frontiers | Frontiers in Energy Research

ORIGINAL RESEARCH
published: 13 July 2022
doi: 10.3389/fenrg.2022.947532

# An AGC Dynamic Optimization Method Based on Proximal Policy Optimization

*Zhao Liu[1]*, Jiateng Li[2], Pei Zhang[1]*, Zhenhuan Ding[3] and Yanshun Zhao[1]*

[1]School of Electrical Engineering, Beijing Jiaotong University, Beijing, China, [2]Department of Artificial Intelligence Applications, China Electric Power Research Institute, Beijing, China, [3]School of Artificial Intelligence, Anhui University, Hefei, China

The increasing penetration of renewable energy introduces more uncertainties and creates more fluctuations in power systems than ever before, which brings great challenges for automatic generation control (AGC). It is necessary for grid operators to develop an advanced AGC strategy to handle fluctuations and uncertainties. AGC dynamic optimization is a sequential decision problem that can be formulated as a discrete-time Markov decision process. Therefore, this article proposes a novel framework based on proximal policy optimization (PPO) reinforcement learning algorithm to optimize power regulation among each AGC generator in advance. Then, the detailed modeling process of reward functions and state and action space designing is presented. The application of the proposed PPO-based AGC dynamic optimization framework is simulated on a modified IEEE 39-bus system and compared with the classical proportional–integral (PI) control strategy and other reinforcement learning algorithms. The results of the case study show that the framework proposed in this article can make the frequency characteristic better satisfy the control performance standard (CPS) under the scenario of large fluctuations in power systems.

**Keywords: automatic generation control, advanced optimization strategy, deep reinforcement learning, renewable energy, proximal policy optimization**

OPEN ACCESS

***Edited by:***
*Bo Yang,*
*Kunming University of Science and Technology, China*

***Reviewed by:***
*Xiaoshun Zhang,*
*Northeastern University, China*
*Jiawen Li,*
*South China University of Technology, China*

****Correspondence:***
*Zhao Liu*
*liuzhao1@bjtu.edu.cn*
*Pei Zhang*
*2512692577@qq.com*

***Specialty section:***
*This article was submitted to Smart Grids, a section of the journal Frontiers in Energy Research*

***Received:*** *18 May 2022*
***Accepted:*** *07 June 2022*
***Published:*** *13 July 2022*

***Citation:***
*Liu Z, Li J, Zhang P, Ding Z and Zhao Y (2022) An AGC Dynamic Optimization Method Based on Proximal Policy Optimization. Front. Energy Res. 10:947532. doi: 10.3389/fenrg.2022.947532*

## INTRODUCTION

Automatic generation control (AGC) is applied to ensure frequency deviation and tie-line power deviation within the allowable range in power systems as a fundamental part of energy management system (EMS) (Jaleeli et al., 2002). Conventional AGC strategies calculate the total adjustment power based on the present information collected from Supervisory Control and Data Acquisition (SCADA) system including frequency deviation, tie-line power deviation, and area control error (ACE), etc., and then allocates the total adjustment to each AGC unit. The control period is generally 2–8 s. Therefore, the key to conventional AGC strategies is to solve two problems: ① how to calculate the total adjustment power based on the online information; ② how to allocate the total adjusted power to each AGC unit with the goal of satisfying the control performance standard (CPS) and minimizing the operation cost. At present, to solve these two problems, scholars have proposed many control strategies. For calculating the total adjustment power, proposed strategies include the classical proportional–integral (PI) control (Concordia and Kirchmayer, 1953), proportional–integral-derivative (PID) control (Sahu et al., 2015; Dahiya et al., 2016), optimal control (Bohn and Miniesy, 1972; Yamashita and Taniguchi, 1986; Elgerd and Fosha, 2007), adaptive control (Talaq and Al-Basri, 1999; Olmos et al., 2004), model predictive control (Atic et al., 2003; Mcnamara and Milano, 2017), robust control (Khodabakhshian and Edrisi, 2004; Pan and Das, 2016), variable structure control (Erschler et al., 1974; Sun, 2017), and intelligent control

technologies such as neural network (Beaufays et al., 1994; Zeynelgil et al., 2002), fuzzy control (Talaq and Al-Basri, 1999; Feliachi and Rerkpreedapong, 2005), and genetic algorithm (Abdel-Magid and Dawoud, 1996; Chang et al., 1998). In terms of allocating total adjustment power to each AGC unit, a baseline allocation approach is proposed according to the adjustable capacity ratio and installed capacity ratio of each unit without considering the differences of dynamic characteristic among units. Additionally, Yu et al. (2011) treated the power allocation as a stochastic optimization problem, which can be discretized and modeled as a discrete-time Markov decision process. Also, the problem is solved by utilizing the Q-learning algorithm of reinforcement learning.

In general, the conventional AGC strategy is designed under a typical feedback-loop structure with the characteristic of hysteresis, which regulates the future output of AGC units based on the present input signal. However, the penetration of large-scale renewable energy introduces high stochastic disturbance to modern power grid due to the characteristic of dramatical fluctuation (Banakar et al., 2008). The phenomenon has not only increased regulation capacity of AGC units but also put forward higher requirement for coordinated control ability of generation units with different dynamic characteristics (such as thermal and hydroelectric units). Nevertheless, the fast regulation capacity of units in power systems is limited. When the load or renewable energy generations is continuously rising or falling, the units with second-level regulation performance will approach its upper or lower regulation limit. At this point, it is hard to ensure the frequency deviation and tie-line power deviation within an allowable range if the fast regulation capacities are insufficient in the system. On the other hand, the adjustment ratio of different units is different, i.e., to be exact, thermal units have minute-level regulation performance, while hydroelectric units have second-level regulation performance. Therefore, these strategies cannot effectively coordinate units with different characteristics, which will cause overshoot or under-adjustment. At present, the goal of AGC strategies is to maintain the dynamic control performance of system to comply with CPS established by the North American Electric Reliability Council (NERC) (Jaleeli and Vanslyck, 1999). CPS pays more attention to the medium- and long-term performance of system frequency deviation and tie-line power deviation, while it no longer requires the ACE to cross zero every 10 min and aims to smoothly regulate the frequency of power systems.

To address the hysteresis issues of conventional AGC strategies and make the dynamic performance satisfy CPS, some scholars put forward the concept of AGC dynamic optimization (Yan et al., 2012). The basic idea can be described as the optimization of the regulation power of AGC units in advance based on ultra-short-term load forecasting and renewable energy generation forecasting information, different security constraints, and objective functions. The strategy aims to optimize the AGC units' regulation power in the next 15 min, and the optimization step is 1 min. From the perspective of dispatching framework formulated by the power grid dispatching center, the AGC dynamic optimization can be viewed as a link between real-time economic dispatch (especially for the next 15 min) and routine AGC (control period is 2–8 s), which can achieve a smooth transition between the two dispatch sections. Compared with economic dispatch, AGC dynamic optimization takes the system's frequency deviation, tie-line power deviation, and ACE and CPS values into account. Compared with conventional AGC strategies, it introduces load and renewable energy forecasting information into account which can better handle renewable energy's fluctuation. Moreover, the dispatch period is 1 min which adapts to the thermal AGC units with minute-level regulation characteristics.

Yan et al. (2012) proposed a mathematic model for AGC dynamic optimal control. It takes the optimal CPS1 index and minimizes ancillary service cost as objective function. The system constraints are considered including system power balance constraints, AGC units' regulation characteristics, tie-line power deviation, and frequency deviation. This model added ultra-short-term load forecasting information into the power balance constraints as well as mapping the relationship between system frequency and tie-line power. Zhao et al. (2018) expanded the model proposed in Yan et al. (2012), taking the ultra-short-term wind power forecasting value and its uncertainties into account and conducted a chance constraint programming AGC dynamic optimization model with probability constraints and expected objectives. An optimal mileage-based AGC dispatch algorithm is proposed in Zhang et al. (2020). Zhang et al. (2021a) further extended the methods in Zhang et al. (2020) with adaptive distributed auction to handle the high participation of renewable energy. A novel random forest-assisted fast distributed auction-based algorithm is developed for coordinated control in large PV power plants in response to the AGC signals (Zhang et al., 2021b). A decentralized collaborative control framework of autonomous virtual generation tribe for solving the AGC dynamic dispatch problem was proposed in Zhang et al. (2016a).

