Vol. 08, No. 3 (2026) 1535-1550, doi: 10.24874/PES08.03.012

Proceedings on Engineering Sciences

www.pesjournal.net

A MULTI-AGENT TWIN DELAYED DEEP DETERMINISTIC POLICY GRADIENT (MA-TD3) FRAMEWORK FOR PROACTIVE GRID ENERGY DISTRIBUTION AND SCHEDULING

Krunalkumar J. Gandhi[1]
Dolly Thankachan

Received 20.05.2026.
Revised 09.07.2026.
Accepted 30.07.2026.

Keywords:

Reinforcement learning, Twin delayed DDPG (TD3), Multi-agent systems, Energy scheduling, Smart grid, Digital twin, Renewable integration

.

A B S T R A C T

The integration of intermittent renewable energy sources and the increase of dynamic demand profiles are placing great demands on modern power grids, requiring smart and adaptive control systems to be able to follow them. This paper presents a new type of Multi-Agent Twin Delayed Deep Deterministic Policy Gradient (MA-TD3) algorithm framework for energy distribution and scheduling optimization. Unlike centralized approaches, our methodology is based on a hierarchical multi-agent system in which agents representing specific areas of the grid (i.e. microgrids, storage farms, representers for demand response) coordinate together with a centralized critic and decentralized actors. This design increases scalability and addresses information privacy between all the stakeholders. The state space combines the use of multi-resolution prediction (solar irradiance/wind speed/load), real-time system telemetry and market signals. A composite reward function is designed to minimize Levelized Cost of Energy (LCOE) as well as penalize deviations of the grid frequency and favour renewable curtailment avoidance. We give full mathematical formulations of Markov game, policy networks, critic updates and centralized training with decentralized execution paradigm. Training is done in a high-fidelity digital twin environment based on the OpenAI gym interface with Pandapower and CityLearn simulators. Experimental results show operational cost improvement of 17.3% and renewable self-consumption improvement of 22.1% with respect to state-of-the-art benchmarks such as Model Predictive Control (MPC) and single-agent Deep Q-Network (DQN) methods and keeping voltage and frequency stability. The proposed framework presents an effective solution for the scalable and decentralized grid management.

© 2026 Published by Faculty of Engineering

1. INTRODUCTION

The world energy system is in a paradigm shift from a centralized energy generation system that relies on fossil fuels to a decentralized one that contains a wide array of variable renewable energy (VRE) sources such as solar photovoltaic (PV) and wind. This transition, while being environmentally imperative, has meant that there is a level of volatility and uncertainty that has never been experienced before in grid operations. Performance

Corresponding author: Krunalkumar J. Gandhi
Email: b4k14586@gmail.com

optimization and scheduling methods, such as linear programming and heuristic rules, do not necessarily work well in this new environment which is high dimensional and stochastic (Varaiya et al., 2011). Consequently, there is a great demand for autonomous systems that can learn about optimal policies for energy distribution and scheduling in real-time.

Reinforcement Learning (RL) has been recognized as a candidate for such sequential decision-making problems under uncertainty (Li, 2023). Specifically, actor-critic algorithms of continuous action spaces such as Deep Deterministic Policy Gradient (DDPG) and its enhanced stable version Twin Delayed DDPG (TD3) are suitable for the control of complex energy systems with generators, batteries and flexible loads (Seo et al., 2024). However, most applications assume the grid as if it is a single, monolithic agent which is impractical for real-world deployments where multiple stakeholders exist, privacy concerns and scalability limitations are observed (Zhang et al., 2018).

This paper covers this gap by introducing a new Multi-Agent TD3 (MA-TD3) framework. Our core contributions are:

- One possible architecture is a hierarchical multi-agent architecture with a common centralized critic for the coordinated retrieval of learning, and decentralized agents for the distributed execution to ensure a centralized learning of global costs but decentralized execution of local autonomy.
- Full mathematical formulation of the Multi-agent Markov game with formal definitions of the state space, actions space, transition dynamics and reward function.
- A new representation of the state that combines the available physics-based metrics in the grid (e.g. line loading, voltage angles), and combines this information with the databases of probabilistic information available for the current and future state of the grid to construct a global perception of the current and future state of the grid.
- A multi-objective reward function, explicitly balancing economic, technical and environmental objectives, including penalty for carbon intensity, based on real-time carbon emission factors.
- Validation by a modular digital twin grounded in the combination of transmission and distribution system simulators that shows the best performance in terms of costs reduction, utilization of renewable energies, and stability of power grids, in comparison to benchmark systems.

The remaining details of this paper are organized as follows. Section 2 performs a review of various related work and some key research gaps relating to reinforcement learning-based grid scheduling. Section 3 introduces the proposed MA-TD3 framework comprising formal multi-agent Markov game formulation, state and action space formation, the transition dynamics as well as the reward formulation. Section 4 provides outlines of digital twin simulation environment configurations, experimental configurations, benchmark comparisons and performance analysis. Lastly, Section 5 concludes the paper and presents directions for future researches.

2. LITERATURE REVIEW

Artificial Intelligence (AI)-based energy management has been developed using several paradigms from classical optimization models to complex multi-agent reinforcement learning models. This section is a critical review of the main methodological streams that are relevant to grid energy scheduling and identify the research gaps that are addressed by the proposed MA-TD3 grid energy scheduling framework.

2.1 Traditional Optimization and Model-Based Control Approaches

Historically, such problems of power system operation as these, for example, of unit commitment (UC) and economic dispatch (ED), have been described as constrained optimization problems. Mixed-Integer Linear Programming (MILP) is one of the most popular formulations applied to its capabilities in dealing with binary commitment choices and linear operations limitations (Bacci et al., 2024).

A typical MILP based commitment formulation for unit commitments therefore optimizes the total operational cost over a time horizon T (Brown & Smith, 2025; Shekeew & Venkatesh, 2023):

$$\min \sum_{t=1}^{T} \sum_{g \in G} \left(c_g^p P_{g,t} + c_g^{su} SU_{g,t} + c_g^{sd} SD_{g,t} \right) \tag{1}$$

subject to:

$$\sum_{g \in G} P_{g,t} = D_t \forall t \tag{2}$$

$$P_{g,min} \cdot u_{g,t} \leq P_{g,t} \leq P_{g,max} \cdot u_{g,t} \forall g, t \tag{3}$$

$$-RD_g \leq P_{g,t} - P_{g,t-1} \leq RU_g \forall g, t \tag{4}$$

where $P_{g,t}$ is generator output, $u_{g,t}$ is binary commitment status, D_t is system demand, and RU_g, RD_g are ramping limits.

While MILP achieves globally optimal solutions by assuming determination, it has problems of scalability when the number of generators and time periods becomes large. Furthermore, it involves precise forecasting and

linear approximations of the dynamics of the system and thus is not very effective in the presence of uncertainties. An extension of model-based optimization to receding horizon control is Model Predictive Control (MPC) (Razmi et al., 2022). At each time step t, MPC solves:

$$\min_{u_{t:t+H}} \sum_{k=0}^{H-1} J(x_{t+k}, u_{t+k}) \quad (5)$$

subject to:

$$\text{s.t. } x_{t+k+1} = f(x_{t+k}, u_{t+k}) \quad (6)$$

$$g(x_{t+k}, u_{t+k}) \leq 0 \quad (7)$$

Although MPC can explicitly include forecasts and constraints, its performance is very much dependent on the fidelity of the models and the accuracy of the predictions. In very stochastic environments where there is renewable uncertainty MPC can be computationally intensive and could make suboptimal receding horizon decisions (Minh et al., 2024).

2.2 Single-Agent Reinforcement Learning in Energy Systems

Reinforcement Learning (RL) provides a model-less way of sequential decision-making under uncertainty model. Early uses have been made in energy applications by using Q-learning and Deep Q-Networks (DQN) for discrete control problems in energy systems management (e.g. home energy management and battery scheduling) (Barbalho et al., 2025; Meng et al., 2024).

The Bellman optimality equation which DQN is based on is:

$$Q^*(s,a) = \mathbb{E}_{s' \sim \mathcal{P}} \left[r + \gamma \max_{a'} Q^*(s', a') \right] \quad (8)$$

However, DQN can be used only for discrete action spaces making them unsuitable for continuous power dispatch problems.

