Automatic Generation Control in a Distributed Power Grid Based on Multi-step Reinforcement Learning

Wenmeng Zhao, Tuo Zeng, Zhihong Liu, Lihui Xie, Lei Xi, *Member*, *IEEE*, and Hui Ma, *Member*, *IEEE*

***Abstract*—The increasing use of renewable energy in the power system results in strong stochastic disturbances and degrades the control performance of the distributed power grids. In this paper, a novel multi-agent collaborative reinforcement learning algorithm is proposed with automatic optimization, namely, Dyna-DQL, to quickly achieve an optimal coordination solution for the multi-area distributed power grids. The proposed Dyna framework is combined with double Q-learning to collect and store the environmental samples. This can iteratively update the agents through buffer replay and real-time data. Thus the environmental data can be fully used to enhance the learning speed of the agents. This mitigates the negative impact of heavy stochastic disturbances caused by the integration of renewable energy on the control performance. Simulations are conducted on two different models to validate the effectiveness of the proposed algorithm. The results demonstrate that the proposed Dyna-DQL algorithm exhibits superior stability and robustness compared to other reinforcement learning algorithms.**

***Index Terms*—Automatic generation control, Dyna framework, distributed power grid, multi-agent, model-based reinforcement learning.**

Received: October 18, 2023
Accepted: January 10, 2024
Published Online: July 1, 2024

Wenmeng Zhao is with the Electric Power Research Institute, Southern Power Grid, Guangzhou 510663, China (zhaowm@csg.cn).

Tuo Zeng is with the Heyuan Power Supply Bureau, Guangdong Power Grid Co., Ltd., Heyuan 517000, China (1881326 1333@139.com).

Zhihong Liu (corresponding author), Lihui Xie and hui Ma are with the College of Electrical Engineering & New Energy, China Three Gorges University, Yichang 443002, China (e-mail: liu zhihong3396@163.com; lihui20042004@163.com; mahuizz119@126.com).

Lei Xi (corresponding author) is with the College of Electrical Engineering & New Energy and Hubei Provincial Key Laboratory for Operation and Control of Cascaded Hydropower Station, China Three Gorges University, Yichang 443002, China (e-mail: xilei 2014@163.com).

DOI: 10.23919/PCMP.2023.000220

I. Introduction

With the rapid development of renewable energy integrated into the traditional power system, the intermittent and highly stochastic characteristics [1] of the renewable energy bring strong random disturbances while causing significant modifications to the grid structure. As a result, the traditional automatic generation control (AGC) methods, such as proportional-integral control and sliding mode control, both of which heavily rely on precise mathematical system models, are no longer applicable.

Therefore, extensive research has been conducted with the knowledge-based artificial intelligence (AI) algorithms which require no precise model of the AGC system. The AI technologies in AGC can be roughly classified into two categories.

The first category focuses on optimizing the parameters of the traditional control methods via swarm intelligence [2] optimization algorithms. In [3], the implementation of the proportional-integral (PI) algorithm for the AGC system is discussed, wherein an electric vehicle cluster is incorporated to assist in frequency modulation. While the influence of electric vehicles generally yields positive outcomes for frequency improvement, it is noteworthy that parameters continue to be the primary influencers in frequency control. In [4], a genetic algorithm is introduced for the optimization of parameters in the PI within AGC. This addresses the challenge of parameter adjustment stemming from modeling complexities to some extent, though the efficacy of the genetic algorithm plays a crucial role in determining the overall effectiveness of the PI control. Reference [5] builds upon these efforts by introducing further enhancements to the genetic algorithm, aiming to achieve superior control performance in AGC. Nevertheless, it is imperative to acknowledge that PI controls heavily rely on accurate system models. This presents challenges in particular in large-scale power systems characterized by their inherent complexities.

The second category primarily involves reinforcement learning algorithms [6]–[12]. Because of their powerful

self-optimization capabilities, simple structure, and lack of the need for physical system models, reinforcement learning algorithms have been extensively investigated in the field of power systems. Value iteration-based Q-learning [13] and its derived model-free reinforcement learning algorithms are widely applied in AGC systems. In [14], the multi-agent deep reinforcement learning algorithm is applied to effectively address the issue of delayed rewards in an AGC system via the principle of "backward estimation". A deep reinforcement learning algorithm with exploration-aware thinking is proposed based on the double deep Q network (DDQN) algorithm in [15]. It can resolve the dimensionality disaster problem caused by high-dimensional states, thus improving system stability. The deep Q network (DQN) algorithm is improved with long short-term memory (LSTM) neural networks and is applied to the coordinated control of complex energy systems with distributed renewable energy sources and chemical energy storage equipment [16]. To improve the issue of poor control performance caused by Q-learning discrete state space in the dynamic power allocation, transfer learning [17] is applied to the real-time training process of Q-learning (QL) to enhance its learning speed [18]. However, transfer learning cannot play a role outside the preset state.

Although the above methods can improve the learning effect of agents to certain extent, model-free reinforcement learning models [19] suffer from low efficiency with the large numbers in the interactive learning process, thus requiring more time to achieve policy convergence. To address this issue, a Dyna framework is proposed to configure the environmental data model via storing the training data [20] for the environmental interaction. This enables faster update of the Q-value matrix with reliance on the model and faster system convergence. However, Q-learning is prone to the problem of overestimating the Q-values [21], where the consecutive updates within the same iteration could result in larger Q-value errors, thus jeopardizing the stability of the agents. In [22], a double Q-learning (DQL) method is introduced to decouple the connection between Q-values and Q-value indexing. This can solve the problem of Q-value overestimation so as to ensure the stability of the agents.

This paper proposes a novel Dyna-DQL algorithm for load frequency control (LFC), one which combines the model-based and model-free iterative updates for the agents so as to avoid Q-value overestimation in nonlinear distributed power systems. Extensive simulations are performed on different models with varied operating conditions to validate the effectiveness of the proposed algorithm. The contributions of the paper are summarized as follows:

1) A model-based reinforcement learning algorithm is proposed to combine the model-free reinforcement learning algorithm and Dyna framework so as to improve the learning efficiency of agents via buffer data playback and frequency control performance in the distributed power systems.

2) A two-area LFC model with superconducting magnetic energy storage (SMES), and a four-area LFC model based on the central china power grid are established for simulation verification.

3) A series of simulations are performed with various models and operating conditions to verify the effectiveness of the proposed algorithm.

The remainder of the paper is organized as follows. Section II explains the proposed Dyna-DQL algorithm in detail together with the Dyna framework and Q-learning. The design of the AGC system model and the related parameter settings are introduced in Section III. Case studies in different scenarios are performed to verify the effectiveness of the proposed algorithm in Section IV. Conclusions are provided in Section V.

II. Frequency Control Model for Distributed Interconnected AGC System

With the integration of large-scale distributed renewable energy sources, the power grids with such structure can be considered as a distributed power grid. Hence, the AGC system [23] is also designed to be distributed where the traditional centralized AGC system dispatching center no longer performs the "central coordination". Instead, the grid system can be divided into multiple distributed areas, each with a distributed "brain" (intelligent controller, also known as an agent). These agents are engaged in the multi-agent interactions to regulate the power of each distributed region. The model of the distributed AGC system is illustrated in Fig. 1. Each agent is responsible for handling load disturbances within its own area. Through the physical and digital information exchange channels, the collective agents aim to achieve coordinated control of the entire power grid, with the objective of ensuring that the tie-line power deviation ΔP_{tie} be zero.