In general, the existing research defined AGC dynamic optimal control as a multistage nonlinear optimization problem that includes objective functions and constraint conditions. To deal with the uncertainties of wind power, some scholars adopted chance-constrained programming method based on the probabilistic model of wind power. However, the accurate probability information of random variables is difficult to model, which limits the accuracy and practicality of this method. Moreover, the stochastic programming model is too complex to solve. Furthermore, these methods cannot take the future fluctuations of wind power into account when making decisions.

Artificial intelligence-based methods have been developed in recent years to address the AGC command dispatch problem, including the lifelong learning algorithm and the innovative combination of the consensus transfer of the Q learning (Zhang et al., 2016b; Zhang et al., 2018). Deep reinforcement learning (DRL) is a branch of machine learning algorithms and an important method of stochastic control based on the Markov decision process, which can better solve sequential decision problems (Sutton and Barto, 1998). Recently, DRL has been

successfully implemented on many applications of power systems, such as optimal power flow (Zhang et al., 2021a), demand response (Wen et al., 2015), energy management system for microgrid (Venayagamoorthy et al., 2016), autonomous voltage control (Zhang et al., 2016b), and AGC (Zhou et al., 2020; Xi et al., 2021). In AGC problems, as stated previously, the presented literatures usually focus on the power allocation problem which still belongs to the conventional AGC strategy. Different from the previous works, this article focuses on AGC dynamic optimization and utilizes 1 minute time resolution wind power and loads forecasting values, which are collected from and used by real wind farms and grid dispatching centers, to regulate the power outputs of AGC units. Unlike the existing optimization model, this article defines AGC dynamic optimization as a Markov decision process and a stochastic control problem and takes the various uncertainties and fluctuations of wind power outputs into account. To better solve the dynamic optimization and support safe online operations, the proximal policy optimization (PPO) deep reinforcement learning algorithm is implemented, with the clipping mechanism of PPO, which can provide more reliable outputs (Schulman et al., 2015).

The key contributions of this article are summarized as follows: ① by formulating the AGC dynamic optimization problem as the Markov decision process with appropriate power grid simulation environment, reasonable state space, action space, and reward functions, the PPO-DRL agent can be trained to learn how to determine the regulation power of AGC units without violating the operation constraints; ② by adopting the state-of-the-art PPO algorithm (Wang et al., 2020), the well-trained PPO-DRL agent could consider the uncertainties of wind power fluctuations in the future when making decisions at the current moment.

The remaining parts of this article are organized as follows: *Introduction* provides the advanced AGC dynamic optimization model considering wind power integration and the details of how to transform advanced AGC dynamic optimization into a multistage decision problem. *Introduction* introduces the principles of reinforcement learning, PPO algorithm, and the procedures of the proposed methodology. In *Introduction*, the IEEE 39-bus system is utilized to demonstrate the effectiveness of the proposed method. Finally, some conclusions are given in *Introduction*.

## ADVANCED AGC DYNAMIC OPTIMIZATION MATHEMATICAL MODEL AND MULTISTAGE DECISION PROBLEM

The essential strategy of AGC dynamic optimization is an advanced control strategy, which aims to optimize the adjustment power of each AGC unit per minute in the next 15 min according to the ultra-short-term load and wind generation forecasting information as well as the current operation condition of each unit, system frequency, and tie-line power. The objective function is to minimize the total adjustment cost, while the system dynamic performance (i.e., frequency, tie-line power deviation, and ACE) is to comply with CPS and satisfy the security constraints. Specifically, the constraints include system power balance, CPS1 and CPS2 indicators, frequency deviation, tie-line power deviation limit, and AGC unit regulation characteristics. The mathematical model of AGC dynamic optimization is formulated as follows:

$$\min f_{AGC} = \sum_{t=1}^{15} \sum_{i=1}^{N_{AGC}} \left[ k_1 \left( P_{G_i}^{max} - P_{G_i}^{min} \right) + k_2 \left| P_{G_i,t} - P_{G_i,t-1} \right| \right] \Delta t, \tag{1}$$

where $N_{AGC}$ is the number of AGC units in the system, $k_1$ and $k_2$ are the cost coefficients, $P_{G_i}^{max}$ and $P_{G_i}^{min}$ are the maximum and minimum output of the AGC unit $i$, $P_{Gi,t}$, and $P_{G_i,t-1}$ are output of AGC unit $i$ at time $t$ and $t-1$, and $\Delta t$ is the time interval, that is, 1 min.

1) Power balance constraints:

$$\sum_{i=1}^{N} P_{G,i,t} + P_{w,t} - P_{L,t} - P_{T,t} - \Delta P_{T,t} - P_{loss,t} = 0, \tag{2}$$

where $P_{w,t}$ and $P_{L,t}$ are, respectively, the predicted power of wind power and load, $P_{T,t}$ is scheduled power of tie-line, $\Delta P_{T,t}$ is forecast deviation of tie-line power, and $P_{loss,t}$ is the transmission loss.

2) CPS1 constraints:

$$\underline{K}_{cps1} \le K_{cps1} \le \bar{K}_{cps1}, \tag{3}$$

where $K_{cps1}$ is the CPS1 index of the system and $\underline{K}_{cps1}$ and $\bar{K}_{cps1}$ are, respectively, the lower and upper limits of the CPS index. $K_{cps1}$ is derived by the following equation:

$$K_{cps1} = \left[ 2 - \frac{\sum_{t=1}^{T} e_{ACE,t} \Delta f_t}{-150 B \varepsilon_{1\,\mathrm{min}}^2} \right] \times 100\%, \tag{4}$$

where $e_{ACE,t}$ is the area control error at time t, $\Delta f_t$ is frequency deviation at time t, B is the equivalent frequency regulation constant for the control area (in MW/0.1Hz), and $\varepsilon_{1\mathrm{min}}$ is the frequency control target and it is usually taken as annual statistic of the root mean square deviation of the interconnection power grid over 1 min period.

3) CPS2 constraints:

$$|E_{ACE-15\,\mathrm{min}}| \le 1.65 \varepsilon_{15\,\mathrm{min}} \sqrt{100 B B_s}, \tag{5}$$

where $E_{ACE-15\,\mathrm{min}}$ is the average ACE over the 15 min period, $\varepsilon_{15min}$ is the annual statistic of the root mean square deviation of the interconnection power grid over 15 min period, and $B$ and $B_s$ are, respectively, the equivalent frequency regulation constants for the control area and the whole interconnection power grid.

4) Power output constraints of units:

$$\underline{P}_{AG,i} \le P_{AG,i,t} \le \bar{P}_{AG,i}, \tag{6}$$

where $P_{AG,i,t}$ is output power of unit $i$ at time $t$ and $\underline{P}_{AG,i}$ and $\bar{P}_{AG,i}$ are the lower and upper limits of output power.

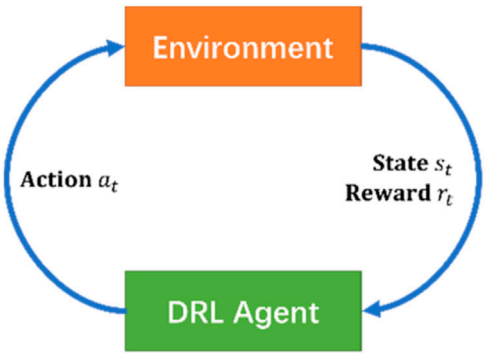


**FIGURE 1 |** Environment–DRL agent interaction loop of reinforcement learning.

5) Ramp power constraints of units:

$$\underline{R}_{AG,i} \le R_{AG,i,t} \le \bar{R}_{AG,i}, \tag{7}$$

where $R_{AG,i,t}$ is the ramp power of unit i at time t and $\underline{R}_{AG,i}$ and $\bar{R}_{AG,i}$ are, respectively, the lower and upper limits of ramp power.

