

Article

Renewable Energy Integrated Power System Load Frequency Control Based on Multi-Agent Actor-Double-Critic Deep Reinforcement Learning

Xinxin Lv *, Xiaodong Wang, Yuxin Yan, Yuyang Weng and Zheng Ge 

School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China; 202522070549@mails.zstu.edu.cn (X.W.); 19012728750@163.com (Y.Y.); 2024333503006@mails.zstu.edu.cn (Y.W.); gezheng01@126.com (Z.G.)

* Correspondence: xinxinlvhhuc@163.com

Abstract

To achieve optimal performance of load frequency control (LFC), a data-driven scheme is proposed for renewable power systems in this paper. A multi-agent Actor-Double-Critic deep reinforcement learning approach is developed to ensure real-time scheduling that complies with system safety operation constraints within the multi-area LFC power system. For implementation, each individual controller only needs local information in its control area to deliver optimal control signals. A Self-Critic and Cons-Critic network is employed to improve the convergence speed during the multi-agent training process. Simulations on two-area and three-area LFC power systems are performed to verify and validate the analytical results. Comparisons with conventional PI and fuzzy PI controllers demonstrate that the presented approach effectively reduces training difficulties, guarantees the satisfaction of system safety constraints, and significantly improves the dynamic frequency regulation performance of the power system.

Keywords: data-driven; deep reinforcement learning; load frequency control; multi-area power system; renewable energy

1. Introduction

The global energy crisis poses a significant challenge, necessitating an accelerated energy transition. Establishing a power system that predominantly relies on renewable energy sources is crucial for ensuring national energy security strategies [1–3]. However, the integration of clean technologies such as wind power and photovoltaic generation into the power grid has diminished the system's frequency regulation capability and exacerbated oscillation issues [4,5]. To regulate grid frequency, an automatic excitation adjustment device is commonly used on generators, but it can increase equipment costs. The most widely utilized technique for frequency regulation is load frequency control (LFC) [6], which plays a vital role in ensuring the stable operation of the power system by strategically managing and adjusting generating units to maintain a harmonious equilibrium between the load and the system's generation capacity [7–9].

Currently, numerous pioneering researchers have focused on improving control schemes for LFC. Techniques such as PID control, model predictive control (MPC), and sliding mode control (SMC) are proposed for LFC to ensure compliance of the control region with LFC commands and to enhance the performance and stability of the power system [10–12]. However, these control methods are highly dependent on the data within



Academic Editor: Babatunde Olubayo

Received: 22 April 2026

Revised: 16 June 2026

Accepted: 17 June 2026

Published: 22 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

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

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

the model itself. Nevertheless, these control methods involve complex mathematical calculations and optimization processes in terms of parameter estimation and optimal parameter tuning [13,14]. Consequently, data-driven control approaches are proposed for a multi-area LFC power system.

With the increasing integration of renewable energy into power systems, the field of LFC faces significant challenges. The traditional centralized control approach, constrained by the inherent inertia of synchronous generators, may lack flexibility required to respond promptly to these challenges. While centralized control imposes minimal constraints on control signals to achieve optimal performance, it suffers from high computational complexity and a substantial computational burden, making it unsuitable for large-scale interconnected power systems with highly integrated subsystems. Furthermore, in multi-area interconnected systems, the control areas may belong to different entities, complicating the acquisition of global information. Therefore, there is a pressing need to design a distributed LFC method. In this way, a two-layer MPC in [15], distributed automatic LFC scheme in [16] and the error in feeder power flow with respect to scheduled value utilizing by the PI controller to estimate the change in power reference of all distributed energy resources units in [17] are studied. In multi-area interconnected systems, control domains may fall under the jurisdiction of different entities, complicating the acquisition of comprehensive global information. Consequently, there is an urgent need to devise a data-driven distributed LFC approach that not only responds to the dynamics and distributed characteristics of modern power systems but also acknowledges inherent limitations and constraints. In multi-area power systems, control domains may fall under the jurisdiction of different entities, complicating the acquisition of comprehensive global information. Consequently, there is an urgent need to devise a data-driven distributed LFC scheme that not only responds to the dynamics and distributed characteristics of modern power systems but also acknowledges inherent limitations and constraints.

In recent years, data-driven control approaches have demonstrated significant potential for coordinated frequency control in multi-area power systems [18–20]. The data-driven control approaches are designed to optimize a control objective directly from the system's observed responses. By employing an appropriate training strategy and defining an effective reward mechanism, data-driven control systems that leverage multi-agent reinforcement learning (RL) can achieve collaborative outcomes even in the absence of comprehensive communication among the control agents. In this way, a deep reinforcement learning model based on a continuous action domain is proposed to enhance the LFC performance of single-area power systems, in [21]. For multi-area interconnected power systems, a data-driven coordinated LFC method based on Multi-agent Deep Reinforcement Learning (MADRL) for LFC is introduced in [22]. In the MADRL framework, each agent, represented by a deep neural network, collaboratively adjusts the generated commands to achieve the global objectives of the multi-area power system. Reference [23] designed an Evolutionary Multi-agent Deep Meta-actor–critic (EMA-DMAC) algorithm, which introduces meta-reinforcement and evolutionary learning to enable rapid collaborative learning of group intelligence, thus improving the robustness and quality of the obtained LFC scheme. Moreover, a hierarchical reinforcement learning method based on the decomposition of value functions, allowing for strategy optimization at multiple levels and thus achieving hierarchical task-solving is presented [24,25]. However, the above-proposed methods demand higher computational resources and incur significant communication overhead.

Recent studies have proposed various constrained and safe reinforcement learning frameworks, such as Lagrangian-based methods, constrained policy optimization approaches, and safety-aware multi-agent reinforcement learning techniques. Furthermore,

several advanced MADRL algorithms have demonstrated promising scalability in large-scale multi-agent environments.