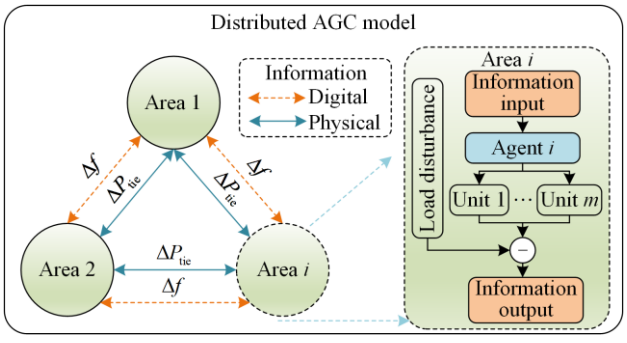


Fig. 1. The distributed AGC system mode.

The difference between the regulation output power $\Delta P_{\text{T},i}$ of a given unit and the increment $P_{\text{D},i}$ in load power during the frequency regulation process can be computed using the subsequent formula:

$$\Delta P_{\text{T},i} - \Delta P_{\text{D},i} = \frac{\text{d}}{\text{d}t} W_i + \Delta P'_{\text{D},i} + \Delta P_{\text{tie},i} \tag{1}$$

where $\mathrm{d}W_i/\mathrm{d}t$ represents the power increment of the generator; $\Delta P'_{\mathrm{D},i}$ stands for the load frequency regulation power; and $\Delta P_{\mathrm{tie},i}$ denotes the exchange power of the tie line of the ith area.

The power flowing through the tie-line from area 1 to area 2 can be calculated using the subsequent formula:

$$L[\Delta P_{\mathrm{tie},12}] = \frac{2\pi}{l} T_{12} \left(L[\Delta f_1] - L[\Delta f_2] \right) \tag{2}$$

where $L[\cdot]$ denotes the Laplace transform; l is the variable for the Laplace transform; Δf_1 and Δf_2 represent the frequency deviations of areas 1 and 2 respectively; and T_{12} is the time constant of the tie line. Consequently, the tie-line exchange power of area i and transmission power of area 'i' and 'j' can be determined as follows:

$$\Delta P_{\mathrm{tie},i} = \frac{2\pi}{l} \left(\sum_{\substack{j=1 \\ j\neq i}}^{N} T_{ij} \Delta f_i - \sum_{\substack{j=1 \\ j\neq i \\ j\neq 1}}^{N} T_{ij} \Delta f_j \right) \tag{3}$$

$$P_{ij} = \frac{|V_i||V_j|}{X_{ij}} \sin(\delta_i - \delta_j) \tag{4}$$

$$\Delta \dot{f}_i = \frac{1}{2H_i} (\Delta P_{\mathrm{m}i} + \Delta P_{\mathrm{RES}i} - \Delta P_{\mathrm{L}i} - \Delta P_{\mathrm{tie},i}) - \frac{D_i}{2H_i} \Delta f_i \tag{5}$$

$$\Delta \dot{P}_{\mathrm{m}i} = \frac{1}{T_{ti}} \Delta P_{\mathrm{g}i} - \frac{1}{T_{ti}} \Delta P_{\mathrm{m}i} \tag{6}$$

$$\Delta \dot{P}_{\mathrm{g}i} = \frac{1}{T_{\mathrm{g}i}} \Delta P_{ci} - \frac{1}{R_i T_{\mathrm{g}i}} \Delta f_i - \frac{1}{T_{\mathrm{g}i}} \Delta P_{\mathrm{g}i} \tag{7}$$

where X_{ij} denotes the equivalent reactance of the tie line; δ_i and δ_j represent the power angles of the equivalent generators in areas i and j, respectively; V_i and V_j denote the output voltages of the equivalent generators in areas i and j, respectively; T_{ti} is the turbine time constant of area i; $T_{\mathrm{g}i}$ is the time constant of area i; $\Delta P_{\mathrm{m}i}$ represents the generation command; $P_{\mathrm{L}i}$ denotes the total load; $\Delta P_{\mathrm{RES}i}$ is the aggregate of all active power transmitted via the tie-line connected to area i; and D_i is the damping coefficient of area i; R_i denotes the primary frequency regulation coefficient of area i; and ΔP_{ci} is the control power deviation of area i.

In large-scale power systems, area control error (ACE) is typically employed to ensure frequency stability and manage power exchange. The model discussed in this paper does not explicitly establish a dispatch center. Instead, it introduces an intelligent agent in each area to independently control operations, with each agent tasked with maintaining the stability of the ACE index within its respective area. The method used to calculate ACE is outlined below:

$$ACE_i = \Delta P_{\mathrm{tie,i}} + B_i \Delta f_i \tag{8}$$

$$B_i = \frac{1}{R_i} + D_i \tag{9}$$

where B_i represents the deviation coefficient of area i.

III. Dyna-DQL Algorithm

A. Dyna Framework

As demonstrated in Fig. 2, Dyna is an AI framework to realize learning, planning and action execution. It includes three components: buffer data containing digital environment information, agent, and the interaction of the agent with the environment. In each iteration, the agent can be updated via real-time data and sampled information from the buffer.

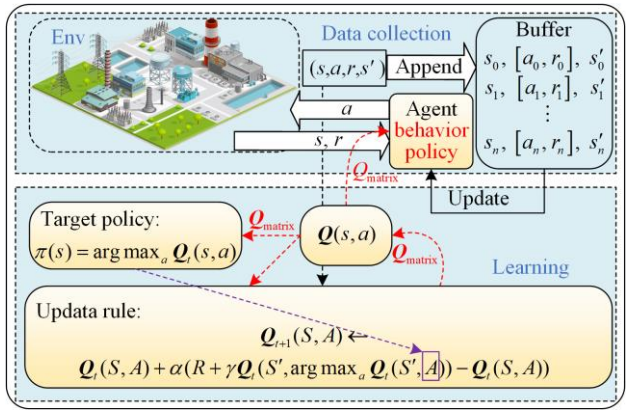


Fig. 2. The Dyna framework.

In reinforcement learning, each data update stems from various strategies, specifically, the off-policy [21] method. It is evident that the prerequisite for utilizing Dyna is a reinforcement learning algorithm based on the off-policy training method. The off-policy method involves using a behavior policy, denoted as 'b', to gather data. This collected data is then employed to optimize a separate target policy, denoted as 'π'. Within the off-policy method, the policy update process incorporates the importance sampling technique. This technique utilizes the sample data generated by the behavior policy to compute the expected value of the probability distribution produced by the target policy. The specific calculation formula is as follows:

$$\rho_{t:T-1} = \frac{\prod_{k=t}^{T-1} \pi(A_k|S_k) p(S_{k+1}|S_k, A_k)}{\prod_{k=t}^{T-1} b(A_k|S_k) p(S_{k+1}|S_k, A_k)} = \frac{\prod_{k=t}^{T-1} \pi(A_k|S_k)}{\prod_{k=t}^{T-1} b(A_k|S_k)} \tag{10}$$

where ρ denotes the probability density; while S and A correspond to the state set and action set of the policy, respectively; the symbol p stands for the state transition probability. Despite the trajectory of the policy execution action being associated with the state transition probability, the ratio of the target policy to the behavior policy can negate the influence of environmental factors. Consequently, ρ is solely determined by the behavior policy, the target policy, and the corresponding trajectory, rendering it independent of the Markov decision process [24].