6) Tie-line power deviation constraints:

$$\Delta\underline{P}_T \le \Delta P_{T,t} \le \Delta\bar{P}_T, \tag{8}$$

where $\Delta P_{T,t}$ is tie-line power deviation at time t and $\Delta\underline{P}_T$ and $\Delta\bar{P}_T$ are, respectively, the lower and upper limits of tie-line power deviation.

7) Frequency deviation constraints:

$$\Delta\underline{f} \le \Delta f_t \le \Delta\bar{f}, \tag{9}$$

where $\Delta f_t$ is the frequency deviation at time t and $\Delta\underline{f}$ and $\Delta\bar{f}$ are, respectively, the lower and upper limits of frequency deviation.

# PROXIMAL POLICY OPTIMIZATION ALGORITHM

## The Framework of Reinforcement Learning

A reinforcement learning framework includes an agent and an environment, as illustrated in **Figure 1**, which aims at maximizing a long-term reward through abundant interactions between the agent and the environment. At each step t, the agent observes states $s_t$ and executes action $a_t$; based on its observation and policy, the environment receives action $a_t$, then emits states $s_{t+1}$, and issues a reward $r_{t+1}$ to the agent. Compared with supervised learning, the actions of RL are not labeled, that is, the agent does not know what the correct action is during training and can only be trained through the trial and error approach to explore the environment and maximize its reward.

The interaction between the agent and environment can be modeled by a Markov decision process, which is a standard mathematical formalism of sequential decision problems. A typical Markov decision is denoted by a tuple $< S, A, P, R, \gamma >$, where $S$ is the state space, and it is the complete description of the environment which is represented by a real-valued vector, matrix, or higher-order tensor. $A$ is often called the action space that is also represented by a real-valued vector matrix or higher-order tensor, whereas different environments allow different kinds of actions, that is, discrete and continuous action spaces. $P$ is the transition probability function, and $P(s'|s,a)$ is the probability of transitioning into state $s'$ by taking action $a$ on state $s$. $R$ is the reward function, and $R(s,a,r)$ is the probability of receiving a reward $r$ from action $a$ and state $s$. $\gamma \in [0,1]$ is the reward discount factor. The agent learns to find a policy to maximize the total discounted reward as presented in **(11)**, and $T$ is the number of time steps in each episode.

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-t-1} R_{t+T}. \tag{10}$$

The policy is a rule used by an agent to decide what actions to take, which maps the action from a given state. A stochastic policy is usually expressed as $\pi_\theta(a_t|s_t)$, in which parameter $\theta$ denotes the weights and biases of the neural network in deep reinforcement learning algorithms.

The state value functions $V^\pi(s)$ is the expected return starting from state $s$ following a certain policy as defined in **(12)**, which is used to evaluate the state:

$$V^\pi(s) = \mathrm{E}(G_t|s_t = s). \tag{11}$$

The action-value function $Q^\pi(s,a)$ is the expected return starting from state $s$, taking action $a$, and then following policy $\pi$, denoted as (13), which is utilized to evaluate the action:

$$Q^\pi(s,a) = \mathrm{E}(G_t|s_t = s, a_t = a). \tag{12}$$

The advantage function $A^\pi(s,a)$ corresponding to policy $\pi$ measures the importance of each action in this state, which is mathematically defined as shown in **(14)**:

$$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s). \tag{13}$$

## Proximal Policy Optimization Algorithm With Importance Sampling and Clipping Mechanism

In general, the DRL algorithms can be divided into the value-based, the policy-based, and the actor-to-critic (A2C) framework. The proximal policy optimization (PPO) algorithm follows the A2C framework with an actor network and a critic network.

The main advantage of applying PPO algorithm to the AGC optimization problem is that the new control action decision updates from the policy network does not change too much from the previous policy and can be restrained within the feasible region by the clipping mechanism. During the off-line training process, the PPO also converges faster than other DRL algorithms. Also, during the on-line operations, the PPO generates smoother, less variance, and more predictable sequential decisions, which is desired for the AGC optimization.

The overall structure of the PPO algorithm is presented in **Figure 2**, including an actor network and a critic network. The AGC training environment sends the experience tuples

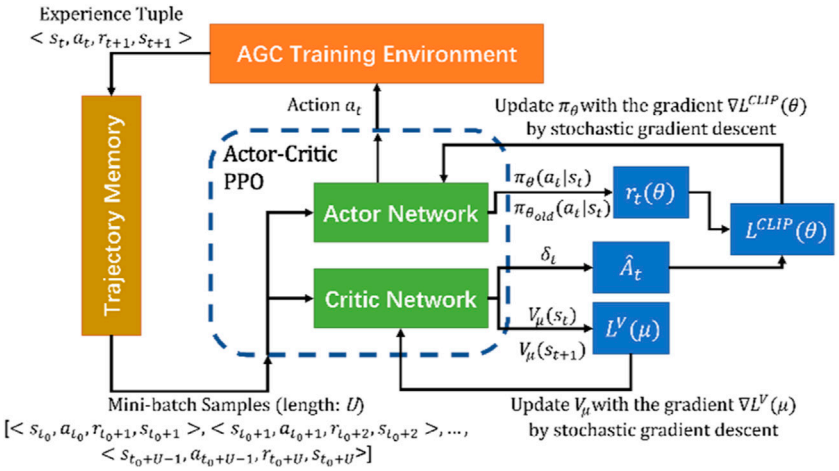


**FIGURE 2 |** Structure of PPO algorithm.

$< s_t, a_t, r_{t+1}, s_{t+1} >$ to the trajectory memory pool to form finite mini-batches of samples and returns to the PPO algorithm.

The actor network and the critic network are realized by deep neural networks (DNNs) with the following equations:

$$O_i = f(W_i I_i + b_i), \quad i = 2, \ldots, n_{layer} + 1, \tag{14}$$

$$OD = O_{n_{layer}}(\ldots O_2(O_1(s_t))), \tag{15}$$

where $I_i$ and $O_i$ represent the input array and output array of the $ith$ layer of the DNN. The layers are connected as (16), $I_{i+1} = O_i$, *and* $I_1 = s_t$. $n_{layer}$ is the number of total layers and $W_i$ and $b_i$ are the weights and bias matrices of the $ith$ layer. The ReLU functions are used as the activation function $f(\cdot)$.

The actor network contains the policy model $\pi_\theta$ with network parameters $< \overline{W_\theta}, \overline{b_\theta} >$. It is responsible for the sequential decisions of AGC optimizations. The rewards are taken in by the critic network $V_\mu$ with parameters $< \overline{W_\mu}, \overline{b_\mu} >$, which is a value function and maps the state $s_t$ to the expected future cumulative rewards.

The conventional policy gradient-based DRL optimizes the following objective function (Wang et al., 2020):

$$L^P(\theta) = \hat{E}\left[log\pi_\theta(a_t|s_t)\hat{A}_t\right], \tag{16}$$

where $\hat{E}[\cdot]$ is the empirical average over a finite mini-batch of samples, $\pi_\theta$ is a stochastic policy, and $\hat{A}_t$ is an estimator of the advantage function at time $t$. In this work, a generalized advantage estimator (GAE) is used to compute the advantage function, which is the discounted sum of temporal difference errors (Schulman et al., 2017).

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1}^V + (\gamma\lambda)^2\delta_{t+2}^V + \ldots + (\gamma\lambda)^{U-t+1}\delta_{U-1}^V, \tag{17}$$

$$\delta_t = r_t + \gamma V_\mu(s_{t+1}) - V_\mu(s_t), \tag{18}$$

where $\gamma \in [0, 1]$ is the discount factor, $\lambda \in [0, 1]$ is the GAE parameter, $U$ is the length of the sampled batch, and $r_t$ is the reward at time $t$. The objective function $L^V(\cdot)$ can be formulated as:

$$L^V(\mu) = \hat{E}[L_t^V(\mu)] = \hat{E}\left[\left|\hat{V}_\mu^{target}(s_t) - V_\mu(s_t)\right|\right], \tag{19}$$

$$\hat{V}_\mu^{target}(s_t) = r_{t+1} + \gamma V_\mu(s_{t+1}), \tag{20}$$

where $\hat{V}_\mu^{target}(\cdot)$ is the target value of time-difference (TD) error. The parameters of the critic network $V_\mu$ can be updated by the stochastic gradient descent algorithm in Duan et al. (2020) according to the gradient $\nabla L^V(\mu)$ with a learning rate $\eta$.