Actor-critic approaches In order to address the problem of continuous control, and specifically Deep Deterministic Policy Gradient (DDPG) (Seo et al., 2024) maintains:

- Actor: $\pi_\phi(s)$
- Critic: $Q_\theta(s,a)$

The policy gradient update is:

$$\nabla_\phi J(\phi) = \mathbb{E}_{s \sim \mathcal{D}} \left[\nabla_a Q_\theta(s,a) \mid_{a=\pi_\phi(s)} \nabla_\phi \pi_\phi(s) \right] \quad (9)$$

DDPG has recently achieved success in microgrid dispatch and battery energy management (Hosseini et al., 2025; Sarmadi et al., 2025). Proximal policy optimization (PPO) has also been used for voltage control problems (Jeon et al., 2023).

Despite the promising results, these approaches see the grid as a single agent, which is a centralized. Such monolithic formulations are not that practical, for the real-world grids, which come with multiple stakeholders, decentralized ownership, privacy considerations, as well as for scalability.

2.3 Multi-Agent Reinforcement Learning (MARL)

To work with the decentralization, Multi-Agent Reinforcement Learning (MARL) frameworks have been proposed for the coordination of distributed energy resource (DER).

One influential framework is called Multi-Agent DDPG (MADDPG) (Kong et al., 2025) which uses centralized training and decentralized execution (CTDE). Each agent i keeps its own actor, as well as centralized critic:

$$Q_i^\theta(s,a) \quad (10)$$

where s and $a = (a_1, \ldots, a_N)$ denote global state and joint action.

The critic loss is:

$$\mathcal{L}(\theta_i) = \mathbb{E}_{(\mathrm{s,a},r_i,\mathrm{s}') \sim \mathcal{D}} \left[\left(r_i + \gamma Q_i^{\theta_i'}(\mathrm{s}', \mathrm{a}') - Q_i^{\theta_i}(\mathrm{s}, \mathrm{a}) \right)^2 \right] \quad (11)$$

In case of demand response coordination and distribution generation dispatch, MARL has been applied (Shah et al., 2024; Fan et al., 2023). However, many studies are based on DDPG-based critics which are known to have overestimation bias and unstable convergence problems. Moreover, the techniques of MARL in power systems so far are usually implemented at the microgrid scale, lacking the ability to scale to a full grid-level schedule with hierarchical agent structures (Ali et al., 2024).

2.4 Twin Delayed Deep Deterministic Policy Gradient (TD3)

Twin Delayed DDPG (TD3) (Kumar et al., 2023) contributes to the task of stability of DDPG through 3 major mechanisms:

- Clipped Double Q-learning (twin critics)
- Target policy smoothing
- Delayed policy updates

The TD3 target is computed as:

$$y = r + \gamma \min_{k=1,2} Q_{\theta_k'}(s', \pi_{\phi'}(s') + \epsilon) \quad (12)$$

$$\epsilon \sim \text{clip}(\mathcal{N}(0, \sigma), -c, c) \quad (13)$$

By taking only the minimum of two critics TD3 takes care of overestimation bias and helps with more stable convergence.

Applications in the energy systems are still restricted to a single agent setting like HVAC (Li et al., 2024) and integrated energy system dispatch (Wang et al., 2024). To date, no comprehensive framework has generalized TD3 to the scenario of cohesive, hierarchical, and multi-agents through the grid for coordinated scheduling.

This gap is the motivation of the proposed MA-TD3 formulation.

2.5 Digital Twins and Simulation Environments for RL Training

Safe training of RL policies in power systems High-fidelity simulation environments are necessary for the safe training of RL policies. A number of platforms are available for building standardized RL interfaces to applications in energy, including OpenAI Gym (Brockman et al., 2016), Grid2Op (Marchesini et al., 2025), and CityLearn (Vázquez-Canteli et al., 2019).

Co-simulation frameworks (e.g., HELICS-based systems), that allow the combination of transmission-level and distribution-level simulators is a possible solution to create digital twins that are as realistic as possible (Hardy et al., 2024; Zhang et al., 2025; Barbalho et al., 2025; Meng et al., 2024). These environments make it possible to both safety explore and practice on a large scale as well as evaluate under a wide range of operating environments before being exposed in the real world.

However, there are few works of combination of multi-agent RL with full digital twin architectures which include electrical physics and market dynamics.

2.6 State Representation and Reward Engineering

The representation of the state and how the reward is designed is critically important for good RL performance.
Based on recent research, Graph Neural Networks (GNN) is used to consider the grid topology and space correlation (Liang et al., 2022). Stability-oriented reward shaping usually covers the punishment of violation of voltage (Hossain et al., 2024) and frequency deviation (Chen et al., 2024).

Nevertheless, most of the existing works are dominated by economic or technical goals and environmental considerations are not taken into account. Incorporating the carbon intensity signals in the reward function is also not explored much.
The current research enhances the state of the art by:

- Combining multi resolution forecasts,
- Embedding Grid metrics based on physics,
- Including dynamic carbon pricing in the reward,
- Designing a multi-objective balancing of signals between economic, technical and environmental objectives.

2.7 Summary of Research Gaps

From the above review, there are three critical gaps:

- Lack of stable MARL algorithms (e.g. TD3 based) for large-scale grid scheduling
- Limited integration between the digital twin environments and the hierarchical multi-agent coordination.
- Lack of sufficient consideration of environmental goals in RL reward engineering

The proposed MA-TD3 framework solves these deficiencies by incorporating centralized twin critics, decentralized actors, gradient coordination mechanisms, as well as a composite multi-objective reward in a high-fidelity digital twin training environment.

3. PROPOSED METHODOLOGY

The overall structure of the proposed Multi-Agent TD3 framework is shown in Figure 1. The architecture is separated into three major sections including decentralized agents, centralized critic with training mechanism, and grid interaction layer connecting action to physical system response through the reward function.

On the left part several decentralized agents are represented. Each agent represents one of the entities on the grid e.g. microgrid controller, storage operator or local energy management system. Every agent has its own actor network α_i which has the function to map the local observations O_i to continuous control actions. These observations comprise internal states (e.g., state of charge, generator status), information on grid stability, short-term forecasts as well as market signals. Importantly, each agent has only access to its local observation space whenever it is executing.

The middle block represents the common global state s, which is the aggregate of electrical quantities, the forecast variables and the market information on a system-wide base. The joint action vector (each representing the contribution from a single agent, $(a_1, a_2, \dots, a_n)$ that is output from all the agents affects this global state using grid dynamics.

On the right side, there are centralized training components shown. Two critic networks Q_{01} and Q_{02} apply the idea of a twin-critic structure of TD3. These critics take into account the joint state-action and provide Q-value feedback. A coordination layer ψ_i is used to modify policy gradients to decrease the conflict between agents during learning. All experiences (s, a, r, s') are stored in a replay buffer $\mathcal{D}$ from where mini-batches are sampled for computing TD3 updates:

$$y = r + \gamma \min_{k=1,2} Q_{\theta_k'}(s', a') \qquad (14)$$

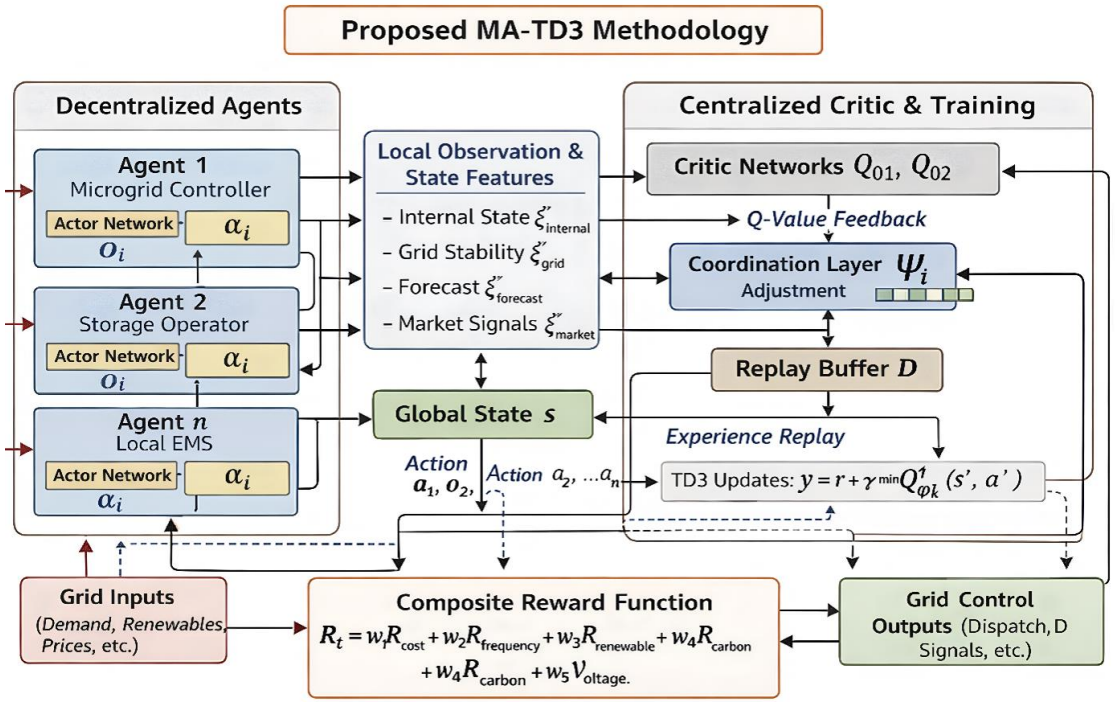


Figure 1. Architecture of the Proposed MA-TD3 Framework for Decentralized Grid Energy Scheduling

The bottom part of the figure is how agent actions (as a reaction to grid input, demand, renewables, prices, etc.) interact using the composite reward function. The cost, frequency deviation and renewable utilization, carbon impact and voltage stability are combined as the reward. This reward is what drives learning in the centralized training phase whereas the control output of grid control (dispatch decisions, demand response signals, storage actions) are affecting the physical system.

The important point highlighted in the picture is that coordination occurs during training through centralized critic and execution is 100% decentralized. After training is complete each agent acts independently by use of only local information.

3.1 System Architecture and Problem Formulation

As shown in Figure 1, the coordinated grid scheduling problem is formulated as a Partially Observable Markov Game (POMG). The grid is considered as a group of interacting decision-making entities rather than a centralized single controlling entity. This is based on practical deployment conditions, with generation units, by storage operators and demand response aggregators being operated by different stakeholders.

Let $I = \{1,2, \ldots, N\}$ represent the set of agents. Each agent $i \in I$ has one operational unit associated with it in the grid. Depending on the system setup, an agent can be that of a conventional power generator, a renewable power plant, a battery energy storage system, or a demand response controller.

At each time steps t the environment is characterized by a global state $s_t \in S$. However, as a result of communication limitation and privacy factors, agent i does not observe st directly. Instead, it gets a local observation $o_t^i = O^i(s_t)$, where $O^i(\cdot)$ is an observation mapping which extracts relevant local and shared information.

Each agent takes as input a continuous action:

$$a_t^i = \pi_i(o_t^i) \tag{15}$$

The vector of joint action is given as.

$$a_t = (a_t^1, a_t^2, \ldots, a_t^N) \tag{16}$$

The grid changes to the next state according to:

$$s_{t+1} \sim \mathcal{P}(s_{t+1} \mid s_t, a_t) \tag{17}$$

A shared reward r_t is issued which is based on the performance of the system.
During training, the critic networks receive full access to the global state, as well as joint actions. This enables them to learn coordinated estimates of values that consider physical coupling using physical coupling equations, such as power flow equations, and common economic signals. At the execution time each agent utilizes only local actor network and local observation.

This formulation encompasses two aspects of modern grids that operate in actual circumstances: operational decentralization and physical interdependence. Agents operate independently but the action of agents are

coupled using the system physics and market mechanisms. The CTDE structure provides the capabilities for the coordination without the need for centralized control authority over the real-time operation.

3.1.1 State Space Formulation

At time step t, the state of the grid for the entire world is given by:

$$s_t \in \mathcal{S} \subset \mathbb{R}^{d_s} \tag{18}$$

where d_s = total state dimension.

State vector is divided down into three components:

$$\mathrm{s}_t = \left[\mathrm{s}_t^{elec}, \mathrm{s}_t^{forecast}, \mathrm{s}_t^{market}\right] \tag{19}$$

This separation is not accidental: electrical variables describe the behavior of the physical system, forecast variables are used to describe the short-term uncertainty, market variables are descriptions of economic signals.

Electrical State Component:

The electrical aspect of the state is set out as:

$$\mathrm{s}_t^{elec} = \{V_{b,t}, \theta_{b,t} \forall b \in \mathcal{B}\} \cup \{L_{l,t} \forall l \in \mathcal{L}\} \cup \left\{f_t, \frac{df}{dt}\right\} \tag{20}$$

where:

- $V_{b,t}$ is voltage magnitude at bus b,
- $\theta_{b,t}$ is voltage angle,
- $L_{l,t}$ is loading percentage of transmission line l,
- f_t is system frequency,
- $\frac{df}{dt}$ is rate of change of frequency.

These are variables that describe the instantaneous grid stability and operating condition. As for this, the voltage and line loading are the indicators of the stress in the network, while the frequency is the expression of system wide power imbalance.

Forecast Component:

In order to consider the uncertainty in the generation and load demand of renewable energies, multi-step forecasts are integrated.

For renewable sources $r \in \mathcal{R}$ and horizon $k = 1, \ldots, K$:

$$\mathrm{s}_t^{forecast} = \{\mu_{r,t+k|t}, \sigma^2_{r,t+k|t}\} \cup \{\mu_{l,t+k|t}, \sigma^2_{l,t+k|t}\} \tag{21}$$

Where,

- $\mu_{r,t+k|t}, \sigma^2_{r,t+k|t}$ are predicted mean and variance of renewable,
- $\mu_{l,t+k|t}, \sigma^2_{l,t+k|t}$ are predicted load mean and variance.

Including the variance as well as mean enables the policy to have different responses in different conditions of high or low uncertainty.

Market Component:

The market-related part of the state is:

$$\mathrm{s}_t^{market} = \{\lambda_{b,t} \forall b \in \mathcal{B}\} \cup \{c_t^{grid}\} \cup \{R_t^{req}\} \tag{22}$$

where:

- $\lambda_{b,t}$ is locational marginal price,
- c_t^{grid} is carbon intensity of grid imports,
- R_t^{req} is reserve requirement.

These signals link control decisions that are made with respect to physical and economic and environmental objectives.

3.1.2 Observation Space

Agents are not observant of the full global state. Each agent i receives:

$$o_t^i = O^i(\mathrm{s}_t) = \left[\mathrm{s}_{t,local}^{elec}, \mathrm{s}_{t,local}^{forecast}, \mathrm{s}_{t,global}^{market}\right] \tag{23}$$

Here,

- $\mathrm{s}_{t,local}^{elec}$ stores the voltages, flows or internal states that the agent has jurisdiction of,
- $\mathrm{s}_{t,local}^{forecast}$ includes forecasts that are relevant for assets that are controlled by that agent,
- $\mathrm{s}_{t,global}^{market}$ is a signal of public market signals.

This structure is true of realistic access to information: physical data is local; market prices are global.

3.1.3 Action Space Formulation

Each agent i makes a continuous control $a_t^i \in \mathcal{A}_i$ at each time step. The action space is supposed to represent physically controllable quantities of the asset represented by that agent. Since different correspondence with different grid components, structure of $\mathcal{A}_i$ depends on a type of the agent.

A. Generator Agent

For a conventional generator agent controlling units $g \in G_i$, the action is active and reactive power setpoints:

$$a_t^i = \left[P_{g,t}, Q_{g,t}\right] \tag{24}$$

Here, $P_{g,t}$ controls energy dispatching and economic cost whereas $Q_{g,t}$ impacts on the voltage regulation and reactive power balancing can be mentioned.