In reinforcement learning that employs value iteration,

the state value function under a given policy, denoted as π, can be computed using the subsequent formula:

$$V^{\pi}(s) = E_{\pi}[G_t \mid S_t = s] \tag{11}$$

where $V^{\pi}(s)$ represents the expected value of the expected reward that can be obtained by starting from state s and following policy π; G_t represents the sum of the rewards obtained by the agent from the beginning of time t to the end of a phase; E_{π} denotes the value expectation under policy π. Regarding state value calculation, the primary computation method hinges on the relationship between the value of the current state and that of the subsequent state. Based on the importance sampling technique and the Bellman equation, the ensuing value formula can be derived:

$$V^{\pi}(s) = \sum_{a \in A} \pi(a \mid s) \sum_{s' \in S} p(s' \mid s,a)[r(s' \mid s,a) + \gamma V^{\pi}(s')] \tag{12}$$

where r denotes the environmental reward; while s and s' correspond to the current and subsequent states of the environment, respectively; the symbol a signifies the agent's action; and γ is the discount factor. Consequently, we can ascertain the state value of any policy, denoted as π, based on the sample data derived from various behavior policies, denoted as b. This establishes the theoretical foundation for the application of the Dyna framework in reinforcement learning algorithms that utilize the off-policy method. Q-learning, a quintessential reinforcement learning algorithm, employs the off-policy method for value updates. Despite its simple architecture and minimal computational resource requirements, it offers extensive application potential and research value.

B. Q-learning

Q-learning is a model-free reinforcement learning algorithm based on value iteration. It maintains a $\boldsymbol{Q}$-value matrix containing state-action values of the real-time interaction with the environment. The $\boldsymbol{Q}$-value matrix is calculated as:

$$\boldsymbol{Q}_{t+1}(s,a) \leftarrow \boldsymbol{Q}_t(s,a) + \alpha(r + \gamma \max_{a \in \mathrm{A}} \boldsymbol{Q}_t(s',a) - \boldsymbol{Q}_t(s,a)) \tag{13}$$

where $\boldsymbol{Q}$ is the state-action value matrix. The optimization process of the $\boldsymbol{Q}$-values is updated as follows:

$$\forall s,a : \boldsymbol{Q}^*_{t+1}(s,a) = \sum_{s'} \boldsymbol{P}^{s'}_{s,a}(r^{s'}_{s,a} + \gamma \max_{a \in \mathrm{A}} \boldsymbol{Q}_t(s',a)) \tag{14}$$

where $\boldsymbol{P}$ represents the state transition probability matrix and $\boldsymbol{Q}^*$ is the optimal value function. For any state s and action a, $\boldsymbol{Q}^*$ is dependent on $\arg\max_a \boldsymbol{Q}_t$. This could result in significant overestimation of the $\boldsymbol{Q}$ value.

In the case of sufficient historical data, the combination of the Dyna framework with Q-learning allows the agent to update various state-action values via the data extracted from the buffer, thereby accelerating the learning process. However, after multiple updates within a single iteration, the accumulation of the $\boldsymbol{Q}$-value overestimation could result in system instability.

To address the problem of $\boldsymbol{Q}$-value overestimation, a DQL is introduced which can decouple the maximum value and related index between different estimators during the agent learning process. The update of the $\boldsymbol{Q}$-value in the DQL is calculated as:

$$\boldsymbol{Q}^{\mathrm{A}}(s,a) \leftarrow \boldsymbol{Q}^{\mathrm{A}}(s,a) + \alpha(s,a)(r + \gamma \max_{a^* \in A} \boldsymbol{Q}^{\mathrm{B}}(s',a^*) - \boldsymbol{Q}^{\mathrm{A}}(s,a)) \tag{15}$$

$$\boldsymbol{Q}^{\mathrm{B}}(s,a) \leftarrow \boldsymbol{Q}^{\mathrm{B}}(s,a) + \alpha(s,a)(r + \gamma \max_{b^* \in A} \boldsymbol{Q}^{\mathrm{B}}(s',b^*) - \boldsymbol{Q}^{\mathrm{B}}(s,a)) \tag{16}$$

After the obtained current state s_t, the agent selects the optimal action from the action set A via the ε-greedy policy, described as:

$$a_t = \begin{cases} \arg\max_{a_k \in A} \boldsymbol{Q}(s_t, a_k), & 1-\varepsilon \\ \boldsymbol{Q}(s_t, a_r), & \varepsilon \end{cases} \tag{17}$$

This indicates that the agent can acquire the action with the highest $\boldsymbol{Q}$-value with a probability of $(1-\varepsilon)$, and select a random action a_r in the action set A with a probability of ε.

C. Dyna-DQL

The flowchart of the Dyna-DQL strategy is illustrated in Fig. 3.

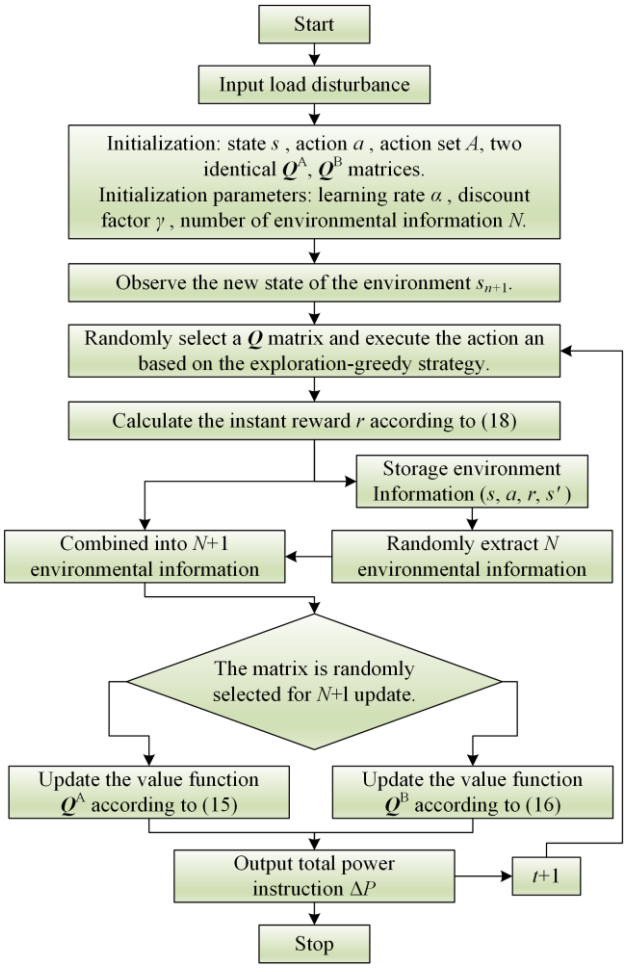


Fig. 3. The flowchart of the Dyna-DQL strategy.

The developed Dyna-DQL can improve the decision-making performance of the agent by employing experience replay and avoiding ***Q***-value overestimation.

During the experience replay, N randomly sampled environmental simulation data are used to update different ***Q***-values. This can significantly improve the efficiency of the ***Q***-table updates. The introduction of the DQL can avoid the problem of ***Q***-value overestimation in each update process, making the algorithm more stable.

IV. Design of AGC System based on Dyna-DQL Algorithm

A. AGC System Control Performance Standards

The control performance standards (CPS) were proposed by the North American Electric Reliability Corporation to evaluate regional power grids [18]. The CPS, ACE and grid frequency deviation (Δf) are monitored in real-time, stored and calculated as the input state variables of the controller in the AGC system. The Dyna-DQL controller can optimize and update the Q-value function based on the state variables and environmental rewards, and outputs the control power commands to maintain grid stability. The specific criteria are summarized as:

1) If CPS1 $\geqslant$ 200% and CPS2 is at any value, the CPS criteria are met.

2) If 100% $\leqslant$ CPS1 < 200% and CPS2 $\geqslant$ 90%, the CPS criteria are met.

3) If CPS1 < 100%, the CPS criteria are not met.

4) If $f \in$ (49.80, 50.20) Hz, the frequency criterion is met.