Motivated by the observation of the aforementioned problems and challenges faced by existing MADRL algorithms, such as lack of constraint consideration and limited processing capability for large-scale problems, a data-driven approach is proposed for renewable power system LFC scheme in this paper. Specifically, a multi-agent Actor-Double-Critic deep reinforcement learning scheme is developed such that real-time scheduling is ensured, which is required for complying with system safety operation constraints within the LFC; Self-Critic and Cons-Critic networks are used to calculate the action value and cost of agents. Compared to PI control and fuzzy PI control methods, the presented approach reduces training difficulties and mitigates the impact of sparse immediate rewards and safety constraint costs, improving the convergence speed in multi-agent training and ensuring real-time scheduling such that system safety operation constraints can be met.

The remaining parts of this paper are structured as follows: the multi-area LFC system with constrained Markov cooperative game model is introduced in Section II. In Section III, the multi-agent Actor-Double-Critic deep reinforcement learning is presented for multi-area LFC power systems. Section IV presents the simulation results and comparative analysis. The conclusion of this work is shown in Section V.

2. Problem Statement

2.1. Load Frequency Control System Modelling

The multi-area LFC flow, which is investigated employing multi-agent Actor-Double-Critic deep reinforcement learning method, is described in Figure 1. The parameters of the i -th control area are represented in Table 1. For the designed LFC power system, the system dynamics of i -th control area are represented as the following:

$$\Delta \dot{f}_i = -\frac{D_i}{M_i} \Delta f_i + \frac{1}{M_i} \Delta P_{mi} - \frac{1}{M_i} \Delta P_{di} - \frac{1}{M_i} \Delta P_{tie-i}, \quad (1)$$

$$\Delta \dot{P}_{mi} = -\frac{1}{T_{chi}} \Delta P_{mi} + \frac{1}{T_{chi}} \Delta P_{vi}, \quad (2)$$

$$\Delta \dot{P}_{vi} = -\frac{1}{R_i T_{gi}} \Delta f_i - \frac{1}{T_{gi}} \Delta P_{vi} + \frac{1}{T_{gi}} ACE_i, \quad (3)$$

$$ACE_i = \beta_i \Delta f_i + \Delta P_{tie-i}, \quad (4)$$

$$\Delta \dot{P}_{w-i} = -\frac{1}{T_{wind-i}} \Delta P_{w-i} + \frac{1}{T_{wind-i}} \Delta \Phi_{w-i}, \quad (5)$$

$$\Delta \dot{P}_{Bi} = -\frac{1}{T_{ES-i}} \Delta P_{Bi} + \frac{1}{T_{ES-i}} \Delta f. \quad (6)$$

2.2. Multi-Area LFC Power System with Constrained Markov Cooperative Game Model

In this paper, the MADRL is applied to enhance the generation speed of scheduling decisions. For the designed multi-area LFC power system, each control area is equipped with its own frequency controller, which is formulated within an MADRL framework. The system is divided into multiple areas, with an agent assigned to each area to make scheduling decisions for resource regulation. A multi-agent Actor-Double-Critic algorithm is designed to train the multi-agent system, ensuring that the scheduling decisions generated by the agents meet the safety and reliability requirements of system operation.

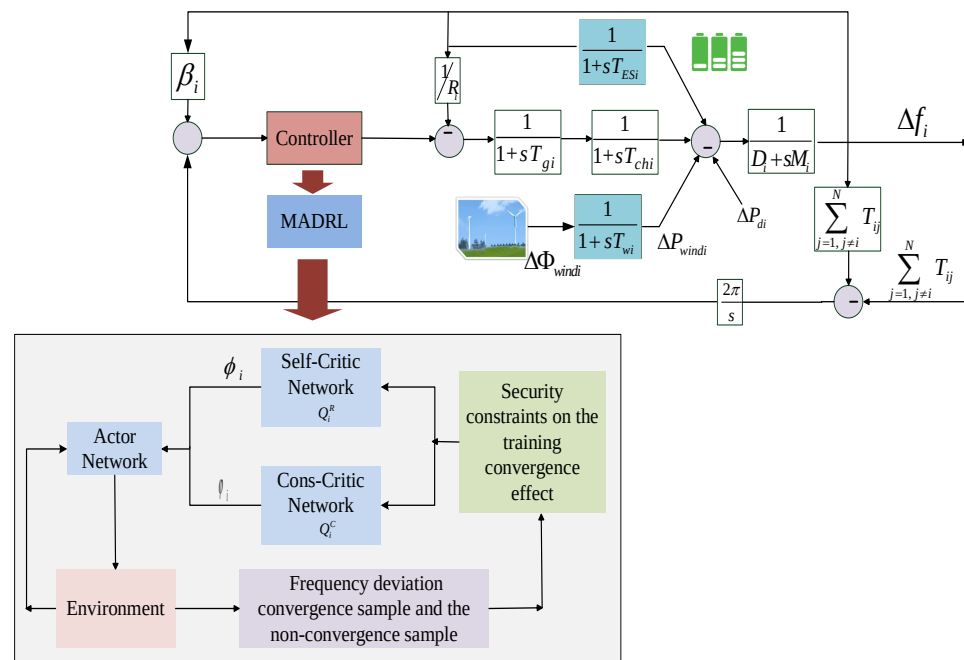


Figure 1. Transfer function model of multi-area power system.

Table 1. Notations.

Symbol	Quantity
ΔP_{di}	Load deviation
ΔP_{mi}	Generator mechanical output deviation
ΔP_{vi}	Valve position deviation
ΔP_{w-i}	Output power fluctuation of the wind turbine generator
ΔP_{Bi}	Output power fluctuation of the battery
Δf_i	Frequency deviation
$\Delta \Phi_{w-i}$	Wind power deviation
M_i	Moment of inertia
D_i	Generator damping coefficient
T_{gi}	Time constant of the governor
T_{chi}	Time constant of the turbine
T_{wi}	Time constant of the wind turbine
T_{ESi}	Time constant of the battery
R_i	Speed drop
β_i	Frequency bias factor
T_{ij}	Tie-line synchronizing coefficient
ACE_i	Area control error
ΔP_{tie-i}	Tie-line power deviation

The system partitioning method influences the composition of adjustment resources within each area, consequently impacting the stability and convergence speed of MADRL training. The algorithm proposed in this paper assumes a predefined system partitioning result, with the system network topology and partitioning remaining unchanged. Additionally, it is assumed that agents can share operational state and action information during the centralized training process.