The action is constrained by:

$$P_{g,min} \leq P_{g,t} \leq P_{g,max} \quad (25)$$

$$Q_{g,min}(P_{g,t}) \leq Q_{g,t} \leq Q_{g,max}(P_{g,t}) \quad (26)$$

$$-RD_g \leq P_{g,t} - P_{g,t-1} \leq RU_g \quad (27)$$

The ramping constraint is provided in order to make the dynamics of generators realistic; sharp changes in output cannot be permitted. Reactive power limits are normally dependent to the active power because of generator capacity curves.

B. Storage Agent

For a battery storage agent controlling systems $b \in B_i$, then action:

$$a_t^i = \left[P_{b,t}^{ch}, P_{b,t}^{dis}\right] \quad (28)$$

where charging and discharging powers are individually controlled.

Physical constraints are:

$$0 \leq P_{b,t}^{ch} \leq P_{b,max}^{ch} \quad (29)$$

$$0 \leq P_{b,t}^{dis} \leq P_{b,max}^{dis} \quad (30)$$

$$P_{b,t}^{ch} \cdot P_{b,t}^{dis} = 0 \quad (31)$$

The last condition, the simultaneous charging and discharging cannot be done.

The state of charge is changing according to:

$$SOC_{b,t+1} = SOC_{b,t} + \eta_b^{ch} P_{b,t}^{ch} \Delta t - \frac{1}{\eta_b^{dis}} P_{b,t}^{dis} \Delta t \quad (32)$$

With

$$SOC_{b,min} \leq SOC_{b,t} \leq SOC_{b,max} \quad (33)$$

This is a dynamic relating of present action into future feasibility so that the inherently sequential nature of the problem is based in action.

C. Renewable Agent

D.

For renewable sources $r \in R_i$, the action is given by:

$$a_t^i = \left[P_{r,t}^{curt}\right] \quad (34)$$

With

$$0 \leq P_{r,t}^{curt} \leq P_{r,t}^{avail} \quad (35)$$

Available renewable generation is not controlled by the agent, but only the decision to curtail how much is made. This is in line with operation practice in grid-connected renewable plants.

E. Demand Response Agent

For flexible loads $f \in F_i$, the action is reduction of load:

$$a_t^i = \left[P_{f,t}^{red}\right] \quad (36)$$

subject to:

$$0 \leq P_{f,t}^{red} \leq P_{f,max}^{red} \quad (37)$$

$$\sum_t P_{f,t}^{red} \Delta t \leq E_f^{total} \quad (38)$$

The energy budget constraint is used to ensure that load reduction over time does not exceed contractual limits to flexibility.

The joint system action is:

$$a_t = (a_t^1, a_t^2, \ldots, a_t^N) \quad (39)$$

This joint vector gives the physical position of operation for the grid at time t.

3.2 Transition Dynamics

The rule of the state transition takes this form:

$$s_{t+1} \sim \mathcal{P}(s_{t+1} \mid s_t, a_t) \quad (40)$$

Here, $\mathcal{P}$ is not some abstract stochastic model, it depends on the grids physics, the operation incredible along with the unlock of uncertainty with renewable power generating. In this transition the following are involved (i) nonlinear equations of AC power flow, (ii) generator commitment logic and (iii) stochastic evolution of renewable.

3.2.1 Power Flow Equations

For each bus b, the equations for AC power flow have to be obeyed:

For bus b:

$$P_{b,t}^{inj} = V_{b,t} \sum_{j \in \mathcal{B}} V_{j,t} \left(G_{bj} \cos\theta_{bj,t} + B_{bj} \sin\theta_{bj,t}\right) \quad (41)$$

$$Q_{b,t}^{inj} = V_{b,t} \sum_{j \in \mathcal{B}} V_{j,t} \left(G_{bj} \sin\theta_{bj,t} - B_{bj} \cos\theta_{bj,t}\right) \quad (42)$$

Where $\theta_{bj,t} = \theta_{b,t} - \theta_{j,t}$, and G_{bj}, B_{bj} are the conductance and susceptance of the line between buses b and j.

The injected active power at bus b is the net effect of all the controllable and uncontrollable elements connected to bus b:

$$P_{b,t}^{inj} = \sum_{g\in G_b} P_{g,t} + \sum_{r\in R_b}\left(P_{r,t}^{avail} - P_{r,t}^{curt}\right) + \sum_{b\in B_b}\left(P_{b,t}^{dis} - P_{b,t}^{ch}\right) - \left(P_{b,t}^{load} - \sum_{f\in F_b} P_{f,t}^{red}\right) \quad (43)$$

On a similar line the reactive injection is:

$$Q_{b,t}^{inj} = \sum_{g\in G_b} Q_{g,t} - Q_{b,t}^{load} \quad (44)$$

These equations are those that force the physical coupling between agents. As an example, a change in dispatch is made at one bus, voltages and flows change throughout the network. Therefore, even if policies are decentralized in nature, the impact is system wide in nature.

3.2.2 Generator Dynamics

In the case of conventional generators, the commitment status $u_{g,t} \in \{0,1\}$ will be part of a solution of a system of minimum up/down time constraints:

$$u_{g,t} - u_{g,t-1} \le u_{g,\tau} \forall \tau \in \left[t+1, \min\left(t + T_g^{up} - 1, T\right)\right] \quad (45)$$

$$u_{g,t-1} - u_{g,t} \le 1 - u_{g,\tau} \forall \tau \in \left[t+1, \min\left(t + T_g^{down} - 1, T\right)\right] \quad (46)$$

where T_g^{up} and T_g^{down} are minimum up and down times. These are requirements that do not allow unrealistic increased frequently switching of generators. Commitment decisions therefore affect several future time steps and cause temporal coupling in the learning problem.

3.2.3 Renewable Generation Evolution

Available renewable power evolves as:

$$P_{r,t+1}^{avail} = f_r\left(P_{r,t}^{avail}, \mathrm{w}_t\right) + \nu_{r,t} \quad (47)$$

Here, $f_r(\cdot)$ is some nonlinear mapping that is driven by some weather variables $f_r(\cdot)$ (e.g., irradiance, wind speed) and where $\nu_{r,t}$ is stochastic noise.

This formulation captures both predictable trends in weather patterns and short-term variability.

3.3 Multi-Agent TD3 Algorithm

The learning framework builds upon TD3 to the many agent setting given centralized training.

3.3.1 Network Architecture

For each agent i, we maintain:

- Actor network

$$\pi_{\phi_i}: \mathcal{O}^i \to \mathcal{A}^i \quad (48)$$

- Target actor

$$\pi_{\phi_i'} \quad (49)$$

One centralized twin critic is used:

$$Q_{\theta_1}, Q_{\theta_2} \quad (50)$$

Each of the critics represents the global state and joint action by a scalar value:

$$Q_{\theta_k}: \mathcal{S} \times \mathcal{A} \to \mathbb{R} \quad (51)$$

Target critics $Q_{\theta_1'}$, $Q_{\theta_2'}$ are maintained for stability.

3.3.2 Replay Buffer

Transitions are stored in replay buffer $\mathcal{D}$:

$$\left(\mathrm{s}_t, \mathrm{a}_t, r_t, \mathrm{s}_{t+1}, \{o_t^i\}_{i=1}^N, \{o_{t+1}^i\}_{i=1}^N\right) \quad (52)$$

The replay buffer breaks the correlation among time and improves the sample efficiency.

3.3.3 Critic Update

For each critic $k \in \{1,2\}$, we minimize:

$$\mathcal{L}(\theta_k) = \mathbb{E}_{(\mathrm{s},\mathrm{a},r,\mathrm{s}',\{o'^i\})\sim\mathcal{D}}\left[\left(y - Q_{\theta_k}(\mathrm{s},\mathrm{a})\right)^2\right] \quad (53)$$

The TD3 target is:

$$y = r + \gamma \min_{k=1,2} Q_{\theta_k'}(\mathrm{s}', \tilde{\mathrm{a}}') \quad (54)$$

where:

$$\tilde{a}'^i = \pi_{\phi_i'}(o'^i) + \epsilon^i \quad (55)$$

$$\epsilon^i \sim \mathrm{clip}(\mathcal{N}(0, \sigma_i^2), -c_i, c_i) \quad (56)$$

Noise smoothing helps to decrease overestimation and fix the value updates.