B. Reward Function and Actions Setting

To fully consider the impact of ACE and CPS criteria, the weighted values of ACE and CPS1 are used as the reward function for the regional power grid, given as:

$$R(t) = -\eta[S_{\mathrm{ACE}}(t)]^2 - \frac{(1-\eta)S_{\mathrm{CPS1}}(t)}{1000} \tag{18}$$

where $S_{\mathrm{ACE}}(t)$ and $S_{\mathrm{CPS}}(t)$ are the instantaneous values of the ACE and CPS1 at time t, respectively; while η is the weighting coefficient (set as 0.5 here). In the standard test case used in this paper, the AGC control period is 4 s, and the secondary delay is 20 s.

For the implementation of the proposed algorithm in the AGC system, the action space of the continuous system is discretized. The agent's action set is established as [−50:10:50] to facilitate effective control.

C. Parameter Settings

The involved parameters are set as follows.

1) Learning rate α $(0<\alpha<1)$: Determine the confidence level of the update process. Here, α is set to 0.1.

2) Discount factor γ $(0<\gamma<1)$: Determine the weight between the current and long-term rewards. Here, γ is set to 0.95.

3) Exploration rate ε $(0<\varepsilon<1)$: The action is selected with the highest ***Q***-value in the current state with a probability of $(1-\varepsilon)$, and a new action with a probability of ε is explored. In the simulation, ε is set to 0.9.

4) Environmental information samples N: The amount of the sampled environmental information and the number of the consecutive updates in each iteration. Here, N is set to 20.

V. Case Study

A. The Improved IEEE Standard Two-area Model

An SMES unit (seen in Fig. 4) is integrated into the IEEE standard two-area model [25] to form an improved two-area LFC model, as demonstrated in Fig. 5.

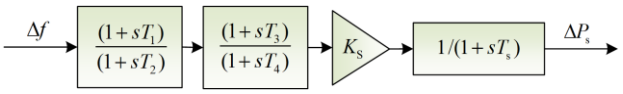


Fig. 4. The simulation model of the SMES unit.

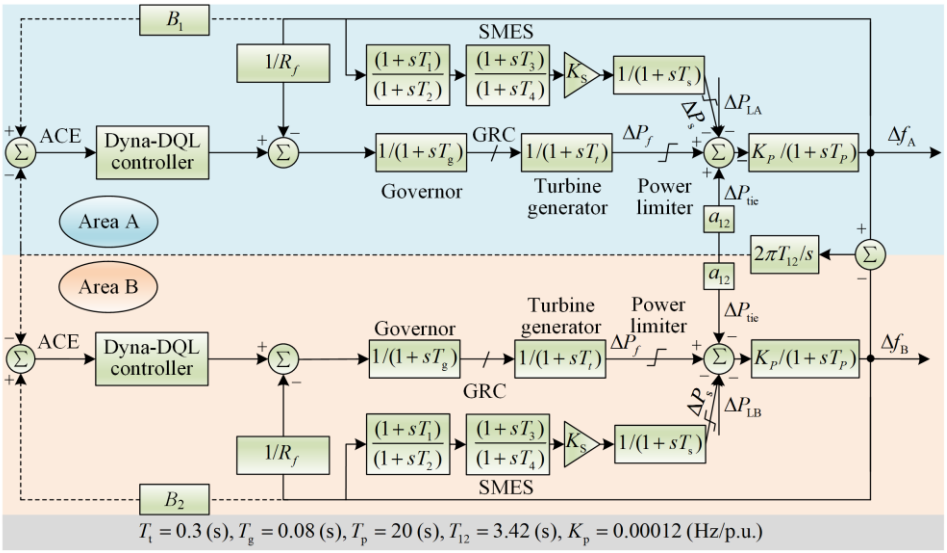


Fig. 5. The improved IEEE standard two-area LFC model.

In Figs. 4 and 5, Δf is the frequency deviation, ΔP_s represents the output power of the SMES unit, $\{T_1, T_2, T_3, T_4, T_s\}$ are the time constants of the SMES unit, and K_s is the gain coefficient of the SMES unit. Detailed parameters are listed in Table I. T_g is the time delay constant of the thermal power unit governor, T_t is the time constant of the thermal power unit, T_p is the time constant of the frequency response function, and K_p is the coefficient of the frequency response function. ΔP_{tie} represents the tie-line exchange power, and T_{12} is the time constant of the tie-line.

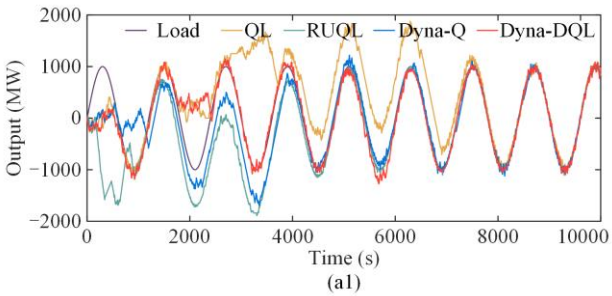
TABLE I
COMPARISON OF UNIT MODEL PARAMETERS

Unit	Parameters
Thermal power unit	$T_g = 0.08,\ T_t = 0.03$
SMES unit	$K_s = 0.0665,\ T_s = 0.001,$ $T_1 = 0.1,\ T_2 = 0.0665,$ $T_3 = 0.0661,\ T_4 = 0.099$

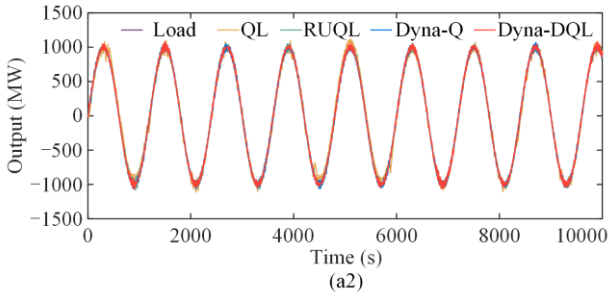
B. Pre-learning

Reinforcement learning is a trial-and-error learning process, so that the Dyna-DQL controller should go through a process of random exploration and trial-and-error learning, known as pre-learning before it is deployed on the online operation. In this paper, the four algorithms of Dyna-DQL, Dyna-Q, RUQL and QL are used for performance comparison. The models of the two areas are subjected to continuous sine signals with a period of 1200 s and amplitude of 1000 MW. The assessment period is set to 30 000 s to guide the agent to make the optimal decisions. The results of the pre-learning process for the four algorithms are illustrated in Fig. 6.

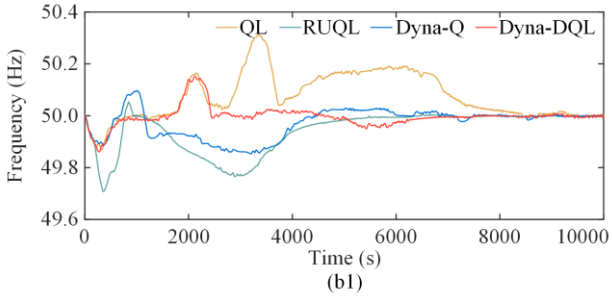
It can be seen from Fig. 6(a1) that during the pre-learning phase, the actual load curve can be well followed after 2500 s, 4000 s, 4000 s and 7000 s via the Dyna-DQL, Dyna-Q, RUQL and Q-learning controllers, respectively. Dyna-DQL has the shortest pre-learning time and a smoother curve. The frequency fluctuation of the Dyna-DQL controller during the initial phase of the pre-learning remains within 0.2 Hz (see Fig. 6(b1)), and after approximately 2500 s of the trial-and-error learning, the frequency deviations in area A and area B stabilize within 0.01 Hz. The actual operational curves in Fig. 6(b2) indicate that Dyna-DQL can maintain the frequency error within 0.005 Hz from the beginning, meeting the frequency criteria within the area. Fig. 6(c1) indicates that the ACE of Dyna-DQL quickly converges within ±5 MW after experiencing two significant fluctuations during the pre-learning phase. In the actual operational phase of Fig. 6(c2), Dyna-DQL can also maintain the ACE within ±2 MW from the starting point. Similarly, the pre-learning curves in Fig. 6(d1) demonstrate that the maximum error of the CPS1 indicator for Dyna-DQL is around 140% or 170%, which is reduced to within 0.005% after approximately 2500 s. During the online operation shown in Fig. 6(d2), Dyna-DQL can also maintain the CPS1 within 200% from the beginning.