The Constrained Markov Cooperative Game (CMCG) model comprises the following elements: a set of agents M ; the system state space S ; the observable state space S_i ; the action space A_i ; the instant reward function R_i of agent i ; the global security constraint cost function C ; the global real-time reward function R ; the system state transition probability P ; and the discount factor γ .

For the interaction process between the agent and the environment, first, agent i observes the operating status $s_{i,t} \in S_i$ of the area it is located in and generates actions $a_{i,t} \in A_i$ according to the policy function, $\pi_i(s_{i,t}|a_{i,t})$ as the regulatory resource scheduling decision of the area in which it is located. Then, according to the joint action a_t of agents, the system feeds back to each agent the immediate reward $r_{i,t} \in R_i$ and the global safety constraint cost $c_t \in C$. Finally, the system transitions to the next state $s_{t+1} \sim \prod(s_{t+1}|s_t, a_t)$. Collect state transition samples including the current state, action, immediate reward, constraint cost and next state of all agents, and store them in the experience playback pool.

In the process of interacting with the environment, each agent utilizes samples from the experience replay pool to optimize the accuracy of the action-value and action-cost evaluation functions. Based on the optimization goals constructed by the action-value and action-cost evaluation functions, each agent updates its own policy function. This interaction process is repeated until all agents converge and no longer change their policy functions. At this point, the multi-agent policy function represents the real-time optimal scheduling strategy model for the system.

For the multi-area LFC power system, the frequency will be influenced by another area LFC scheme. Therefore, the proposed MADRL approach for the multi-area LFC power system employs centralized learning, wherein all agents being updated in each iteration. For each area's LFC scheme, the frequency deviation, ACE and external disturbances are represented as the state space, as shown in the following equations ($i = 1 \dots N$):

$$\begin{cases} s_1 = \begin{bmatrix} \Delta f_1 & ACE_1 & \Delta P_{d1} \end{bmatrix} \\ s_2 = \begin{bmatrix} \Delta f_2 & ACE_2 & \Delta P_{d2} \end{bmatrix} \\ \vdots \\ s_N = \begin{bmatrix} \Delta f_N & ACE_N & \Delta P_{dN} \end{bmatrix} \end{cases} \quad (7)$$

For action, in relation to control action a_i , the gradient of the centralized action-value function Q is quantified. For agent i , taking the gradient with respect to a_i , the following can be obtained:

$$\begin{aligned} \frac{\partial Q_i}{\partial a_i} &\approx -2B_i \Delta f_i \frac{\partial f_i}{\partial a_i} - 2\Delta P_{tie-i} \frac{\partial P_{tie-i}}{\partial a_i}, \\ &\quad - \sum_{j \neq i}^n (2\Delta P_{tie-j} \frac{\partial P_{tie-j}}{\partial a_i}), \\ \frac{\partial P_{tie-i}}{\partial a_i} &\approx 2\pi \left(\sum_{j \neq i}^n T_{ij} \frac{\partial f_j}{\partial a_i} - \sum_{j \neq i}^n T_{ij} \frac{\partial f_j}{\partial a_i} \right), \\ \frac{\partial P_{tie-j}}{\partial a_i} &\approx 2\pi \left(\sum_{k \neq i}^n T_{jk} \frac{\partial f_j}{\partial a_i} - \sum_{k \neq i}^n T_{ik} \frac{\partial f_k}{\partial a_i} \right). \end{aligned} \quad (8)$$

The gradient of frequency deviation in various locations with respect to the control actions a_i needs to be estimated based on the above formula.

During the interaction process, the system environment translates the actions output by the agent into scheduling decisions based on the established mapping relationship. This ensures that the scheduling decisions consistently adhere to the constraints of the regulating resources.

For a multi-area LFC system, frequency deviation and ACE affect frequency regulation performance. In this way, the reward function for MADRL agent in different area LFC systems is designed as follows:

$$r(t) = \begin{cases} \mu_{11}|\Delta f(t)| + \mu_{12}|ACE(t)|, & |\Delta f(t)| \leq 0.05 \\ -\mu_{21}|\Delta f(t)| - \mu_{22}|ACE(t)|, & 0.05 < |\Delta f(t)| \leq 0.1 \\ -\mu_{31}|\Delta f(t)| - \mu_{32}|ACE(t)|, & 0.1 < |\Delta f(t)| \leq 0.15 \\ -\mu_{41}|\Delta f(t)| - \mu_{42}|ACE(t)|, & 0.15 < |\Delta f(t)| \leq 0.2 \\ -\mu_{51}|\Delta f(t)| - \mu_{52}|ACE(t)|, & |\Delta f(t)| > 0.2 \end{cases} \quad (9)$$

where $\mu_{11}, \mu_{12}, \mu_{21}, \dots, \mu_{52}$ represent the weighting factors of the reward functions corresponding to different control areas and control objectives.

The global optimization objective of the multi-agent cooperation alliance can be expressed as the following:

$$\max_{\pi} J(\pi) = E_{\pi} \left[\sum_{t=1}^T \gamma^{t-1} r(t) \right]. \quad (10)$$

3. Multi-Agent Actor-Double-Critic Deep Reinforcement Learning Algorithm for LFC Power System

Because $r(t)$ cannot accurately reflect the contribution of i -th agent's strategy π_i to the global optimization goal, optimizing each agent's individual strategy during training may reduce the overall optimization objective. This can lead to conflicts between agents and reduce the training convergence speed. Therefore, using the global instantaneous reward $r(t)$ as the immediate reward for each agent is not feasible for fostering a cooperative relationship among agents to maximize $J(\pi)$.

To improve the convergence of MADRL training, applying the cost sharing idea of Vickrey–Clark–Groves (VCG) auction mechanism into the design of agent instant reward function r_i to accurately reflect the contribution of agent's strategy π_i to the global optimization goal. The VCG mechanism maximizes social welfare by considering the losses incurred by bidders due to the participation of other bidders as auction costs. This approach ensures that the optimal bidding strategy, which maximizes the utility of each individual bidder, also maximizes the overall social welfare of the bidder group.