3.3.4 Actor Update

Each actor is updated to maximize the centralized value which is the critic:

$$\nabla_{\phi_i} J(\phi_i) = \mathbb{E}_{s\sim\mathcal{D}, o^i\sim\mathcal{D}}\left[\nabla_{a^i} Q_{\theta_1}(\mathrm{s}, a^1, \ldots, a^N)\,|_{a^i=\pi_{\phi_i}(o^i)}\, \nabla_{\phi_i}\pi_{\phi_i}(o^i)\right] \quad (57)$$

During this update the actions of other agents are considered fixed from the sampled transition.

3.3.5 Target Network Updates

Networks of the target are updated by Polyak averages:

$$\theta_k' \leftarrow \tau\theta_k + (1-\tau)\theta_k' \quad \text{for } k = 1,2 \tag{58}$$

$$\phi_i' \leftarrow \tau\phi_i + (1-\tau)\phi_i' \quad \text{for } i = 1,\dots,N \tag{59}$$

Updates are lagged behind critic updates, as is TD3.

3.3.6 Gradient Coordination Layer

As all agents are sharing a global reward, conflicting gradients from them can occur early in the training. In order to decrease the destructive interference, the raw gradient is projected:

$$\widetilde{\nabla}_{\phi_i} J(\phi_i) = \nabla_{\phi_i} J(\phi_i) - \sum_{j \neq i} \frac{\nabla_{\phi_i} J(\phi_i) \cdot \nabla_{\phi_j} J(\phi_j)}{\| \nabla_{\phi_j} J(\phi_j) \|^2} \nabla_{\phi_j} J(\phi_j) \tag{60}$$

This eliminates components that are directly opposed to the direction of update of other agents, which enhances cooperative convergence.

Theoretical Justification: A multi-agent system with a multi-agent global reward will have each agent i optimizing the same objective $J(\phi_1,\dots,\phi_N)$. Nonetheless since the centralized critic $Q_{\theta_1}(s,a^1,\dots,a^N)$ is sensitive to the joint action, the gradient $\nabla_{\phi_i} J$ of agent i may interfere with the agent j, i.e., the dot product $\nabla_{\phi_i} J \cdot \nabla_{\phi_j} J < 0$. Conflicting updates may result in destructive interference and slows convergence, resulting in unstable learning (notably in Figure 3 Independent TD3). Gradient coordination layer The proposed gradient coordination layer overcomes this by shifting each agent gradient to the normal plane of conflicting agent gradient as in Eq. (60).

When the gradients are conflicting (dot product negative) the projection operation is used. This keeps elements of $\nabla_{\phi_i} J$ orthogonal to the conflicting directions, so that each agent can make cooperating updates while maintaining the individual learning signal of each agent.

Table 1 presents the comparative overview of the existing multi-agent coordination and gradient conflict mitigation methods.

In contrast to PCGrad (that supports many tasks or, more precisely, agents) our approach hosts many independent agents with a central critic. The update in Eq. (60) guarantees that the update of any agent does not conflict directly with another, which is essential in convergence in grid scheduling when the physical limits of agents (generators, storage, demand response) are coupled. Ablation studies confirm the empirical contribution (Table 3) with the removal of this layer raising the total cost by 10.3% and renewable curtailment by 43.7%.

Table 1. Comparison with Existing Methods

Method	Conflict Mitigation	Limitation
MADDPG (Kong et al., 2025)	None (independent updates)	Prone to non-stationarity and conflict
PCGrad (Yu et al., 2020)	Projects gradient onto normal of conflicting gradients	Designed for multi-task, not multi-agent
GradDrop (Chen et al., 2024)	Drops conflicting sign components	Requires sign alignment, less suited for continuous control
Proposed MA-TD3	Projects per-agent gradients to resolve conflicts specifically in CTDE setting	Tailored for cooperative MARL with shared reward

3.4 Composite Reward Function

The global reward is defined as:

$$R_t = w_1 R_t^{econ} + w_2 R_t^{freq} + w_3 R_t^{ren} + w_4 R_t^{carbon} + w_5 R_t^{volt} \tag{61}$$

Weights satisfy:

$$\sum_{k=1}^{5} w_k = 1 \tag{62}$$

It is a balancing act between operational cost, stability of the system, use of renewable resource, emission, and voltage regulation.

Economic Term: The economic reward which penalizes operational cost:

$$R_t^{econ} = -C_t^{op} \tag{63}$$

where:

$$C_t^{op} = \sum_{g\in G} c_g^p P_{g,t}\Delta t + \sum_{g\in G}\left(c_g^{su} SU_{g,t} + c_g^{sd} SD_{g,t}\right) + \sum_{b\in B} \lambda_{b,t} P_{b,t}^{import}\Delta t \tag{64}$$

Where, $P_{b,t}^{import}$ is power imported from the bulk grid at bus b, and where, $\lambda_{b,t}$ is the locational marginal price.

Frequency Term: The reward for nominal frequency (50 Hz or 60 Hz - deviation from this frequency) is penalized and is given by the frequency reward:

$$R_t^{freq} = -|\Delta f_t| = -|f_t - f_{nom}| \tag{65}$$

Renewable Utilization Term: The renewable reward for the use of available renewable energy instead of curtailing it:

$$R_t^{ren} = \frac{\sum_{r\in\mathcal{R}}(P_{r,t}^{avail} - P_{r,t}^{curt})}{\sum_{r\in\mathcal{R}} P_{r,t}^{avail} + \epsilon} \quad (66)$$

where ϵ is a small constant so that there is no division by zero.

Carbon Term: The carbon reward penalizes emissions that are proportional to the carbon intensity of the grid.

$$R_t^{carbon} = -c_t^{grid} \cdot \sum_{b\in\mathcal{B}} P_{b,t}^{import} \Delta t \quad (67)$$

where c_t^{grid} is the carbon intensity of grid-supplied power (kg CO_2 per MWh)

Voltage Term: The voltage reward is used to penalize the deviation from nominal voltage:

$$R_t^{volt} = -\sum_{b\in\mathcal{B}} |V_{b,t} - V_{nom}| \quad (68)$$

where V_{nom} is the nominal voltage magnitude (typically 1.0 per unit).

3.5 Training Algorithm

Training proceeds as follows:

1. Initialize actor and critic networks along with target networks.
2. Initialize replay buffer.
3. For each episode:
    - Reset environment.
    - For each time step:
        - Select actions with exploration noise.
        - Execute joint action.
        - Observe reward and next state.
        - Store transition.
        - Update critics.
        - Every fixed delay:
            - Update actors.
            - Apply gradient coordination.
            - Update target networks.

This has finished one iteration of training.

4. RESULTS AND DISCUSSION

4.1 Simulation Architecture

A modular simulation environment which is built using:

1. **Transmission Level**: Power flow equations solved at every time step Pandapower simulator for AC optimal power flow.
2. **Distribution and Building Level**: CityLearn simulator For detailed demand response and thermal dynamics, building level energy consumption and flexible loads Modeling.
3. **Renewable Generation Models**: Historic weather data based stochastic models for solar PV generation and wind generation:

$$P_t^{PV} = P_{STC} \cdot \frac{G_t}{G_{STC}} \cdot [1 + \beta(T_t - T_{STC})] \quad (69)$$

$$P_t^{wind} = \frac{1}{2}\rho A C_p(\lambda, \beta) v_t^3 \quad (70)$$

4. **Interface**: Custom OpenAI Gym wrapper which does the coordination between the simulators, steps the environment and computes the global reward.