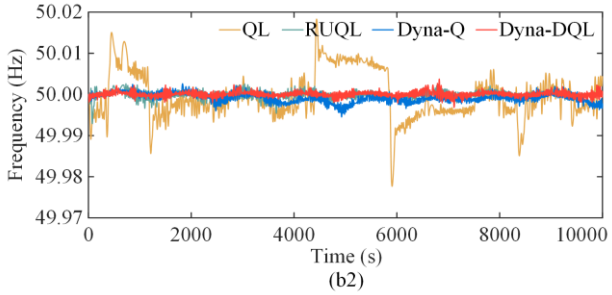
(a1)



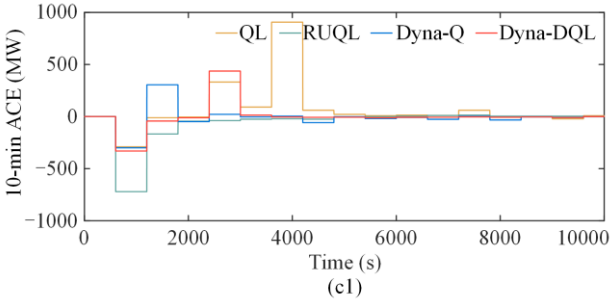
(a2)



(b1)



(b2)



(c1)

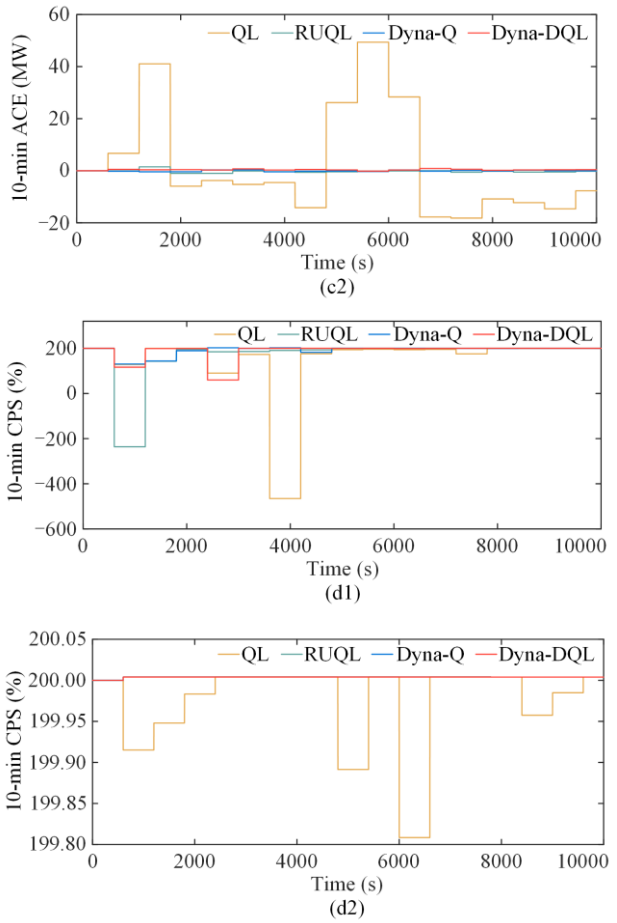


Fig. 6. Comparison of the pre-learning performance of different algorithms. (a1) The pre-learning load curves. (a2) The load curves during the actual operation. (b1) The pre-learning of the frequency curves. (b2) The frequency curves during the actual operation. (c1) 10-min average ACE during the pre-learning. (c2) 10-min average ACE during the actual operation. (d1) 10-min average CPS1 during the pre-learning. (d2) 10-min average CPS1 during the actual operation.

In summary, Fig. 6 indicates that there are significant fluctuations in the frequency, ACE, and CPS1 indices during the formal operation phase via the Dyna-Q, RUQL and Q-learning algorithms. While noticeable instability can be observed via the Q-learning algorithm, Dyna-DQL has powerful and stable frequency control performance. To systematically evaluate the control performance of the proposed algorithm with a broader spectrum of load scenarios, we incorporate step and square wave load disturbances in the subsequent simulations to validate the efficacy of the algorithm.

C. Step Disturbance

Considering the scenario of sudden load increase during actual grid operation, step disturbances are introduced to simulate and test the four algorithms. Figure 7 displays the load tracking curves with step disturbance, where Dyna-DQL exhibits faster rise time and more stable load tracking performance in Figs. 7(a) and (b). After the stabilization, Dyna-DQL can maintain the minimum overshoot with the reduced absolute value of the frequency deviation ($|\Delta f|$/Hz) up to 42.24%. Figure 7(c) demonstrates that the absolute value of the ACE control deviation ($|\mathrm{ACE}|$/MW) is reduced by up to 64.95%, while Dyna-DQL can also achieve a superior CPS1 value, as seen in Fig. 7(d).

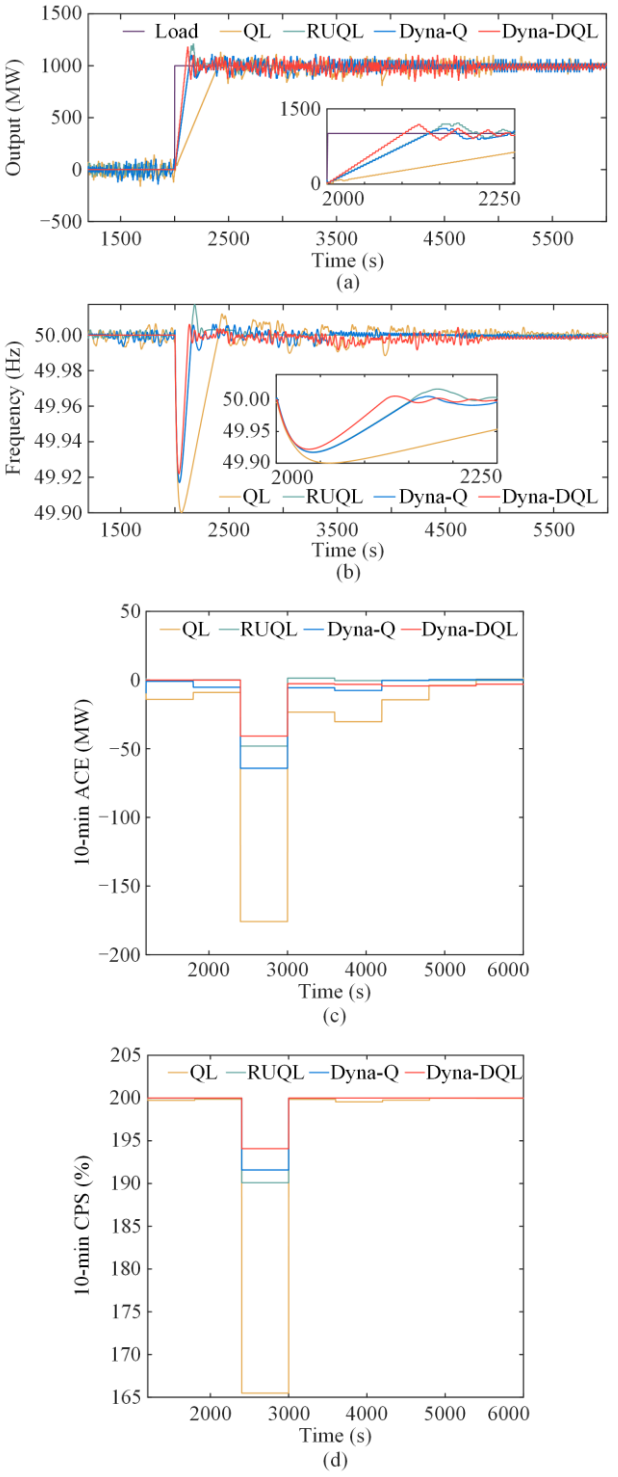


Fig. 7. Control performance comparison with step disturbance with four algorithms. (a) Load tracking curves. (b) Frequency variation curves. (c) 10-min average ACE curves. (d) 10-min average CPS1 curves.