In this paper, taking the frequency deviation and ACE in the area where the agent i is located are considered as the benefits of its participation in cooperation.

$$\hat{J}_i(\pi_i) = E_{\pi} \left[- \sum_{t=1}^T \gamma^{t-1} (\Delta f_i(t) + ACE_i(t)) \right]. \quad (11)$$

The VCG cost $VCG_i(\pi_{-i})$ of agent i is the difference between the global optimization objective $J(\pi_{-i})$ when it does not participate in real-time scheduling and the value created by other agents in the alliance when it does participate. Since agents make decisions based solely on the operating status of their respective regions, their action strategies are independent of one another. Therefore, whether agent i participates in scheduling does not change the strategies of other agents. The $VCG_i(\pi_{-i})$ can be expressed as follows:

$$\begin{aligned} VCG_i(\pi_{-i}) &= J(\pi_i) - J(\pi) + \hat{J}_i(\pi_i) \\ &= E_{\pi} \left[\sum_{t=1}^T \gamma^{t-1} (F_p^{-i}(t) + F_p(t)) \right], \end{aligned} \quad (12)$$

where π_{-i} represents the policy set that does not contain agent i , and $F_p^{-i}(t)$ is the global immediate reward and penalty term when agent i does not participate in scheduling.

The utility of agent i participating in the cooperative alliance is the difference between revenue and VCG cost. Taking the utility regarded as the value function $J_i(\pi_i)$ of agent i ,

$$J_i(\pi_i) = \hat{J}_i(\pi_i) - VCG_i(\pi_{-i}). \quad (13)$$

Therefore, the immediate reward function of agent i can be obtained as

$$r_i(t) = F_p(t) - F_p^{-i}(t) - \Delta f_i(t) - ACE_i(t). \quad (14)$$

It can be seen from the above formula that it reflects the impact of a single agent on the global optimization goal while considering other agents. This approach can alleviate conflicts when each agent optimizes its own strategy in a distributed manner and promote convergence in MADRL training.

Remark 1. The VCG mechanism does not theoretically guarantee convergence; however, it serves as a reward-shaping strategy that improves the alignment between local and global objectives, thereby facilitating cooperative learning and enhancing training stability. The CTDE framework has been widely adopted in multi-agent reinforcement learning because it balances cooperative learning capability and practical deployment requirements.

4. Construction of Multi-Agent Distributed Optimization

When the multi-agents reach the Nash equilibrium state, the joint strategy can be regarded as the optimal solution of the global optimization objective function:

$$\begin{aligned} \pi_i &= \arg \max_{\pi_i} J_i(\pi_{-i}, \pi_i) \\ &= E_{(\pi_{-i}, \pi_i)} \left[\sum_{t=1}^T \gamma^{t-1} \eta_t \gamma_{i,t} \right], \end{aligned} \quad (15)$$

$$s.t. \quad J(\pi_{-i}, \pi_i) = E_{(\pi_{-i}, \pi_i)} \left[\sum_{t=1}^T \gamma^{t-1} (1 - \eta_t) c_t \right] \leq \delta, \quad (16)$$

where δ is a minimum value limit to 0, δ denotes the safety threshold that defines the acceptable operating region of the system, and $E_{(\pi_{-i}, \pi_i)}[\cdot]$ represents the expected value of the cumulative global security constraint cost under the joint strategy π . In this paper, taking the 0–1 state quantity η_t to characterize the system power frequency regulation in period t , 1 represents system power frequency regulation convergence, and 0 represents no convergence.

To solve the above equation, the Lagrange multiplier method is applied, converting it into an unconstrained minimax problem as shown in the following equation:

$$\min_{\pi_i} \max_{\lambda \geq 0} L_i(\pi_i, \lambda) = -J_i(\pi) + \lambda J(\pi_{-i}, \pi_i) - \lambda \delta \quad (17)$$

where λ is the Lagrange multiplier,

$$\tilde{J}(\pi) = E_{\pi} \left[\sum_{t=1}^T \gamma^{t-1} c(t) \right] \leq \delta$$

The solution process of the above equation is as follows:

$$\begin{aligned}\pi_i^{k+1} &= \arg \min_{\pi_i} (-J_i(\pi_i^k, \pi_i) + \lambda^k J(\pi_i^k, \pi_i) - \lambda^k \delta), \\ \lambda^{k+1} &= \arg \max_{\lambda \geq 0} (\lambda \tilde{J}(\pi^{k+1}) - \lambda \delta).\end{aligned}\quad (18)$$

π_i and λ are updated iteratively until all agents no longer change the policy function. In this section, apply Self-Critic network $Q_i^R(s_t, a_t; \phi_i)$ and Cons-Critic network $Q_i^C(s_t, a_t; \varphi_i)$ to evaluate the action-value and action-cost of the agent, respectively. Here, ϕ_i and φ_i are network weight parameters. $Q_i^R(s_t, a_t)$ and $Q_i^C(s_t, a_t)$ represent the evaluation results output of the Self-Critic and Cons-Critic networks, respectively.

To mitigate the impact of sparse characteristics of immediate rewards and security constraints on the training convergence effect, $Q_i^R(s_t, a_t)$ and $Q_i^C(s_t, a_t)$ can be rewritten as follows:

$$\begin{aligned}Q_i^R(s_t, a_t) &= \sum_{k=t}^T \left(E_{\pi} [\gamma^{k-t} r_k | s_t, a_t, \rho_k = 1] P(\rho_k = 1) \right), \\ Q_i^C(s_t, a_t) &= \sum_{k=t}^T \left(E_{\pi} [\gamma^{k-t} c_k | s_t, a_t, \rho_k = 0] P(\rho_k = 0) \right).\end{aligned}$$

These formulas can be derived as

$$Q_i^R(s_t, a_t) |_{\rho_t=1} = r_t + \gamma E_{\pi} [Q_i^R(s_{t+1}, a_{t+1})], \quad (19)$$

$$Q_i^C(s_t, a_t) |_{\rho_t=0} = c_t + \gamma E_{\pi} [Q_i^C(s_{t+1}, a_{t+1})]. \quad (20)$$

Therefore, during the training process, only the frequency deviation and ACE samples are used to update the Self-Critic network, while the non-convergent samples are used to update the Cons-Critic network. The proposed Self-Critic and Cons-Critic structures are designed to alleviate the adverse effects of sparse learning signals by providing complementary local and coordination-oriented value evaluations. This dual-critic mechanism supplies richer feedback during policy learning, which contributes to improved learning efficiency and convergence behavior.