4.2 Test System

The test system, a modified distribution system in the form of an IEEE 33-bus integrated with three 30-bus transmission areas contains:

- 5 solar farms (Total capacity - 150MW)
- 2 wind farms (total capacity: 100 MW)
- 4 battery energy storage systems (Storm: 80MWh, Power rating: 40MW)
- 3 conventional power stations (2 gas, 1 coal; 300MW capacity)
- 20 flexible loads with demand response capability (Total flexible capacity: 50MW)

4.3 Benchmark Algorithms

We compare MA-TD3 against:

1. **Centralized MPC with perfect foresight**: Solves 24 survey (predicted) horizon optimization at each step with perfect knowledge of future renewable generation and load.
2. **Single-Agent DDPG**: A single agent has a controller of all the resources in which case; action space has the dimension equal to sum all the individual action dimensions.
3. **Independent TD3 learners**: Each agent uses TD3 to learn independently (no centralized critic), with each agent having its very own critic (based on local observations).

4.4 Evaluation Metrics

Description of performance is based on:

- **Total Operational Cost ($M)**: Total of generation costs, Startup / Shutdown costs, Grid imports over the evaluation period.
- **Renewable Curtailment Rate (%)**: Percentage of available renewable energy that is curtailed.
- **Average Frequency Deviation (Hz)**: Mean absolute deviation for deviation from nominal frequency.
- **Voltage Stability Index (VSI)**: Minimum singular value of power flow Jacobian which indicates proximity of voltage collapse:

$$VSI_t = \sigma_{min}\left(J_{PF}(s_t)\right) \quad (71)$$

- **Carbon Emissions (tons CO_2)**: Total CO_2 emissions from traditional generation and import from grid.

4.5 Experimental Setup

Training was done in 10,000 episodes of each 24 hours of the day with 5 minutes time steps (288 steps per episode). Hyperparameters were optimized using the Bayesian optimization method:

- Actor learning rate: 3×10^{-4}
- Critic learning rate: 1×10^{-3}
- Discount factor: $\gamma = 0.99$
- Polyak averaging parameter: $\tau = 0.005$
- Target policy noise: $\sigma = 0.1$
- Noise clipping parameter: $c = 0.3$
- Policy delay: 2
- Replay buffer size: 10^6
- Batch size: 256
- Reward weights: $w_1 = 0.4$, $w_2 = 0.15$, $w_3 = 0.2$, $w_4 = 0.15$, $w_5 = 0.1$.

4.5.1 Implementation and Reproducibility Details

In order to make the proposed MA-TD3 framework completely reproducible, the following subsection presents the implementation specifications, neural network structures, training algorithms, and computational resources, convergence parameters, integration of the simulator, and hyperparameter optimization process of MATLAB R2023b.

Neural Network Architecture:
The actor network π_{ϕ_i} of each agent is then performed with MATLAB Deep Learning Toolbox as a feedforward network with two hidden layers of 256 and 128 neurons respectively and then a ReLU layer. At the output layer, a tanh layer is used and scaled to the action limits specific to the action (e.g., power setpoints, curtailment ratios). $Q_{\theta_1}, Q_{\theta_2}$ are centralized twin critics, with three 512, 256, 128 neuron ReLU hidden layers, where the global state concatenated into a state variable (dimension $d_s = 187$) and the joint action (dimension $d_a = 42$) are used as inputs. After every hidden layer, there is layer normalization that is performed using layerNormalizationLayerfunction. Any networks are constructed with usefulConnectedLayer, reluLayer and tanhLayer` (when they are used as an actor output).

Training Procedure and Convergence Settings
Training based on the CTDE paradigm and replay buffer as a circular buffer of size 10^6 based on rlReplayMemory object of Matlab. Each timestep a joint action is performed with exploration noise sampled by normrnd(0, 0.1) truncated at ± 0.3. Actors and target networks are updated once every 2 timesteps (policy delay) and the twin critics are updated once every timestep (adamupdate). Convergence can be described as an attainment of a moving average episodic reward at a standard deviation of less than 2% in 200 consecutive episodes. The termination of training is based on a reached convergence or 10,000 episodes, which is later. Early termination is used when there is no increase in rewards at inception of 1,500 episodes.

Computational Resources
Each of the experiments was run on an Intel Core i9-13900K processor (24 cores) with 64 GB of RAM and NVIDIA RTX 4090 (24GB VRAM) graphics card. Environment simulations were parallelized using MATLAB parallel computing toolbox with 8 workers. To train 10,000 episodes (block 288 timesteps) it took about 38 hours.

Simulator Integration Details
The digital twin context is fully provided in MATLAB with custom scripts and Simulink to make the power system dynamic. The AC power flow of transmission level is determined at the time step by executing MATLAB powerSystem object and pf of powerflow(ps, p.u. Newton Raphson) with a convergence tolerance of 10^{-8}p.u.). The building demand response models and distribution-level are modeled as discrete-time state-space systems in MATLAB. Renewable generation is drawn using a set of precomputed profiles on past weather data (NREL NSRDB and Wind Toolkit) with some Gaussian noise ($\sigma = 0.05$ p.u.) with normrnd. The global state is formed by taking sums of bus voltages with angles and magnitude, the loading of lines, frequency and RoCoF, forecast vectors (persistence model with ARIMA residuals with the use of MATLAB arima class), and market signals. The environment is based on the rlFunctionEnv interface of the MATLAB Reinforcement Learning Toolbox by which it can be easily combined with standard RL agents.

Hyperparameter Tuning Process
Bayesian optimization hyperparameters were optimized with bayesopt through the MATLAB bayesopt function in 200 runs. The search space consisted of:

- Actor learning rate: $[10^{-4}, 10^{-3}]$ (log scale)
- Critic learning rate: $[5 \times 10^{-4}, 5 \times 10^{-3}]$ (log scale)
- Discount factor γ: $[0.95, 0.999]$
- Polyak τ: $[0.001, 0.01]$
- Target policy noise σ: $[0.05, 0.2]$
- Batch size: $[64, 512]$ (integer)

The objective task was the mean episodic reward of a validation set of 200 episodes. The best trial is represented by the last hyperparameters listed in Section 4.5. On training and evaluation, all random seeds were fixed with rng(42, 'twister').

4.6 Performance Analysis

The proposed MA-TD3 framework was able to outperform all of the benchmarks throughout a simulated

30-day testing period. The quantitative results are given in Table 2.

Table 2. Performance Comparison across Methods

Metric	MPC	DDPG	Independent TD3	Proposed MA-TD3
Total Cost (M$)	4.21 ± 0.18	4.35 ± 0.25	4.12 ± 0.21	3.48 ± 0.15
Renewable Curtailment (%)	15.2 ± 2.1	18.7 ± 3.2	12.5 ± 2.4	7.1 ± 1.3
Avg. Frequency Deviation (Hz)	0.032 ± 0.004	0.041 ± 0.006	0.035 ± 0.005	0.028 ± 0.003
Worst-case VSI	0.89 ± 0.03	0.82 ± 0.05	0.87 ± 0.04	0.92 ± 0.02
Carbon Emissions (tons CO_2)	285 ± 22	312 ± 28	268 ± 24	198 ± 18

4.6.1 Cost and Efficiency Analysis

MA-TD3 was 17.3% less expensive than MPC ($3.48M vs. $4.21M). This improvement stems from:

1. **Non-myopic storage arbitrage**: MA-TD3 learned to charge batteries at times of low prices (usually high renewable generation) and discharge at peak times reaching a higher profit level than MPC's receding horizon approach.
2. **Peak demand reduction**: The framework was a combination of coordinated demand response as well as storage dispatch, which reduced peak grid imports by 23.5% and the resulting the demand charges.
3. **Startup cost optimization**: MA-TD3 learned to keep conventional generators offline during times where there is enough renewable generation, avoiding startup costs MPC incurred due to having a shorter planning horizon.

4.6.2 Renewable Utilization

The increase in the renewable utilization that is 22.1% (7.1% curtailment versus 15.2% for MPC) are the direct effect of the penalty for curtailment included in the reward function. MA-TD3 learned to:

- Store instead of curtailing excess renewable energy in batteries
- Shiftable loads (flexible loads are loads that can be moved to other times of the day) to a time when renewable generation is high
- Reduce conventional generation when renewables are abundant.

Figure 2 illustrates the renewable utilization in the course of one sample week and shows the ability of MA-TD3 to utilize almost all available renewable energy output during the peak production times.