The input of the actor network is the observation state $s_t$, and the outputs are the normal distribution mean value and standard deviation of the actions, that is, the strategy distribution $\pi_\theta(a_t|s_t)$. The importance sampling is used to obtain the expectation of samples gathered from an old policy $\pi_{\theta_{old}}(a_t|s_t)$ under the new policy $\pi_\theta(a_t|s_t)$. This process converts the PPO algorithm from an on-policy method to an off-policy method, which means that the actor network is updated asynchronously to further stabilize the performance of AGC actions. The following surrogate object function is being maximized:

$$\begin{aligned} L^{CPI}(\theta) &= \hat{E}\left[\left(\pi_\theta(a_t|s_t)/\pi_{\theta_{old}}(a_t|s_t)\right)\hat{A}_t\right] \\ &= \hat{E}\left[r_t(\theta)\hat{A}_t\right], \end{aligned} \tag{21}$$

$$s.t. \quad \hat{E}\left[KL\left[\pi_{\theta_{old}}(\cdot|s_t), \pi_\theta(\cdot|s_t)\right]\right] \le \xi, \tag{22}$$

where $KL[\cdot]$ is the Kullback–Leibler (KL) divergence, $r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{old}}(a_t|s_t)$ denotes the ratio of the probability of action $a_t$ under the new and old policies, and $\xi$ is a small number. In order to simplify the penalty by the KL divergence to a first-order algorithm and attain the data efficiency and robustness, a clipping mechanism, $clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t$, is introduced to modify the surrogate objective by clipping $r_t(\theta)$. It removes the incentive of moving $r_t$ outside of the interval $(1-\epsilon, 1+\epsilon)$. The objective function with the $clip(\cdot)$ function is defined as,

$$\begin{aligned} L^{CLIP}(\theta) &= \hat{E}\left[L_t^{CLIP}(\mu)\right] \\ &= \hat{E}\left[\min\left(r_t(\theta)\hat{A}_t,\ clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right]. \end{aligned} \tag{23}$$

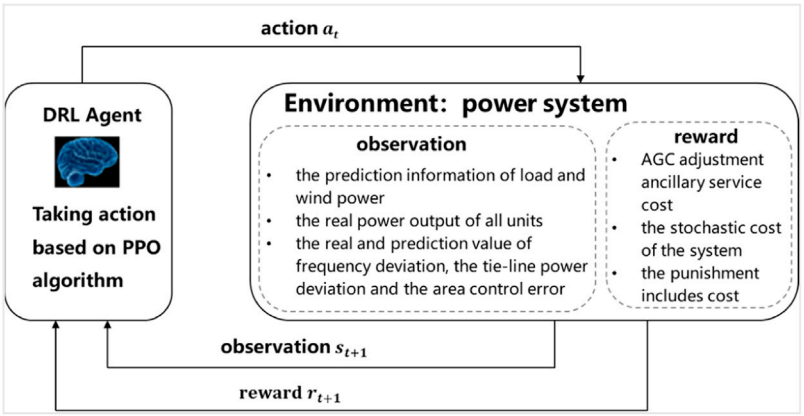


**FIGURE 3 |** Framework of grid environment interacting with an agent.

The importance sampling and clipping function help the PPO DRL algorithm achieve better stability and reliability for AGC online operations, better data efficiency and computation efficiency, and better overall performance.

## AGC OPTIMIZATION STRATEGY BASED ON REINFORCEMENT LEARNING

If the regulation power of each AGC unit is regarded as the action of the agent and the real power system is regarded as the environment of the agent, then the AGC dynamic optimization model considering the uncertainty of wind power can be transformed into a typical random sequential decision problem. Combining the description of the aforementioned AGC dynamic optimization mathematical model, the 15-min control cycle can be divided into a 15-stage Markov process. The framework is shown in **Figure 3**.

The agent can be trained offline through historical data and massive simulations and then applied online in the real power grid. This section mainly focuses on the efficient offline training process of such agent, which introduces the design of several important components.

### State and Action Spaces

State space S: the setting of the state space should consider the factors that may affect the decision as much as possible. In this work, the state space is determined as a vector of system information representing the current system condition at time t and prediction system information at time $t+1$. Specifically, the former includes the real power output of all units (AGC and non-AGC unit) $P^r_{G,t}$, the frequency deviation $\Delta f^r_t$, the power deviation of the tie-line $\Delta P^r_{T,t}$, and the area control error $ACE^r_t$. The latter includes the prediction system information of load $P^f_{l,t+1}$, wind power $P^f_{w,t+1}$, frequency deviation $\Delta f^f_{t+1}$, power deviation of tie-line $\Delta P^f_{T,t+1}$, and area control error $ACE^f_{t+1}$. It is set as follows:

$$S: \left\{ \begin{matrix} P^r_{G,t}, \Delta f^r_t, \Delta P^r_{T,t}, ACE^r_t, P^f_{l,t+1}, \\ P^f_{w,t+1}, \Delta f^f_{t+1}, \Delta P^f_{T,t+1}, ACE^f_{t+1} \end{matrix} \right\}. \tag{24}$$

Action space A: action space is the decision variable in the optimization model, including the ramp direction and ramp power. In this article, to avoid the lack of generality, the action is defined as power increments of AGC units $\Delta P^a_{AG,t}$ at each optimization time, which are subjected to the ramp power limits of corresponding AGC units. $A$ is set as

$$A: \left\{ \Delta P^a_{AG,t} \right\}. \tag{25}$$

### Reward Function Design

The design of reward function is crucial in DRL. It generates reward $r_t$ at time $t$ in each decision cycle, which evaluates the agent's actions based on the AGC control performance under the impacts of uncertainties in the system variables. In this work, the values of load, wind generations, frequency deviations, the tie-line power deviations, and ACE are used to formulate the reward function, which consists of cost terms, punishment terms, and performance terms. The reward $r_t$ is calculated by the formula:

$$r_t = F_{cost} + r_{penel} + f_{cps}, \tag{26}$$

where the cost term $F_{cost}$ represents the total cost of the system. It includes AGC adjustment ancillary service cost and the load shedding cost, which are calculated as follows:

$$F_{cost} = -c_1 f_{AGC} - c_1 P^2_{c,t}, \tag{27}$$

where $c_1$ and $c_2$ are the corresponding cost coefficients. The load shedding $P_{c,t}$ must be set reasonably as follows:

$$P_{c,t} = \left\{ \begin{matrix} 0, & |\Delta f| \le 0.2, \\ (|\Delta f| - 0.1) \cdot \Delta P_t, & |\Delta f| > 0.2. \end{matrix} \right. \tag{28}$$

Here, the real power deviations $\Delta P_t$ are utilized to reflect the stochastic process caused by the load and wind power fluctuations. At time $t$, the power deviations in the system are calculated as follows:

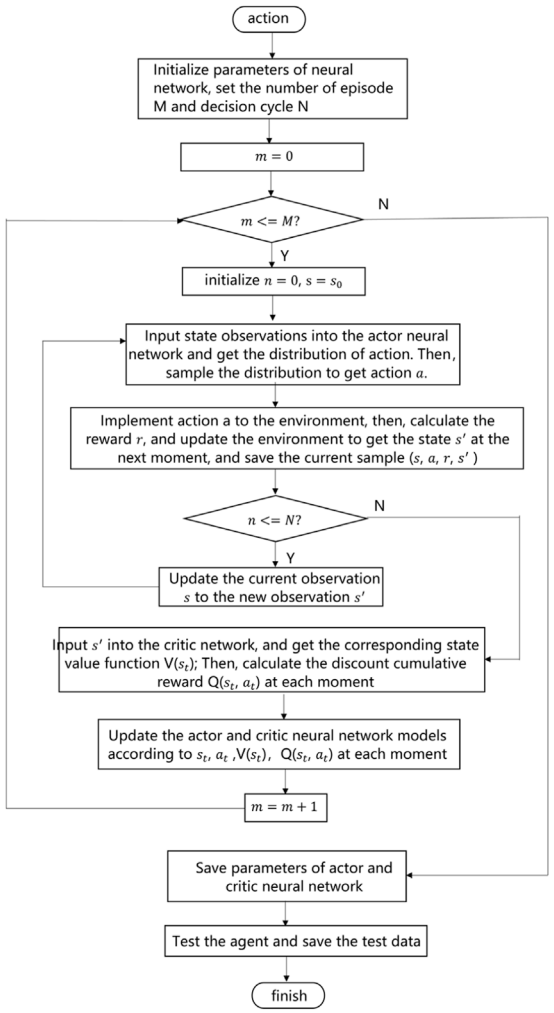


**FIGURE 4 |** AGC dynamic optimization problem based on the PPO algorithm.

$$\Delta P_t = \sum_{i=1}^{N} P_{G,i,t} + P^r_{w,t} - P^r_{L,t} - P^r_{T,t} - P^r_{loss,t}, \tag{29}$$

where $N$ is the total number of thermal power units in the system, including AGC and non-AGC units, $P_{G,i,t}$ is the power output of thermal power unit $i$ at $t$ period, and $P^r_{w,t}$, $P^r_{L,t}$, and $P^r_{T,t}$ are the power output of wind power, load, and tie-line power. Note that the power flowing out of the system is taken to be positive. $P^r_{loss,t}$ is the system loss at time $t$.

The power output of any unit (AGC or non-AGC unit) relates to the frequency deviation, tie-line power deviation, and ACE. Taking an interconnection power grid of two areas as an example, the system contains region $A$ and region $B$. The control strategy of the two areas is the tie-line bias frequency control. It is assumed that $\Delta P_{LA,t}$ and $\Delta P_{LB,t}$, and $\Delta P_{GA,t}$ and $\Delta P_{GB,t}$ are the change of load and the change of power output of units in region $A$ and region $B$ at time $t$, respectively, and $K_A$ and $K_A$ are the frequency regulation constants of region $A$ and region $B$. We define $\Delta P_{A,t}$ and $\Delta P_{B,t}$ as power imbalance of the two areas which can be calculated as follows:

$$\Delta P_{A,t} = \Delta P_{LA,t} - \Delta P_{GA,t}, \tag{30}$$

$$\Delta P_{B,t} = \Delta P_{LB,t} - \Delta P_{GB,t}. \tag{31}$$

Frequency deviation, tie-line power deviation, and area control error can be calculated as follows:

$$\Delta f_t = -\frac{\Delta P_{A,t} + \Delta P_{B,t}}{K_A + K_B}, \tag{32}$$

$$\Delta P_{T,t} = \frac{K_A \cdot \Delta P_{B,t} - K_B \cdot \Delta P_{A,t}}{K_A + K_B}, \tag{33}$$

$$e_{ACE,t} = \Delta P_{T,t} - 10B \cdot \Delta f_t, \tag{34}$$

where $B$ is the equivalent frequency regulation constant for the control area in MW/0.1Hz and the value is negative.

The punishment term $r_{penel}$ formulates the operation and control limits in AGC dynamic optimization, including generation unit power output limits, CPS1, frequency deviation limits, and tie-line power deviation limits and given as:

$$r_{penel} = r_1 + r_2 + r_3 + r_4. \tag{35}$$

The AGC units participate in both primary and secondary frequency control; thus, the outputs of AGC units at time $t + 1$ are calculated as

$$P'_{AG,i,t+1} = PAG, i, t + 1 + \Delta P^a_{AG,t} - K_{Gi}\left(\Delta f^r_{t+1} - \Delta f^r_t\right), \tag{36}$$

where $\Delta P^a_{AG,t}$ is the regulated power of AGC unit $i$ at time $t$, that is, the power increment of secondary frequency control; and $K_{Gi}\left(\Delta f^r_{t+1} - \Delta f^r_t\right)$ is the primary frequency control power of AGC unit $i$, where $K_{Gi}$ is the frequency regulation constant of unit $i$, and $\Delta f^r_{t+1}$ and $\Delta f^r_t$ are system frequency deviations at time $t + 1$ and $t$, respectively.

Accordingly, the power outputs of non-AGC units at time $t +$ 1 are calculated as

$$P'_{NG,i,t+1} = P_{NG,i,t} - K_{Gi}\left(\Delta f^r_{t+1} - \Delta f^r_t\right). \tag{37}$$

The outputs of AGC and non-AGC units are subjected to the corresponding maximum and minimum power limits:

$$P_{AG,i,t+1} = \begin{cases} P_{G,i,min}, P'_{G,i,t+1} < P_{G,i,min}, \\ P_{G,i,max}, P'_{G,i,t+1} > P_{G,i,\max}, \\ P'_{G,i,t+1} \text{ else}, \end{cases} \tag{38}$$

$$r_1 = \begin{cases} 0, & P_{AG,i,\min} < P_{G,i,t} < P_{AG,i,\max}, \\ k_1, & \text{else}, \end{cases} \tag{39}$$

where $k_1$ is the punishment coefficient. The CPS1-related punishment term is formulated as,

$$r_2 = \begin{cases} 0, & K_{cps1} \geq 200\%, \\ -k_2\left|e_{ACE} - e^*_{ACE}\right|, & 100\% \leq K_{cps1} \leq 200\%, \\ -k_3\left|K_{cps1} - K^*_{cps1}\right|, & K_{cps1} < 100\%, \end{cases} \tag{40}$$

where $k_2$ and $k_3$ are the punishment coefficients of ACE and CPS1 and $e^*_{ACE}$ and $K^*_{cps1}$ are, respectively, ideal values of ACE and CPS1. In this article, the ideal values of ACE and CPS1 are 0 and 200%.

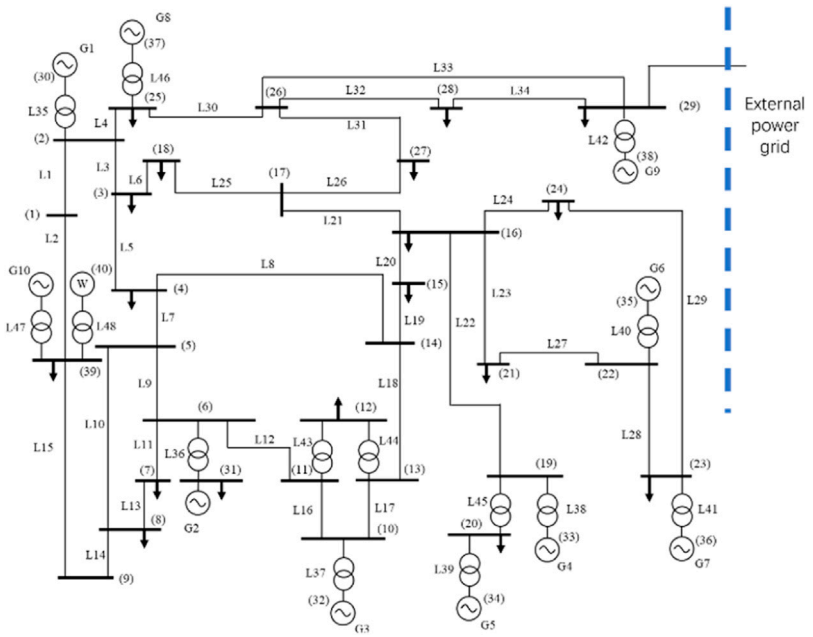


**FIGURE 5 |** Modified IEEE 39 bus system.

**TABLE 1 |** Information of AGC units.

| Symbol | Quantity | Unit 1 | Unit 2 | Unit 3 |
|---|---|---|---|---|
| $n_{AG,i}$ | Bus number | 31 | 38 | 39 |
| $\bar{P}_{N,i}$, MW | Rated power | 800 | 860 | 1,100 |
| $\underline{R}_{AG,i}$ and $\bar{R}_{AG,i}$, MW/min | Limits of lower and upper ramp power | -30 and 30 | -45 and 45 | -60 and 60 |
| $k_{AG,i}$, ¥/(MW min) | Cost coefficient of frequency regulation | 0.5 | 0.5 | 0.25 |
| $K_{AG,i}$ (per unit) | Frequency regulation constant of unit | 25 | 25 | 25 |

The following two functions denote the frequency deviation and tie-line power transfer deviation punishments, respectively:

$$r_3 = \begin{cases} 0, & \Delta f_{\min} \le \Delta f \le \Delta f_{\max}, \\ k_4, & \text{else}, \end{cases} \tag{41}$$

$$r_4 = \begin{cases} 0, & \Delta P_{T\min} \le \Delta P_{T,t} \le \Delta P_{T\max}, \\ k_5, & \text{else}, \end{cases} \tag{42}$$

where $k_4$ and $k_5$ are the corresponding punishment coefficients.