Self-Critic network and Cons-Critic network can be updated as the following equation:

$$\begin{aligned}\phi_i^{k+1} &= \arg \min_{\phi_i} L_i^R(\phi_i) = E \left[\left(Q_i^R(s_t, a_t; \phi_i) - y_i^R \right)^2 \right], \\ y_i^R &= r_{i,t} + \gamma Q_i^{R'}(s_t, a_t; \phi_i') |_{a_{j,t+1}=\mu_j'(s_{j,t+1}, \theta_j'), \forall j \in M'} \\ \varphi_i^{k+1} &= \arg \min_{\varphi_i} L_i^C(\varphi_i) = E \left[\left(Q_i^C(s_t, a_t; \varphi_i) - y_i^C \right)^2 \right], \\ y_i^C &= r_{i,t} + \gamma Q_i^{C'}(s_t, a_t; \varphi_i') |_{a_{j,t+1}=\mu_j'(s_{j,t+1}, \theta_j'), \forall j \in M'}.\end{aligned}$$

where D^S and D^C are respectively the frequency deviation convergence sample set and the non-convergence sample set in the experience playback pool D ; y_i^R and y_i^C represent the target evaluation results of Self-Critic and Cons-Critic respectively.

The policy gradient of Actor network for each agent can be expressed as the following equation:

$$\begin{aligned} \nabla_{\theta_i} L_i^{\pi}(\theta_i) = & - E_{DS} \left[\nabla_{\theta_i} \mu_i(s_{i,t}; \theta_i) \nabla_{a_{i,t}} Q_i^R(s_{i,t}, a_{i,t}, \right. \\ & \left. a_{-i,t}; \phi_i) \mid_{a_{i,t}=\mu_i(s_{i,t}; \theta_i)} \right] \\ & + \lambda E_{DC} \left[\nabla_{\theta_i} \mu_i(s_{i,t}; \theta_i) \nabla_{a_{i,t}} Q_i^C((s_{i,t}, i, \right. \\ & \left. a_{i,t}, a_{-i,t}, \phi_i) \mid_{a_{i,t}=\mu_i(s_{i,t}; \theta_i)} \right]. \end{aligned} \quad (21)$$

Then, applying the updated Cons-Critic network of each agent to update the Lagrange multiplier λ of the distributed optimization problem. To ensure that $\lambda \geq 0$ is always true, the variable $x = \log \lambda$ is introduced. Then, it yields the following:

$$x^{k+1} = \arg \max_x E_{(DC, j \in M)} \left[x Q_j^C(s_t, a_t; \phi_i^{k+1}) - x \delta \right]. \quad (22)$$

Remark 2. Compared with the multi-agent Actor–Critic algorithm, the proposed algorithm allows for the decoupled evaluation of action-value and action-cost, reducing the complexity of fitting a single value network. Additionally, it mitigates the impact of sparse immediate rewards and safety constraint costs on evaluation accuracy, thereby enhancing the convergence of MADRL training. Equation (21) reflects the influence of the agent’s strategy on action-value and action-cost under different system operating states. This ensures that the frequency decisions generated by the multi-agent system meet the safety operation constraints while maximizing the system’s performance optimization goals.

5. Case Study and Discussion

To prove the effectiveness of the proposed method, the two-area LFC power systems and three-area LFC power systems are applied in this section.

5.1. Case 1: Two-Area LFC Power System with Multi-Agent Actor-Double-Critic Deep Reinforcement Learning Algorithm

In this case, a two-area LFC power system utilizing multi-agent Actor-Double-Critic deep reinforcement learning algorithm is implemented to demonstrate the efficacy of the newly developed LFC approach. The system parameters are detailed in Table 2. To evaluate the performance of the designed LFC scheme based on the MADRL framework for a two-area power system, comparative experiments are conducted with PI control and fuzzy PI control, respectively.

Table 2. Parameters of two-area LFC scheme.

Area	R	M	D	T_g	T_{ch}	T_{12}
1	0.05	10.0	1.0	0.1	0.3	0.1986
2	0.05	12.0	1.5	0.17	0.4	0.1986

Based on the partitioning of the simulation system, a fully connected layer neural network is employed. The input layer of the agent’s Actor network contains three neurons, corresponding to the observed dimensions of its respective area. This network features two hidden layers, with 128 and 200 neurons, respectively. The output layer comprises a single neuron. The learning rate for the Actor network is set at 0.001. The agent’s Critic network integrates both state and action pathways, which subsequently pass through a ReLU layer followed by a fully connected layer. The state pathway incorporates two hidden layers with 128 and 200 neurons, respectively, utilizing the ReLU function as the activation function.

The action pathway includes a single hidden layer with 200 neurons. The learning rate for the Critic network is uniformly set at 0.0001.

In this case, the total simulation time is set to 150 s. To analyze the designed LFC scheme performance, the load disturbances of $\Delta P_d = 0.03$ Hz, $\Delta P_d = 0.1$ Hz, and $\Delta P_d = 0.08$ Hz at $t = 10$ s, $t = 50$ s, and $t = 100$ s are set, respectively. The simulation results are shown in Figure 2. The frequency deviation response curves for Area 1 and Area 2 are shown in Figure 2a and Figure 2b, respectively. The study quantifies the maximum absolute values and the average absolute values of frequency deviations under different controllers, as shown in Table 3.