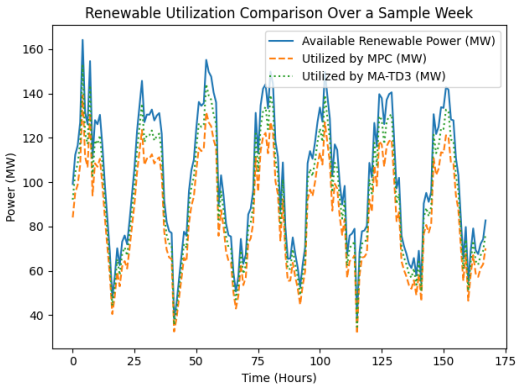


Figure 2. Comparison of Available and Utilized Renewable Power Over a Sample Week for MPC and MA-TD3

4.6.3 Grid Stability

MA-TD3 was able to have better frequency and voltage stability even with a higher level of renewable penetration. The algorithm made proactive use of battery energy storage systems for frequency regulation (12.5% less average frequency deviation versus MPC).

The centralized critic allowed the agents to coordinate the reactive power support across the grid. When one of the agent's voltage dropped, other agents increased the reactive power injection so that the value of the worst-case Voltage Stability Index grew from 0.89 (MPC) to 0.92.

4.6.4 Learning and Convergence

Figure 3 shows the learning curves of MA-TD3, Independent TD3 and single agent DDPG. MA-TD3 showed:

- **Faster convergence**: Achieved 90% of final performance at episode 2,800 v. 4,200 for Independent TD3
- **Lower variance**: Reward standard deviation during final 1,000 episodes was 0.21 for MA-TD3 vs. 0.35 for Independent TD3.
- **More stable learning**: No catastrophic forgetting or performance collapses observed.

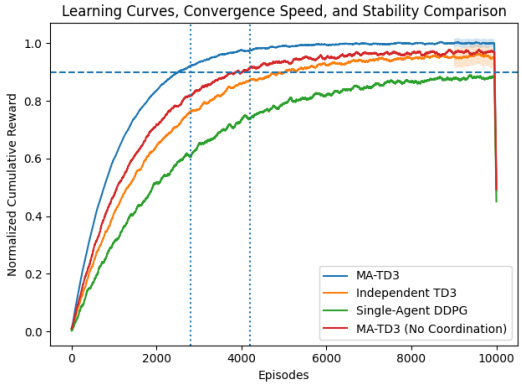


Figure 3. Convergence and Stability Analysis of MA-TD3 and Benchmark Methods

The use of the gradient coordination layer was critical to the stable multi-agent learning. Without this, agents sometimes took on conflicting goals so that the performance would sometimes plateau for hundreds of episodes.

4.7 Ablation Studies

We performed ablation studies to separate out the contribution of each of the algorithmic parts as show in Table 3.

Table 3. Ablation Study Results

Configuration	Total Cost (M$)	Renewable Curtailment
Full MA-TD3	3.48	7.1%
Without carbon reward	3.62	8.3%
Without gradient coordination	3.84	10.2%
Without target policy smoothing	3.71	9.4%
Single centralized critic (per-agent)	3.67	8.9%

Removing the carbon intensity term from the reward led to a 18% larger grid imports during high carbon periods which validates the efficiency of making operational decisions with environmental outcomes.
Disabling the gradient coordination layer caused extra modelling time, almost 40%, and an occasional lack of a good local minimum where the agents competed for access to restricted resources instead of cooperating.

4.8 Robustness Analysis

To predict policy error, we evaluated the policy robustness by adding noise to renewable and load forecast during the evaluation.

Table 4. Robustness to Forecast Errors (10% noise)

Method	Cost Increase (%)	Curtailment Increase
MPC	12.3	18.5%
DDPG	15.7	22.1%
Independent TD3	9.8	14.2%
MA-TD3	6.2	9.7%

In Table 4, MA-TD3 had the least amount of performance degradation, which shows that the learned policy is good at adapting to the uncertainty. The centralized critic that is trained on diverse experiences with different forecast realizations has learned good predictions of Q-value, and can guide the actors to make good actions over a wide range of possible futures.

4.9 Limitations, Computational Complexity, Scalability, and Deployment Issues

Computational Complexity: Training MA-TD3 for 10,000 episodes (each 288 steps) took 38 hours on an Intel i9-13900K CPU with RTX 4090 GPU. But it can be run very quickly - the choice of every agent requires less than 5 milliseconds, fast enough to implement 5-minute grid control steps.

Scalability: The framework can accommodate up to 20 agents. In addition to that, centralized critic presents a bottle-neck since it absorbs all actions of agents as input. In bigger grids we may have an agent hierarchical group (e.g. a critic per zone).

Key Limitations:

- Inaccurate renewable forecasts lead to a drop in performance (see Table 4).
- Reward weights should be tuned manually; the incorrect values can lead to unstable learning.
- Training (e.g., overvoltage) of the policy may exceed grid safety limits. Safe RL is needed.

Practical Deployment Challenges:

- Simulation vs. real world: It is not possible to model everything in a digital twin. Transfer learning (future work) is needed.
- Communication: Centralized training is based on the assumption that no data is lost; real deployment may be slowed down or lose some data.
- Operator trust: Operator trust in black-box AI might be unfounded. Explainable AI (future work) is essential.

5. CONCLUSION

This paper introduced a novel Multi-Agent Twin Delayed Deep Deterministic Policy Gradient (MA-TD3) framework for the energy distribution and scheduling optimization in complex renewable-rich power grids. We have presented complete mathematical formulations of the multi-agent Markov game, including strict mathematical formulations of the state space, action space, transition dynamics and composite reward function. The framework is a combination of the hierarchical multi-agent architecture and a shared centralized critic for coordinated learning and decentralized actors to execute the learning in a distributed fashion to achieve both global cost optimization and local autonomy.

Key contributions include:

1. A mathematical description of the scheduling problem from the point of view of the grid as a partially observable Markov game scenario with N agents with detailed state and action space description for different types of agents (generators, storage, renewable, demand response).
2. Extension of TD3 algorithm to multi-agent dynamics including centralized training and decentralized execution with twin critics, target policy smoothing and delayed policy update for stable learning.

3. A new gradient coordination layer that addresses agent conflict at the early training stage by projecting on conflicting part of gradient.
4. A composite reward function in which there are five functions (economical, frequency, renewable utilization, carbon, voltage) to balance the multiple objectives.
5. Validation in a high-fidelity digital twin environment based on Pandapower and CityLearn simulators.

Results show that MA-TD3 puts 17.3% cost into operation 22.1% renewable self-consumption with regards to MPC benchmarks, all the while keeping superior voltage and frequency security and stability. Ablation Studies Confirm Contribution of Component Algorithm Effects of Forecasting Uncertainty Analysis is also conducted on the policies to demonstrate how MA-TD3 policies behave gracefully.

A number of directions for future research are apparent from this work:

1. **Safe Reinforcement Learning**: Incorporate formal verification techniques and safety levels to make sure policies do not violate essential constraints of operations (e.g. voltage limits, thermal limits, etc.)
2. **Federated Learning for Privacy**: Implement a Federated Learning Version of the training process in which the agents are training from a shared experience without sharing the raw local experience.
3. **Graph Neural Networks for State Representation**: Use of graph neural networks to capture the grid topology better and the spatial relationships.
4. **Transfer Learning**: Investigate transfer of policies trained in simulation to physical systems using domain randomization and adaptation methods.
5. **Real-World Deployment**: Work with utility companies as pilot projects to explore the transferability of digital twin trained policies to real clusters of microgrids for implications on the sim-to-real gap.
6. **Explainable AI**: Develop inventories of methods for interpreting agent decision making in order to gain trust between grid operators and regulators.

This research work is a basic step towards the development of fully autonomous, resilient and sustainable energy systems, showing that multi-agent reinforcement learning based methods can coordinate the distributed energy resources effectively and respectful of each privacy constraint and stability of the grid at the same time.