In this article, we added an additional performance evaluation term $f_{cps}$ with a coefficent $c_3$ to the reward function which makes the PPO-based DRL algorithm has the capability of further improving the long-term AGC performance:

$$f_{cps} = -c_3\left(2 - K_{cps1}\right)^2. \tag{43}$$

## Other Parameter Setting

State transition probability $P$: in this work, the reinforcement learning algorithm based on the model-free method is utilized, so the state of the agent at the next time and rewards can be obtained by the interaction with the environment, and they make up state transition probability $P$ including environmental stochasticity.

Discount factor $\gamma$ ($\gamma \epsilon [0, 1]$) determines the importance of rewards in future to current reward. When $\gamma = 0$, it means that the impact of current decisions on the future system operating status is not considered, and only the operating cost of the current control period is optimized; when $\gamma = 1$, it means that the impact of current decisions on the operating status of the system at every moment in the future is equally considered. For AGC dynamic optimization control, the decision at the current moment will have an important impact on the future operating state of the system, and the closer the distance to the current decision period, the greater the impact.

## Detailed PPO Training Algorithm in Solving the AGC Dynamic Optimization Problem

Based on the aforementioned analysis, this article transforms the AGC dynamic optimization problem into a sequential decision issue and utilizes the PPO deep reinforcement learning algorithm to solve the proposed problem. The

**TABLE 2 |** Information of the test system.

| Symbol | Quantity | Value |
|---|---|---|
| $f_N$/Hz | Nominal frequency | 50 |
| $f_O$/Hz | Initial frequency | 50 |
| $P_{T,N}$/MW | Nominal power of tie-line | 100 |
| $P_{T,O}$/MW | Initial power of tie-line | 100 |
| $\varepsilon_1$ and $\varepsilon_{15}$ | Target bound for the 12-month RMS value of the 1-/15-minute average frequency error, in Hz | 0.04 and 0.021 |
| $B$ and $B_s$ | Target bound for the 12-month RMS value of the 1-/15-min average frequency error, in MW/0.1 Hz | −38 and 50 |
| $\Delta\underline{f}$ and $\Delta\bar{f}$, Hz | Limits of frequency deviation | −0.05 and 0.05 |
| $\Delta\underline{P}_T$ and $\Delta\bar{P}_T$, MW | Limits of transmission power deviation of tie-line | −20 and 10 |