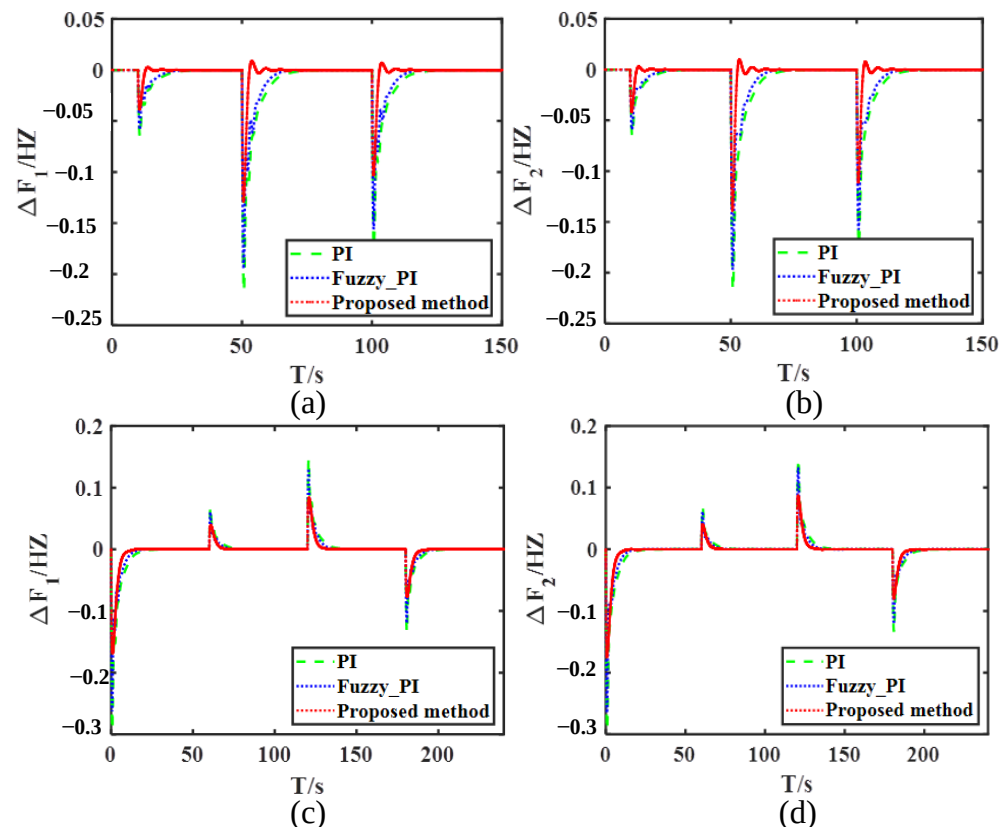


Figure 2. Frequency deviations of two-area power system with disturbances: (a) frequency deviations of Area 1 with disturbances, (b) frequency deviations of Area 2 with disturbances, (c) frequency deviations of Area 1 with random disturbances, (d) frequency deviations of Area 2 with random disturbances.

Table 3. Comparison of system performance under load disturbances.

Methods	Max $ \Delta f_1 $	Mean $ \Delta f_1 $	Max $ \Delta f_2 $	Mean $ \Delta f_2 $
Proposed method	0.12871	0.00325	0.13883	0.00329
Fuzzy PI	0.19467	0.00872	0.19684	0.00872
PI	0.21622	0.01172	0.21638	0.01171

As described in Table 3, the proposed method shows improvements over the Fuzzy PI algorithm (by 33.9%, 62.7%, 29.5%, and 62.3% respectively) and over the PI algorithm (by 40.5%, 72.3%, 35.8%, and 71.9% respectively), in terms of the maximum absolute value and the average absolute value of the system frequency deviations. The proposed method demonstrates significant improvements in managing frequency deviations. Under given load disturbances, the proposed method can quickly suppress disturbances, responding faster than the two conventional methods and resulting in significantly smaller frequency

deviations. This allows the system to rapidly return to a stable state. As illustrated in Figure 2a,b, the proposed method results in smaller frequency fluctuations and smoother frequency curves when compared to other algorithms.

To further demonstrate the effective performance of the designed LFC scheme based on MADRL, random disturbances are introduced into the two-area interconnected power systems model. Every 60 s, a random load disturbance is generated and simultaneously applied to both Area 1 and Area 2, with the upper limit of the load disturbance with $\Delta P_{dmax} = 0.2$ Hz. The frequency deviation response curve for Area 1 and Area 2 are described in Figure 2c and Figure 2d, respectively. As presented in Figure 2c,d, the proposed method significantly improves over the Fuzzy PI algorithm (by 36.0%, 24.8%, 34.3%, and 24.8%, respectively) and PI algorithm (by 42.4%, 54.2%, 40.2%, and 44.2%, respectively).

The maximum absolute values and average absolute values of the system frequency deviations under different controllers are quantified, as shown in Table 4.

Table 4. Comparison of system performance under random load disturbances across different algorithms.

Methods	Max $ \Delta f_1 $	Mean $ \Delta f_1 $	Max $ \Delta f_2 $	Mean $ \Delta f_2 $
Proposed method	0.16809	0.00537	0.17442	0.00537
Fuzzy PI	0.26259	0.00714	0.26546	0.00714
PI	0.29158	0.01172	0.29174	0.00962

Compared to other methods, the proposed approach results in smaller frequency fluctuations and less impact on the power grid. Under various load disturbances, it can rapidly reduce the frequency deviations across multiple areas to zero, thereby achieving stability.

5.2. Case 2: Three-Area LFC Power System with Multi-Agent Deep Reinforcement Learning

To further scrutinize the effectiveness of the designed LFC scheme based on MADRL. In the three-area LFC power system model, the total simulation time is set at 300 s, with a controller sampling time of 0.03 s.

The study involves introducing fixed load disturbances to different areas at specified times: a disturbance $\Delta P_{d1} = 0.3$ Hz is added to Area 1 at time $t = 2$ s, a disturbance $\Delta P_{d2} = 0.8$ Hz is added to Area 2 at time $t = 80$ s, and a disturbance $\Delta P_{d3} = 0.5$ Hz is added to Area 3 at time $t = 190$ s. The simulation results are shown in Figure 3.

It is observed that, the proposed method significantly improves over the Fuzzy PI algorithm (by 30.9%, 59.5%, 30.8%, 52.0%, 24.8%, and 66.1%, respectively) and the PI algorithm (by 39.2%, 59.5%, 35.7%, 67.0%, 40.2%, and 74.4%, respectively). The maximum absolute values and average absolute values of the system frequency deviations under different controllers are quantified, as shown in Table 5.

