing virtual power plant management: A novel MILP algorithm to minimize levelized cost of energy, technical losses, and greenhouse gas emissions. Energies, 17(16), 4075. https://doi.org/10.3390/en17164075

[9] Guo, B., Li, F., Yang, J., Yang, W., & Sun, B. (2024). The application effect of the optimized scheduling model of virtual power plant participation in the new electric power system. Heliyon, 10(11). https://doi.org/10.1016/j.heliyon.2024.e31748

[10] Yoon, S. J., Ryu, K. S., Kim, C., Nam, Y. H., Kim, D. J., & Kim, B. (2024). Optimal Bidding Scheduling of Virtual Power Plants Using a Dual-MILP (Mixed-Integer Linear Programming) Approach under a Real-Time Energy Market. Energies, 17(15), 3773. https://doi.org/10.3390/en17153773

[11] Arévalo, P., Ochoa-Correa, D., Villa-Ávila, E., Iñiguez-Morán, V., & Astudillo-Salinas, P. (2025). Systematic Review of Hierarchical and Multi-Agent Optimization Strategies for P2P Energy Management and Electric Machines in Microgrids. Applied Sciences, 15(9), 4817. https://doi.org/10.3390/app15094817

[12] Guerra, P., Gil, E., & Hinojosa, V. H. (2024). Improving the Computational Efficiency of the Unit Commitment Problem in Hydrothermal Systems by Using Multi-Agent Deep Reinforcement Learning. IEEE Access, 12, 53266-53276. https://doi.org/10.1109/access.2024.3383442

[13] Michailidis, P., Michailidis, I., & Kosmatopoulos, E. (2025). Reinforcement learning for optimizing renewable energy utilization in buildings: A review on applications and innovations. Energies, 18(7), 1724. https://doi.org/10.3390/en18071724

[14] Liu, Z. (2015). Global energy interconnection. Academic Press. DOI: 10.1016/C2015-0-01255-2

[15] Chatzivasileiadis, S., Ernst, D., & Andersson, G. (2013). The global grid. Renewable Energy, 57, 372-383. https://doi.org/10.1016/j.renene.2013.01.032

[16] Darvishi, M., Tahmasebi, M., Shokouhmand, E., Pasupuleti, J., Bokoro, P., & Raafat, J. S. (2023). Optimal operation of sustainable virtual power plant considering the amount of emission in the presence of renewable energy sources and demand response. Sustainability, 15(14), 11012. https://doi.org/10.3390/su151411012

[17] Jendoubi, I., & Bouffard, F. (2023). Multi-agent hierarchical reinforcement learning for energy management. Applied Energy, 332, 120500. https://doi.org/10.1016/j.apenergy.2022.120500