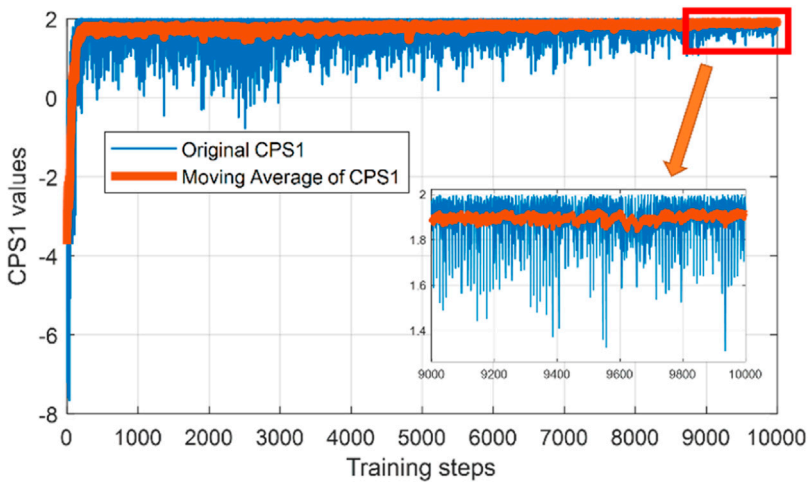


**FIGURE 6 |** Results of CPS1 index in the training process using 10,000 episodes.

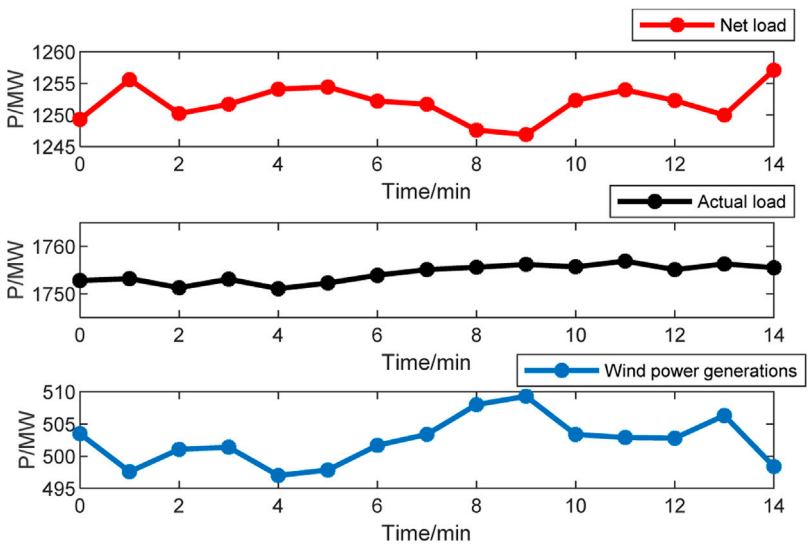


**FIGURE 7 |** Curve of load with wind power fluctuations.

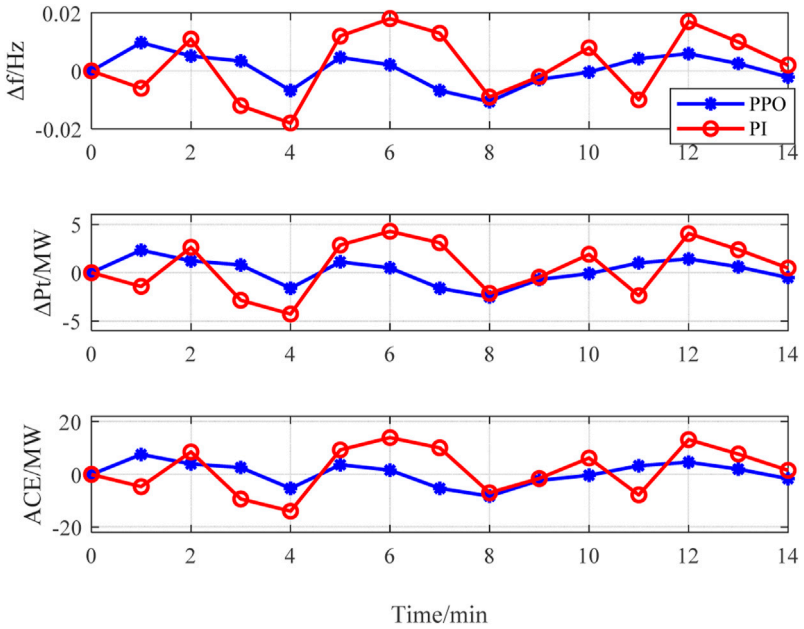


**FIGURE 8 |** Results of the optimization method and PPO algorithm.

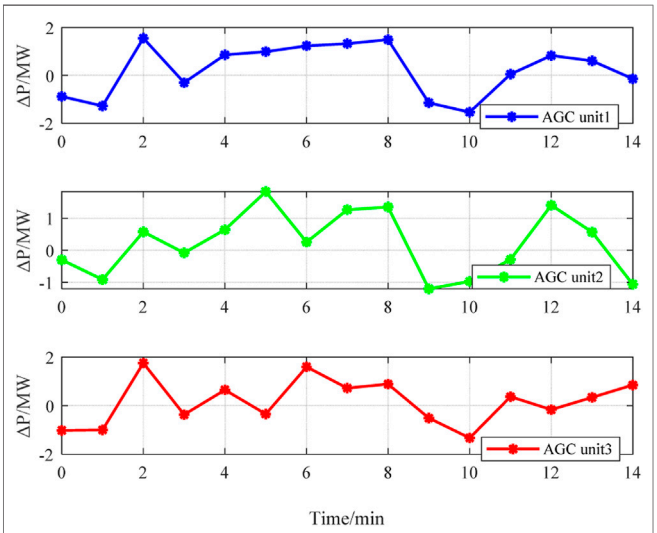


**FIGURE 9 |** Results of AGC regulation curves.

AGC dynamic optimization problem based on the PPO algorithm is shown in **Figure 4**. The specific process is described as follows:

1) Initialize the weight and bias of the neural network, actor and critic neural network learning rate, reward discount factor $\gamma$, and hyperparameters $\varepsilon$ and other parameters. Set the number of episode $M$ and decision cycle $N$.
2) Initialize the initial observation value at the first moment from the power system environment.
3) Input state observation $s$ into the actor neural network and get the distribution of action $a$. Then, sample the distribution to get action $a$ by importance sampling.
4) Implement action $a$ to the environment, then calculate the reward $r$, and update the environment to get the state $s'$ at the next moment, and save the current sample $(s, a, r, s')$. Update the current observation $s$ to the new observation $s'$.
5) Input $s'$ into the critic network, and get the corresponding state value function $V(s_t)$. Then, calculate the discount cumulative reward $Q(s_t, a_t)$ at each moment based on **(35)**,

$$Q(s_t, a_t) = r_t + \gamma r_{t+1} + \cdots + \gamma^{T-t-1} r_{T-1} + \gamma^{T-t} V(s_T). \quad (44)$$

6) Update the actor and critic neural network models according to $s_t$, $a_t$, $V(s_t)$, and $Q(s_t, a_t)$ at each moment.
7) Repeat steps 2–6 until the number of training episodes is equal to the set number $M$.
8) Save the parameters of actor and critic neural networks. Utilize the trained agent on the test data.

# CASE STUDY

## Test System and Data

In this article, the PPO agent for AGC dynamic optimization control is tested on the modified IEEE 39 bus system model which includes three AGC units and seven non-AGC units. The tie-line is connected to bus 29, and a wind farm with 130 MW

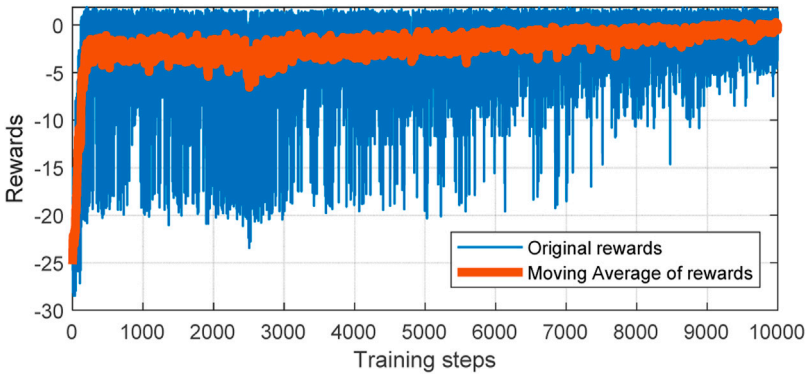


**FIGURE 10 |** Performance of the PPO agent in the training process using 10,000 episodes.

installations is connected to bus 39. A single-line diagram of the system is shown in **Figure 5**.

The forecasting and actual data of load and wind power come from New England power grid[1]. The basic parameters of three AGC units and test system are shown in **Tables 1**, **2**. The control period is set at 15 min, and it is assumed that the deviation of frequency and tie-line transmission power at the initial time are 0.

The action space refers to the regulation power of AGC units at each optimization moment, which is determined by ramp power limits of each AGC unit. The per unit action space of the three AGC units is set as follows:

$$\begin{aligned} A1 &= \{-0.3, 0.3\}, \\ A2 &= \{-0.45, 0.45\}, \\ A3 &= \{-0.6, 0.6\}. \end{aligned} \tag{45}$$

In addition, the state space dimension is 35 according to the preceding description, which includes information on 19 forecasting loads at time t+1, actual output power of 10 units at time t, and actual and forecasting values of system frequency deviation, transmission power deviation, and ACE separately at time t and t+1. The dimensions of state space and action space, respectively, correspond to the neural numbers of input and output layers. Therefore, this work sets up three hidden layers both in actor and critic neural networks, and the number of neurons in each layer is 64, 128, and 32, respectively. The activation function in each hidden layer is the ReLU function. A larger learning rate $\alpha$ accelerates the convergence of the algorithm, while a smaller $\alpha$ tends to enhance the stability. In this article, learning rate $\alpha$ both in actor and critic networks is set to 0.0001.

## Evaluation of the Test Results

Based on the preceding model and significant parameters, the PPO agent is coded using the TensorFlow framework with Python 3.7. The results of CPS1 index are shown in **Figure 6**.

The $x$-axis represents the number of episodes being trained, while the $y$-axis represents the value of CPS1 index in each episode. It can be observed that the CPS1 values of the first few hundreds of episodes are relatively low and unstable. As training episodes increase, CPS1 values are kept within a stable range around 191.3%, which fits CPS. In comparison, the deep Q learning (DQL) algorithm and the duel deep Q learning (DDQL) algorithm are also implemented. The average CPS1 values are 187.4 and 184.5%. This shows that the PPO architecture for AGC unit dynamic optimization proposed in this article can effectively learn the growing uncertainties in the power system. Once the agent is trained, it can make proper decisions based on its trained strategy combined with environmental observation data feedback. Specifically, the agent trained in this work receives data from the power system, including actual information of unit output power, frequency, tie-line transmission power, ACE, and forecasting information of load, wind power, frequency, tie-line transmission power, and ACE as its observation, and then makes decisions for the regulation power of AGC units at time t, that is, advanced control of AGC units, in order to reduce the frequency deviation at time t+1.

In addition, taking a typical control period of the system as an example, **Figure 7** shows the actual load, the wind power generations, and the net load by subtracting the wind generations from the actual load. The load at each bus is allotted in proportion to the load of the original IEEE-39 node system.

Using PI hysteresis control and PPO algorithm for frequency control in this period, the results of system frequency deviation, transmission power deviation of tie-line, and ACE are represented in **Figures 8A–C**, respectively.

It is observed in **Figure 8A** that the outputs of both the PPO agent and optimization method can meet the requirements of frequency deviation (i.e., ±0.05 Hz). Moreover, maximum frequency deviation of the system controlled by the PPO agent is 0.0175 Hz, which is superior to -0.044 Hz that is controlled by the optimization method. This demonstrated that the dynamic optimization strategy of AGC units based on PPO algorithm is able to mitigate the frequency fluctuation of the system efficiently by advanced control of AGC units.

[1]https://www.iso-ne.com/isoexpress/web/reports.

**Figure 8B** shows the transmission power deviation of tie-line. The power deviation under the optimization method is relatively large, which contains three times the off-limit conditions owing to the AGC resources in the system that are insufficient. While under PPO agent controller, the power deviation fluctuation is smaller without over-limit time. It proves that the system operation is more stable when using the PPO agent controller optimization method. In addition, as shown in **Figure 8C**, the agent performs much better than the optimization method when calculating the values of ACE. **Figure 9** shows the AGC power regulation curve of all AGC units.