Table 5. Comparison of system performance under random load disturbances across different algorithms for case 2.

Methods	Max $ \Delta f_1 $	Mean $ \Delta f_1 $	Max $ \Delta f_2 $	Mean $ \Delta f_2 $	Max $ \Delta f_3 $	Mean $ \Delta f_3 $
Proposed method	0.0204	0.0069	0.0386	0.0062	0.0139	0.0064
Fuzzy PI	0.0267	0.0110	0.0549	0.0094	0.0175	0.0106
PI	0.0284	0.0110	0.0524	0.0104	0.0195	0.0112

The proposed method effectively controls frequency deviations caused by different disturbances, exhibiting smaller fluctuations. Although the Fuzzy PI and PI algorithms can also manage frequency deviations in the three-area power system, they respond more

slowly and require longer adjustment periods. Figure 3 demonstrates that the proposed method effectively reduces the impact of load disturbances on the power system, with faster response times and smoother frequency deviation curves.

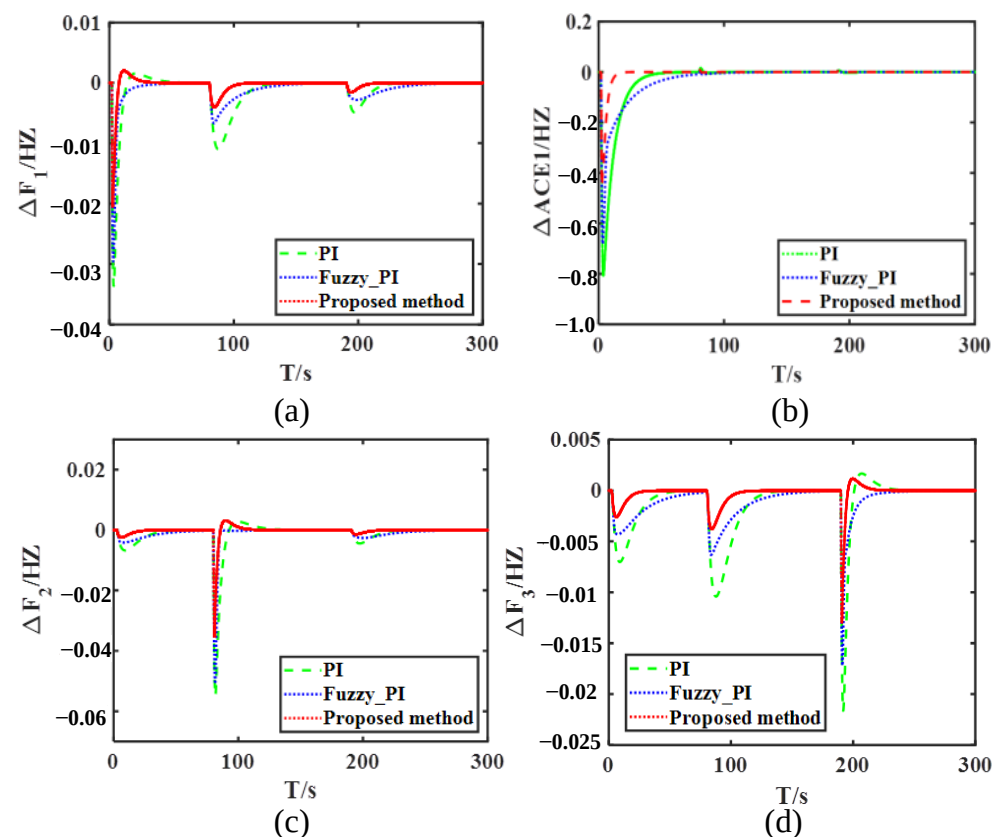


Figure 3. Frequency deviations and ACE of three-area power system with disturbances: (a) frequency deviations of Area 1 with disturbances, (b) ACE of Area 1 with disturbances, (c) frequency deviations of Area 1 with disturbances, (d) frequency deviations of Area 2 with random disturbances.

6. Conclusions

A multi-agent Actor-Double-Critic deep reinforcement learning algorithm has been developed for LFC power system in this paper. This framework ensures that the scheduling decisions generated by the multi-agent system meet the safe and reliable operation requirements of the power system. To enhance the convergence of the MADRL training process, the cost sharing idea of the VCG auction mechanism has been applied to the design of the agent's instant reward function r_i to reflect the contribution of the agent strategy π_i to the global optimization goal. It concludes that compared with traditional LFC scheme, the designed novel LFC scheme based on MADRL can further mitigate the impact of sparse immediate rewards and safety constraint costs, ensuring real-time scheduling decisions that comply with system safety operation constraints.

However, the current study assumes a centralized-training and decentralized-execution environment and does not explicitly consider communication failures, system partitioning, or fully distributed learning scenarios.

Future work will focus on extending the proposed framework to high-renewable-penetration power systems, evaluating its performance under severe renewable-energy uncertainties and communication-constrained environments, incorporating explainable reinforcement learning techniques, and developing rigorous stability and safety analysis methods to further enhance its practical applicability.

Author Contributions: Conceptualization, X.L. and X.W.; methodology, X.L. and Y.Y.; software, Y.W.; validation, Y.Y. and Z.G.; formal analysis, X.L. and Z.G.; writing—review and editing, X.L. and X.W.; funding acquisition, X.L. and Z.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China under Grants LQ24F030006 and LTGS23E050002, in part by the Fundamental Research Funds of Zhejiang Sci-Tech University under Grant 25222175-Y.