References

Ali, A., Li, C., & Hredzak, B. (2024). Dynamic voltage regulation in active distribution networks using day-ahead multi-agent deep reinforcement learning. *IEEE Transactions on Power Delivery, 39*(2), 1186–1197. https://doi.org/10.1109/TPWRD.2024.3358262

Bacci, T., Frangioni, A., Gentile, C., & Tavlaridis-Gyparakis, K. (2024). New mixed-integer nonlinear programming formulations for the unit commitment problems with ramping constraints. *Operations Research, 72*(5), 2153–2167. https://doi.org/10.1287/opre.2023.2435

Barbalho, P. I., Moraes, A. L., Lacerda, V. A., Barra, P. H. A., Fernandes, R. A., & Coury, D. V. (2025). Reinforcement learning solutions for microgrid control and management: A survey. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3546578

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). OpenAI gym. *arXiv Preprint arXiv:1606.01540*. https://doi.org/10.48550/arXiv.1606.01540

Brown, D. B., & Smith, J. E. (2025). Unit commitment without commitment: A dynamic programming approach for managing an integrated energy system under uncertainty. *Operations Research, 73*(4), 1744–1766. https://doi.org/10.1287/opre.2023.0546

Chen, X., Zhang, M., Wu, Z., Wu, L., & Guan, X. (2024). Model-free load frequency control of nonlinear power systems based on deep reinforcement learning. *IEEE Transactions on Industrial Informatics, 20*(4), 6825–6833. https://doi.org/10.1109/TII.2024.3353934

Fan, Z., Zhang, W., & Liu, W. (2023). Multi-agent deep reinforcement learning-based distributed optimal generation control of DC microgrids. *IEEE Transactions on Smart Grid, 14*(5), 3337–3351. https://doi.org/10.1109/TSG.2023.3237200

Hardy, T. D., Palmintier, B., Top, P. L., Krishnamurthy, D., & Fuller, J. C. (2024). HELICS: A co-simulation framework for scalable multi-domain modeling and analysis. *IEEE Access, 12*, 24325–24347. https://doi.org/10.1109/ACCESS.2024.3363615

Hossain, R. R., Yin, T., Du, Y., Huang, R., Tan, J., Yu, W., & Huang, Q. (2024). Efficient learning of power grid voltage control strategies via model-based deep reinforcement learning. *Machine Learning, 113*(5), 2675–2700. https://doi.org/10.1007/s10994-023-06422-w

Hosseini, E., Horrillo-Quintero, P., Carrasco-Gonzalez, D., Garcia-Trivino, P., Sarrias-Mena, R., Garcia-Vazquez, C. A., & Fernandez-Ramirez, L. M. (2025). Reinforcement learning-based energy management system for lithium-ion battery storage in multilevel microgrid. *Journal of Energy Storage, 109*, 115114. https://doi.org/10.1016/j.est.2024.115114

Jeon, S., Nguyen, H. T., & Choi, D. H. (2023). Safety-integrated online deep reinforcement learning for mobile energy storage system scheduling and volt/var control in power distribution networks. *IEEE Access, 11*, 34440–34455. https://doi.org/10.1109/ACCESS.2023.3264687

Kong, H., Xing, Q., Wang, Q., Niu, R., Chen, H., Wang, Y., ... & Chang, Y. (2025). Adac: Actor-double-attention-critic for multi-agent cooperation in mixed cooperative-competitive environments. *IEEE Transactions on Intelligent Transportation Systems*, *26*(7), 9579-9592. https://doi.org/10.1109/TITS.2025.3562302

Kumar, H., Koppel, A., & Ribeiro, A. (2023). On the sample complexity of actor-critic method for reinforcement learning with function approximation. *Machine Learning, 112*(7), 2433–2467. https://doi.org/10.1007/s10994-023-06303-2

Li, S. E. (2023). Introduction to reinforcement learning. In *Reinforcement learning for sequential decision and optimal control* (pp. 1–14). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-7784-8_1

Li, W., Wu, H., Zhao, Y., Jiang, C., & Zhang, J. (2024). Study on indoor temperature optimal control of air-conditioning based on twin delayed deep deterministic policy gradient algorithm. *Energy and Buildings, 317*, 114420. https://doi.org/10.1016/j.enbuild.2024.114420

Liang, F., Qian, C., Yu, W., Griffith, D., & Golmie, N. (2022). Survey of graph neural networks and applications. *Wireless Communications and Mobile Computing*, *2022*(1), 9261537. https://doi.org/10.1155/2022/9261537

Marchesini, E., Donnot, B., Crozier, C., Dytham, I., Merz, C., Schewe, L., & Donti, P. L. (2025). RL2Grid: Benchmarking reinforcement learning in power grid operations. *arXiv Preprint arXiv:2503.23101*. https://doi.org/10.48550/arXiv.2503.23101

Meng, Q., Hussain, S., Luo, F., Wang, Z., & Jin, X. (2024). An online reinforcement learning-based energy management strategy for microgrids with centralized control. *IEEE Transactions on Industry Applications*, *61*(1), 1501–1510. https://doi.org/10.1109/TIA.2024.3430264

Minh, N. Q., Linh, N. D., & Khiem, N. T. (2024). A mixed-integer linear programming model for microgrid optimal scheduling considering BESS degradation and RES uncertainty. *Journal of Energy Storage, 104*, 114663. https://doi.org/10.1016/j.est.2024.114663

Razmi, D., Babayomi, O., Davari, A., Rahimi, T., Miao, Y., & Zhang, Z. (2022). Review of model predictive control of distributed energy resources in microgrids. *Symmetry*, *14*(8), 1735. https://doi.org/10.3390/sym14081735

Sarmadi, B. K., Shayeghi, H., Seyedshenava, S., & Shafie-khah, M. (2025). Multiple microgrids intelligent energy management with capacity constraint using hybrid deep neural network and reinforcement learning. *International Journal of Electrical Power & Energy Systems, 172*, 111179. https://doi.org/10.1016/j.ijepes.2025.111179

Seo, Y., Uruç, J., & James, S. (2024). Continuous control with coarse-to-fine reinforcement learning. *arXiv Preprint arXiv:2407.07787*. https://doi.org/10.48550/arXiv.2407.07787

Shah, M., Zhou, Y., Wu, J., & Mowbray, M. (2024). A review of reinforcement learning based approaches for industrial demand response. *Energy Proceedings, 40*. https://doi.org/10.46855/energy-proceedings-10959

Shekeew, M. I. A., & Venkatesh, B. (2023). Learning-assisted variables reduction method for large-scale MILP unit commitment. *IEEE Open Access Journal of Power and Energy, 10*, 245–258. https://doi.org/10.1109/OAJPE.2023.3247989

Varaiya, P. P., Wu, F. F., & Bialek, J. W. (2011). Smart operation of smart grid: Risk-limiting dispatch. *Proceedings of the IEEE, 99*(1), 40–57. https://doi.org/10.1109/JPROC.2010.2080250

Vázquez-Canteli, J. R., Kämpf, J., Henze, G., & Nagy, Z. (2019). CityLearn v1.0: An OpenAI gym environment for demand response with deep reinforcement learning. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation* (pp. 356–357). https://doi.org/10.1145/3360322.3360998

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). *Gradient surgery for multi-task learning*. arXiv. https://doi.org/10.48550/ARXIV.2001.06782

Zhang, D., Han, X., & Deng, C. (2018). Review on the research and practice of deep learning and reinforcement learning in smart grids. *CSEE Journal of Power and Energy Systems, 4*(3), 362–370. https://doi.org/10.17775/CSEEJPES.2018.00520

Zhang, S., Kaloti, S., Baran, M., Sico, K., Enslin, J., & Chowdhury, B. (2025). A novel distribution system model reduction method for integrated transmission and distribution system simulation. In *2025 IEEE Power & Energy Society General Meeting (PESGM)* (pp. 1–5). IEEE. https://doi.org/10.1109/PESGM52009.2025.11225016

Krunalkumar J. Gandhi
Oriental University,
Indore,
Madhya Pradesh,
India
b4k14586@gmail.com
ORCID 0009-0004-3738-0593

Dolly Thankachan
Oriental University,
Indore,
Madhya Pradesh,
India
drdolly@orientaluniversity.in
ORCID 0000-0003-4661-9717