D. Square Wave Disturbance

To simulate both regular and sudden load variations, a square wave disturbance with a period of 2400 s is introduced for the simulation tests, and Fig. 8(a) illustrates the load tracking curves for the four algorithms. It can be seen that Q-learning experiences a clear load shedding phenomenon. From Figs. 8(b), 8(c) and 8(d), Dyna-DQL has significant advantages in maintaining the frequency, ACE and CPS1 stability.

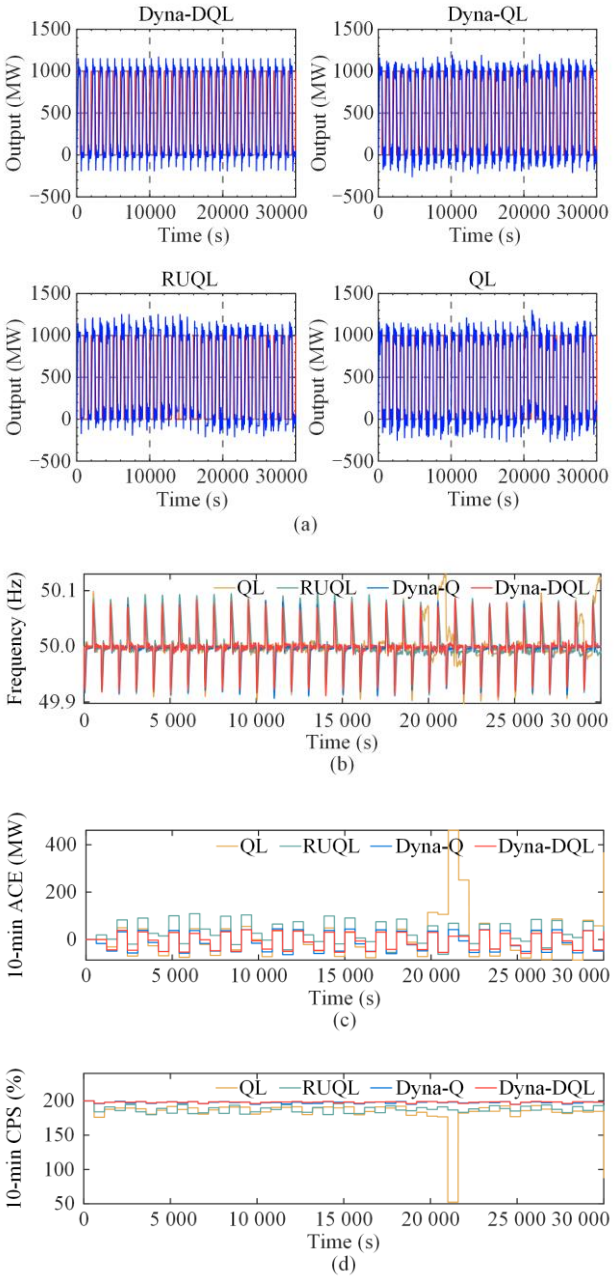


Fig. 8. Control curves with square wave disturbance. (a) Load tracking curves with square waves. (b) Frequency variation curves. (c) 10-min average ACE curves. (d) 10-min average CPS1 curves.

Compared to the other three algorithms, Dyna-DQL can reduce $|\Delta f|$ between 9.74% to 43.03%, reduce $|\mathrm{ACE}|$ between 1.71% to 44.38% and achieve higher CPS1 values. Hence, Dyna-DQL can maintain stable control performance even under load uncertainties and unclear energy output.

E. Distributed LFC Model of Central China Power Grid in Four Areas

Here, a distributed power grid model is established for the four areas of Hubei, Hunan, Henan, and Jiangxi in central China, as demonstrated in Fig. 9. The model incorporates various energy sources, i.e., wind, solar, thermal and hydro power, diesel generators, biomass power generators and energy storage equipment. For simplicity, the wind and solar power, and energy storage units do not receive frequency control signals from the AGC controller, and are only considered as disturbances. The daily wind data and solar power generation data are obtained from[26], while the data of the thermal and hydro power, biomass power generators, diesel generators and energy storage units are modeled via the typical transfer function models from[27]. The detailed parameters of each unit are listed in Table II.

TABLE II
COMPARISON OF UNIT MODEL PARAMETERS

Unit	Parameters
Biomass power generator	$T_{SB}=10,\ T_{GB}=0.08,$ $T_{WB}=5,\ T_{MB}=0.3$
Diesel generator	$T_{SD}=7,\ T_{GD}=2,$ $T_{WF}=1,\ T_{MD}=3$
Hydro power generator	$T_{gh}=5,\ T_{rs}=-1,$ $T_{rh}=0.5,\ T_{wls}=-5,\ T_{wlz}=0.513$

Considering more realistic operating conditions of the power generators, white noise disturbances are introduced in the periodic signals to simulate a series of small random fluctuations in the power grid. With an assessment period of 30 000 s, the control performance of the four algorithms is tested, and Fig. 10 depicts the corresponding load tracking curves. As seen, Dyna-DQL has a smaller fluctuation range which can better track the load disturbances. Comparing the performance indices in Table III, Dyna-DQL can reduce $|\Delta f|$ from 4.21% to 79.69%, reduce $|\mathrm{ACE}|$ up to 90.04%, and consistently maintain a superior CPS1 value. Hence, Dyna-DQL can achieve rapid coordinated control of the distributed power grids, and balance the electric rate with standard performance satisfaction.