Data Availability Statement: All data generated or analysed during this study are included in this published article. Also, additional information (i.e., extra datasets) in this study used during the current study is available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Yu, Y.; Othman, M.; Chu, Y. Analytical design and performance of MSTOGI-PLL in unified power quality conditioner (UPQC) for anti-disturbance enhancement in power quality mitigation. *Arab. J. Sci. Eng.* **2026**, *18*, 11555–11576. [\[CrossRef\]](#)
2. Wang, Y.; Zhang, D.; Li, Y.; Jiao, W.; Wang, G.; Zhang, J. Enhancing power grid resilience with blockchain-enabled vehicle-to-vehicle energy trading in renewable energy integration. *IEEE Trans. Ind. Appl.* **2024**, *60*, 2037–2052. [\[CrossRef\]](#)
3. Almutairi, S.; Anayi, F.; Packianather, M.; Shouran, M. An innovative LFC system using a fuzzy FOPID-enhanced via PI controller tuned by the catch fish optimization algorithm under nonlinear conditions. *Sustainability* **2025**, *17*, 5966. [\[CrossRef\]](#)
4. Wang, B.; Sun, H.; Zhao, B.; Wu, P.; Li, W.; Xu, S.; Chen, J. Calculation method of renewable maximum penetration considering frequency response of renewable energy. *Electr. Power Syst. Res.* **2024**, *237*, 110984. [\[CrossRef\]](#)
5. Shukla, S.; Jha, S.; Singh, S. Optimizing load frequency control in diverse power generation systems using hybrid metaheuristic algorithms. *Arab. J. Sci. Eng.* **2025**, *50*, 18015–18036. [\[CrossRef\]](#)
6. Lu, J.; Hu, J.; Yu, J.; Cao, J. A dynamic demand response control strategy for isolated microgrid with primary frequency regulation. *Electr. Power Syst. Res.* **2023**, *224*, 109691. [\[CrossRef\]](#)
7. Lv, X.; Qiao, F.; Sun, Y.; Dinavahi, V.; Liu, P.X. Improved memory event-triggered load frequency control in multi-area power system with renewable energy. *IEEE Trans. Circuits Syst. Express Briefs* **2024**, *71*, 311–315. [\[CrossRef\]](#)
8. Huang, C.; Yang, M.; Ge, H.; Deng, S. DMPC-based load frequency control of multi-area power systems with heterogeneous energy storage system considering SoC consensus. *Electr. Power Syst. Res.* **2024**, *228*, 110064. [\[CrossRef\]](#)
9. Pati, S.; Subudhi, U. Stability analysis of a Mmulti-area renewable system and frequency control with improved chaotic harris hawk optimization algorithm. *Arab. J. Sci. Eng.* **2024**, *49*, 6531–6550.
10. Zhao, X.; Ma, Z.; Li, S.; Zou, S. Robust load frequency control of power systems with wind power under packet losses and communication delays. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2022**, *22*, 135–148. [\[CrossRef\]](#)
11. Mu, C.; Tang, Y.; He, H. Improved sliding mode design for load frequency control of power system integrated an adaptive learning strategy. *IEEE Trans. Ind. Electron.* **2017**, *64*, 6742–6751. [\[CrossRef\]](#)
12. Wang, X.; Ding, D.; Ge, X.; Dong, H. Neural-network-based control with dynamic event-triggered mechanisms under DoS attacks and applications in load frequency control. *IEEE Trans. Circuits Syst. Regul. Pap.* **2022**, *69*, 5312–5324. [\[CrossRef\]](#)
13. Oshnoei, S.; Aghamohammadi, M.R.; Khooban, M. Model-free predictive frequency control under sensor and actuator FDI attacks. *IEEE Trans. Ind. Inform.* **2024**, *71*, 2434–2438. [\[CrossRef\]](#)
14. He, X.; Ge, S.; Liu, H.; Xu, Z.; Mi, Y.; Wang, C. Frequency regulation of multi-microgrid with shared energy storage based on deep reinforcement learning. *Electr. Power Syst. Res.* **2023**, *214*, 108962. [\[CrossRef\]](#)
15. Oshnoei, A.; Kheradmandi, M.; Muyeen, S.M. Robust control scheme for distributed battery energy storage systems in load frequency control. *IEEE Trans. Power Syst.* **2020**, *35*, 4781–4791. [\[CrossRef\]](#)
16. Chen, X.; Zhao, C.; Li, N. Distributed automatic load frequency control with optimality in power systems. *IEEE Trans. Control Netw. Syst.* **2021**, *8*, 307–318. [\[CrossRef\]](#)
17. Pandi, V.; Al-Hinai, A.; Feliachi, A. Coordinated control of distributed energy resources to support load frequency control. *Energy Convers. Manag.* **2015**, *105*, 918–928. [\[CrossRef\]](#)
18. Mishra, A.; Mishra, P.; Mathur, H. A deep learning assisted adaptive nonlinear deloading strategy for wind turbine generator integrated with an interconnected power system for enhanced load frequency control. *Electr. Power Syst. Res.* **2023**, *214*, 108960. [\[CrossRef\]](#)

19. Gong, X.; Wang, X.; Joos, G. An online data-driven method for microgrid secondary voltage and frequency control with ensemble Koopman modeling. *IEEE Trans. Smart Grid* **2023**, *14*, 68–81. [[CrossRef](#)]
20. Wang, Q.; Li, F.; Tang, Y.; Xu, Y. Integrating model-driven and data-driven methods for power system frequency stability assessment and control. *IEEE Trans. Power Syst.* **2019**, *34*, 4557–4568. [[CrossRef](#)]
21. Yan, Z.; Xu, Y. Data-driven load frequency control for stochastic power systems: A deep reinforcement learning method with continuous action search. *IEEE Trans. Power Syst.* **2019**, *34*, 1653–1656. [[CrossRef](#)]
22. Yan, Z.; Xu, Y. A multi-agent deep reinforcement learning method for cooperative load frequency control of a multi-area power system. *IEEE Trans. Power Syst.* **2020**, *35*, 4599–4608. [[CrossRef](#)]
23. Li, J.; Zhou, T. Evolutionary multi-agent deep meta reinforcement learning method for swarm intelligence energy management of isolated multi-area microgrid with internet of things. *IEEE Internet Things J.* **2023**, *10*, 12923–12937. [[CrossRef](#)]
24. Huang, L.; Zhang, M.; He, M.; Guo, Y.; Duan, J. Closed loop iterative learning control for consistency tracking in lower limb rehabilitation robotic system with initial state deviations. *Sci. Rep.* **2025**, *15*, 9593. [[CrossRef](#)] [[PubMed](#)]
25. Dietterich, T.G. Hierarchical reinforcement learning with the MAXQ value function decomposition. *J. Artif. Intell. Res.* **2000**, *13*, 227–303. [[CrossRef](#)]

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