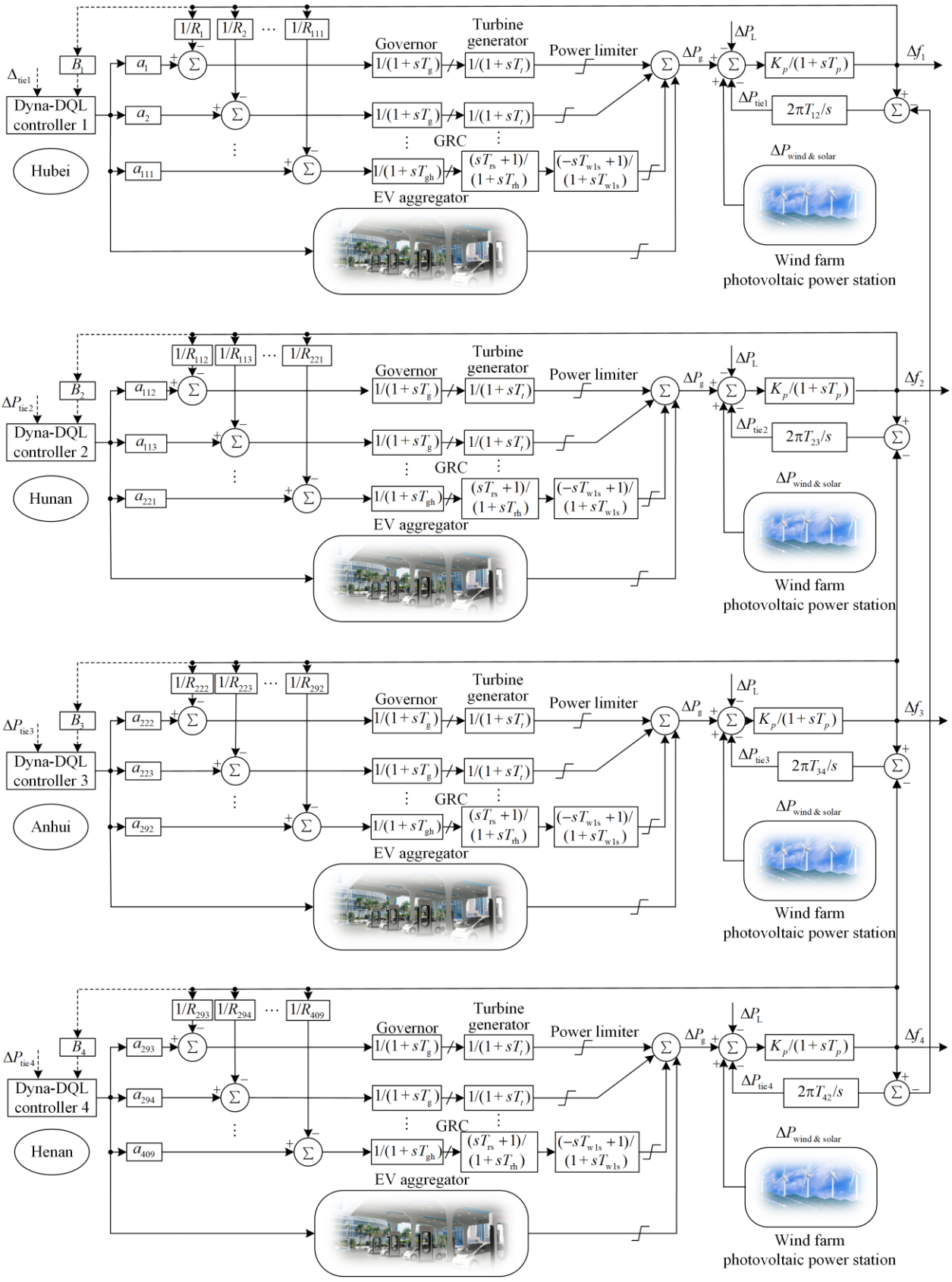


Fig. 9. Distributed LFC model of the four areas based on the central china power grid.

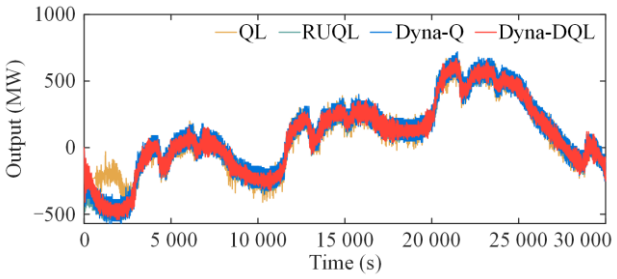


Fig. 10. Load tracking curves under random white noise disturbance.

TABLE III
CONTROL PERFORMANCE OF THE FOUR AREAS WITH RANDOM WHITE NOISE DISTURBANCE

Area	Algorithms	$\|\Delta f\|$ (Hz)	$\|ACE\|$ (MW)	CPS1(%)	CPS2(%)	CPS(%)
Hubei	Dyna-DQL	**0.0010**	**4.2754**	**199.9804**	**99.88**	**100**
	Dyna-Q	0.0015	10.3517	199.9493	99.86	100
	RUQL	0.0019	5.8392	199.9622	99.88	100
	QL	0.0044	7.3248	199.9184	98.89	100
Hunan	Dyna-DQL	**0.0013**	**5.0522**	**199.7524**	**99.56**	**99.77**
	Dyna-Q	0.0029	12.6432	199.2985	98.95	99.38
	RUQL	0.0025	8.8556	199.5929	99.40	99.74
	QL	0.0064	52.6368	191.6530	92.09	93.30
Anhui	Dyna-DQL	**0.0012**	**4.5032**	**199.8771**	**99.42**	**100**
	Dyna-Q	0.0014	4.6688	199.8770	99.39	100
	RUQL	0.0046	20.2630	193.3394	98.33	98.82
	QL	0.0047	8.8208	199.8483	98.31	100
Henan	Dyna-DQL	**0.00091**	3.9201	199.9807	**99.89**	**100**
	Dyna-Q	0.00095	**3.1900**	**199.9868**	99.80	100
	RUQL	0.0016	4.1563	199.9952	99.48	100
	QL	0.0040	5.0694	199.9625	99.88	100

VI. CONCLUSION

This paper proposes a model-based reinforcement learning algorithm, namely, Dyna-DQL, from the perspective of an AGC for coordinated control of distributed power grids. In the developed Dyna-DQL algorithm, the Dyna framework is used to store the environmental parameter information and employ experience replay to update the DQL algorithm at a high frequency. This improves the utilization efficiency of the environmental information and accelerates the training speed of the agent. The improved IEEE standard two-area LFC model and the distributed power grid model of the four areas in Central China are used for simulation verification. Different types of disturbances are applied to the distributed multi-area distributed power grid model to confirm that Dyna-DQL can effectively address the challenging control of the multi-area and distributed power grids with high penetration of renewable energy.

Nevertheless, the Q-learning-based algorithm demonstrates ineffectiveness when dealing with high-latitude state-action environments. Thus, in a forthcoming investigation, we will shift the focus to the use of neural networks and policy gradient algorithms in the realm of integrated energy system scheduling. This strategic pivot aims to specifically tackle and improve on facing the challenges posed by such high-latitude state-action environments.

ACKNOWLEDGMENT

The authors extend their sincere appreciation to the College of Electrical Engineering & New Energy and the Hubei Provincial Key Laboratory for Operation and Control of Cascaded Hydropower Station at China Three Gorges University (the first affiliation), the Electric Power Research Institute of China Southern Power Grid (the second affiliation), and the Heyuan Power Supply Bureau of Guangdong Power Grid Co., Ltd. (the third affiliation), for their invaluable support in theoretical analysis and experimental work.

AUTHORS' CONTRIBUTIONS

Wenmeng Zhao: methodology, formal analysis and software. Tuo Zeng: visualization and writing original draft. Zhihong Liu: writing, reviewing, editing and validation. Lihui Xie and Hui Ma: data collection and management. Lei Xi: conceptualization and supervision. All authors read and approved the final manuscript.

FUNDING

This work is supported by the National Natural Science Foundation of China (No. 52277108) and Guangdong Provincial Department of Science and Technology (No. 2022A0505020015).

DECLARATIONS

Competing interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

AUTHORS' INFORMATION

Wenmeng Zhao received Ph.D. degree of electrical power system and automation from South China University of Technology in 2016. He is currently working as a senior engineer at the Research Institute of Southern Power Grid. He has been granted 26 authorized patents and has published 40 research papers. His research interests include optimization operation of power systems, power markets, and virtual power plants.

Tuo Zeng received the bachelor degree in electrical engineering and automation from Huizhou University, Huizhou, China in 2013. He is currently working the director of scientific and technological progress in the production technology department of Heyuan Power Supply Bureau of Guangdong Power Grid Co., Ltd. His research interests include intelligent transmission and

transformation equipment technology, flexible intelligent distribution network technology, equipment operation and maintenance technology.

Zhihong Liu received the bachelor degree in electronic information science and technology from Chongqing Normal University, Chongqing, China in 2020. He is currently pursuing the M.Sc. with the College of Electrical Engineering and New Energy, China Three Gorges University, Yichang, China. His research interests include smart generation control and artificial intelligence techniques.

Lihui Xie received M.S. degree of electrical engineering from China Three Gorges University in 2013. Now, he is pursuing the Ph.D. in the College of Electrical Engineering and New Energy, China Three Gorges University. His research interests include the smart generation control and artificial intelligence techniques.

Lei Xi received the M.S. degree in control theory and control engineering from the Harbin University of Science and Technology, Harbin, China, in 2009, and Ph.D in electrical engineering from the South China University of Technology, Guangzhou, China, in 2013. He is currently a full professor with the College of Electrical Engineering and New Energy, China Three Gorges University, Yichang, China. His research interests include load frequency control, artificial intelligence techniques, automatic generation control and network attack and defense.

Hui Ma was born in Kaifeng, Henan Province, China in 1985. He received the Ph.D. degree in power electronics from the South China University of Technology, Guangzhou, China, in 2016. He is currently an associate professor in the College of Electrical Engineering & New Energy, China Three Gorges University, Yichang. His current research interests include the high-power density rectifiers, multilevel converters, and electric energy conversion control strategies in various industrial fields.

REFERENCES

[1] L. Liu, X. Hu, and J. Chen *et al*., "Embedded scenario clustering for wind and photovoltaic power, and load based on multi-head self-attention," *Protection and Control of Modern Power Systems*, vol. 9, no. 1, pp. 122-132, Jan. 2024.

[2] N. Kumari, P. Aryan, and G. L. Raja *et al*., "Dual degree branched type-2 fuzzy controller optimized with a hybrid algorithm for frequency regulation in a triple-area power system integrated with renewable sources," *Protection and Control of Modern Power Systems*, vol. 8, no. 3, pp. 1-29, Jul. 2023.

[3] Y. L. Abdel-Magid and M. M. Dawoud, "Optimal AGC tuning with genetic algorithms," *Electric Power Systems Research*, vol. 38, no. 3, pp. 231-238, Sept. 1996.

[4] C. Chang, W. Fu, and F. Wen, "Load frequency control using genetic-algorithm based fuzzy gain scheduling of PI controllers," *Electric Machines & Power Systems*, vol. 26, no. 1, pp. 39-52, Jan. 1998.

[5] Q. Zhang, Y. Li, and C. Li *et al*., "Grid frequency regulation strategy considering individual driving demand of electric vehicle," *Electric Power Systems Research*, vol. 163, pp. 38-48, Oct. 2018.

[6] Q. Zhang, H. Liu, and C. Li, "A hierarchical dispatch model for optimizing real-time charging and discharging strategy of electric vehicles," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 13, no. 4, pp. 537-548, Apr. 2018.

[7] Q. Zhang, Y. Tan, and J. Cai *et al*., "Negotiation strategy for discharging price of EVs based on fuzzy bayesian learning," *IET Generation Transmission & Distribution*, vol. 12, no. 20, pp. 4396-4406, Nov. 2018.

[8] L. Xi, H. Li, and J. Zhu *et al*. (2022, Aug.), "A novel automatic generation control method based on the large-scale electric vehicles and wind power integration into the grid," *IEEE Transactions On Neural Networks and Learning Systems*, [Online], pp. 1-11. Available: 10.1109/TNNLS.2022.3194247

[9] L. Xi, L. Zhou, and Y. Xu *et al*., "A multi-step unified reinforcement learning method for automatic generation control in multi-area interconnected power grid," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 2, pp. 1406-1415, Apr. 2021.

[10] L. Yin, Q. Gao, and L. Zhao *et al*., "Expandable deep learning for real-time economic generation dispatch and control of three-state energies based future smart grids," *Energy*, vol. 191, Jan. 2020.

[11] L. Yin, S. Li, and H. Liu, "Lazy reinforcement learning for real-time generation control of parallel cyber-physical-social energy systems," *Engineering Applications of Artificial Intelligence*, vol. 88, Feb. 2020.

[12] L. Yin, T. Yu, and X. Zhang *et al*., "Relaxed deep learning for real-time economic generation dispatch and control with unified time scale," *Energy*, vol. 149, pp. 11-23, Apr. 2018.

[13] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3, pp. 279-292, 1992.

[14] J. Li, T. Yu, and H. Cui *et al*., "A multi-agent deep reinforcement learning-based 'Octopus' cooperative load frequency control for an interconnected grid with various renewable units," *Sustainable Energy Technologies and Assessments*, vol. 51, Jun. 2022.

[15] L. Xi, L. Yu, and Y. Xu *et al*., "A novel multi-agent DDQN-AD method-based distributed strategy for automatic generation control of integrated energy systems," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 4, pp. 2417-2426, Oct. 2020.

[16] L. Yin and J. Xie, "Multi-temporal-spatial-scale temporal convolution network for short-term load forecasting of power systems," *Applied Energy*, vol. 283, Feb. 2021.

[17] F. Ouyang, J. Wang, and H. Zhou, "Short-term power load forecasting method based on improved hierarchical transfer learning and multi-scale CNN-BiLSTM- Attention," *Power System Protection and Control*, vol. 51, no. 2, pp. 132-140, Jan. 2023. (in Chinese)

[18] X. Zhang, Q. Li, and T. Yu *et al*., "Consensus transfer q-learning for decentralized generation command dispatch based on virtual generation tribe," *IEEE Transactions on Smart Grid*, vol. 9, no. 3, pp. 2152-2165, May 2018.

[19] L. Yin and Y. Wu, "Mode-decomposition memory reinforcement network strategy for smart generation control in multi-area power systems containing renewable energy," *Applied Energy*, vol. 307, Feb. 2022.

[20] R. S. Sutton, "Dyna, an integrated architecture for learning, planning and reacting," *ACM Sigart*, *Bulletin*, vol. 2, no. 4, pp. 160-163, July 1991.

[21] X. Zhang, D. Meng, and J. Cai *et al.*, "A swarm based double Q-learning for optimal PV array reconfiguration with a coordinated control of hydrogen energy storage system," *Energy*, vol. 266, Mar. 2023.

[22] H. V. Hasselt, "Double Q-learning," in *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, Vancouver, Canada, Jan. 2010, pp. 2613-2321.

[23] L. Yin and Z. Su, "Multi-step depth model predictive control for photovoltaic power systems based on maximum power point tracking techniques," *International Journal of Electrical Power & Energy Systems*, vol. 131, Oct. 2021.

[24] R. Bellman, "A Markovian decision process," *Journal of mathematics and mechanics*, vol. 134, no. 4, pp. 679-684, 1957.

[25] J. Li, T. Qian, and T. Yu, "Data-driven coordinated control method for multiple systems in proton exchange membrane fuel cells using deep reinforcement learning," *Energy Reports*, vol. 8, pp. 290-311, Nov. 2022.

[26] Z. Zhuo, N. Zhang, and J. Yang *et al.*, "Transmission expansion planning test system for AC/DC hybrid grid with high variable renewable energy penetration," *IEEE Transactions on Power Systems*, vol. 35, no. 4, pp. 2597-2608, Jul. 2020.

[27] G. Fei, B. Blunier, and D. Chrenko *et al.*, "Multirate fuel cell emulation with spatial reduced real-time fuel cell modeling," *IEEE Transactions on Industry Applications*, vol. 48, no. 4, pp. 1127-1135, Aug. 2012.