## Convergence of Algorithm

In the training process, the cumulative reward of each episode is recorded. Then, the results and the filtered curve are shown in **Figure 10**.

Due to the loads and wind power fluctuations being different in each episode, the needs of frequency regulation in each episode are also different. Therefore, it is normal that there exists slight oscillation of cumulative rewards for each episode. As the training process continues, the cumulative rewards tend to converge as shown in Fig.

# CONCLUSION AND FUTURE WORKS

To effectively mitigate frequency control issues under growing uncertainties, this article presents a novel solution, the PPO architecture for AGC dynamic optimization, which transformed the traditional optimization problem into a Markov decision process and utilized deep reinforcement learning algorithm for frequency control.

Through the design of state, action, and reward functions, the continuous multiple time step control can be implemented with the goal of maximizing cumulative rewards. The model utilized the way of interaction between the agent and the environment to improve the parameters, which is adaptive to the uncertainties in the environment and avoids the modeling of uncertain variables. The model proposed in this article is tested on the modified IEEE 39 bus system. The results demonstrate that the PPO architecture for AGC dynamic optimization can achieve the goal of frequency control with satisfactory performance compared to other methods. It is verified that the method proposed in this article can effectively solve the stochastic disturbance problem caused by large-scale integration of renewable energy into power grid and ensure the safety and stability of system frequency.

From the lessons learned in this work, the directions of future works are discussed here. First, the deep learning-based algorithms suffered from poor interpretability, which is undesired for control engineering problems. With the developments of explainable artificial intelligence, future works are needed on this direction. Second, better exploration mechanisms for DRL algorithms need to be developed to further improving the training efficiency and avoiding the local optimal solutions.

# DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

# AUTHOR CONTRIBUTIONS

ZL and JL wrote the manuscript. The simulations were performed by JL and PZ. ZD and YZ provided major and minor revisions to the final version of the submitted manuscript. All of the aforementioned authors contributed to the proposed methodology.

# FUNDING

This work is supported by the Fundamental Research Funds for the Central Universities under Grant No. 2021JBM027 and the National Natural Science Foundation of China under Grant No. 52107068. Both funds are supportive to open access publication fees.

# REFERENCES

Abdel-Magid, Y. L., and Dawoud, M. M. (1996). Optimal AGC Tuning with Genetic Algorithms. *Electr. Power Syst. Res.* 38 (3), 231–238. doi:10.1016/s0378-7796(96)01091-7

Atic, N., Rerkpreedapong, D., Hasanovic, A., and Feliachi, A. (2003). "NERC Compliant Decentralized Load Frequency Control Design Using Model Predictive Control[C]," in Power Engineering Society General Meeting (Piscataway, NJ, USA: IEEE).

Banakar, H., Luo, C., and Ooi, B. T. (2008). Impacts of Wind Power Minute-To-Minute Variations on Power System Operation. *IEEE Trans. Power Syst.* 23 (1), 150–160. doi:10.1109/tpwrs.2007.913298

Beaufays, F., Widrow, B., Abdel-Magid, Y., and Widrow, B. (1994). Application of Neural Networks to Load-Frequency Control in Power Systems. *Neural Netw.* 7 (1), 183–194. doi:10.1016/0893-6080(94)90067-1

Bohn, E., and Miniesy, S. (1972). Optimum Load-Frequency Sampled-Data Control with Randomly Varying System Disturbances. *IEEE Trans. Power Apparatus Syst.* PAS-91 (5), 1916–1923. doi:10.1109/tpas.1972.293519

Chang, C. S., Fu, W., and Wen, F. (1998). Load Frequency Control Using Genetic-Algorithm Based Fuzzy Gain Scheduling of Pi Controllers. *Electr. Mach. Power Syst.* 26 (1), 39–52. doi:10.1080/07313569808955806

Concordia, C., and Kirchmayer, L. K. (1953). Tie-Line Power and Frequency Control of Electric Power Systems. *Power Apparatus Syst. Part III Trans. Am. Inst. Electr. Eng.* 72 (2), 562–572. doi:10.1109/aieepas.1953.4498667

Dahiya, P., Sharma, V., and Naresh, R. (2016). Automatic Generation Control Using Disrupted Oppositional Based Gravitational Search Algorithm Optimised Sliding Mode Controller under Deregulated Environment. *IET Gener. Transm. & Distrib.* 10 (16), 3995–4005. doi:10.1049/iet-gtd.2016.0175

Duan, J., Shi, D., Diao, R., Li, H., Wang, Z., Zhang, B., et al. (2020). Deep-Reinforcement-Learning-Based Autonomous Voltage Control for Power Grid Operations. *IEEE Trans. Power Syst.* 35 (1), 814–817. doi:10.1109/TPWRS.2019.2941134

Elgerd, O. I., and Fosha, C. E. (2007). Optimum Megawatt-Frequency Control of Multiarea Electric Energy Systems[J]. *IEEE Trans. Power Apparatus Syst.* PAS-89 (4), 556–563.

Erschler, J., Roubellat, F., and Vernhes, J. P. (1974). Automation of a Hydroelectric Power Station Using Variable-Structure Control Systems. *Automatica* 10 (1), 31–36. doi:10.1016/0005-1098(74)90007-7

Feliachi, A., and Rerkpreedapong, D. (2005). NERC Compliant Load Frequency Control Design Using Fuzzy Rules. *Electr. Power Syst. Res.* 73 (2), 101–106. doi:10.1016/j.epsr.2004.06.010

Jaleeli, N., VanSlyck, L. S., Ewart, D. N., Fink, L. H., and Hoffmann, A. G. (2002). Understanding Automatic Generation Control[J]. *IEEE Trans. Power Syst.* 7 (3), 1106–1122. doi:10.1109/59.207324

Jaleeli, N., and Vanslyck, L. S. (1999). NERC's New Control Performance Standards. *IEEE Trans. Power Syst.* 14 (3), 1092–1099. doi:10.1109/59.780932

Khodabakhshian, A., and Edrisi, M. (2004). A New Robust PID Load Frequency Controller. *Control Eng. Pract.* 19 (3), 1528–1537. doi:10.1016/j.conengprac.2007.12.003

Mcnamara, P., and Milano, F. (2017). Model Predictive Control Based AGC for Multi-Terminal HVDC-Connected AC Grids[J]. *IEEE Trans. Power Syst.* 2017, 1.

Olmos, L., de la Fuente, J. I., Zamora Macho, J. L., Pecharroman, R. R., Calmarza, A. M., and Moreno, J. (2004). New Design for the Spanish AGC Scheme Using an Adaptive Gain Controller. *IEEE Trans. Power Syst.* 19 (3), 1528–1537. doi:10.1109/tpwrs.2004.825873

Pan, I., and Das, S. (2016). Fractional Order AGC for Distributed Energy Resources Using Robust Optimization. *IEEE Trans. Smart Grid* 7 (5), 2175–2186. doi:10.1109/TSG.2015.2459766

Sahu, B. K., Pati, S., Mohanty, P. K., and Panda, S. (2015). Teaching-learning Based Optimization Algorithm Based Fuzzy-PID Controller for Automatic Generation Control of Multi-Area Power System. *Appl. Soft Comput.* 27, 240–249. doi:10.1016/j.asoc.2014.11.027

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal Policy Optimization Algorithms[J]. *arXiv:1707.06347*.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. (2015). High-dimensional Continuous Control Using Generalized Advantage Estimation. *arXiv Prepr. arXiv:1506.02438*.

Sun, L. (2017). "Analysis and Comparison of Variable Structure Fuzzy Neural Network Control and the PID Algorithm," in 2017 Chinese Automation Congress (CAC), Jinan, 3347–3350. doi:10.1109/CAC.2017.8243356

Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 3–23.

Talaq, J., and Al-Basri, F. (1999). Adaptive Fuzzy Gain Scheduling for Load Frequency Control. *IEEE Trans. Power Syst.* 14 (1), 145–150. doi:10.1109/59.744505

Venayagamoorthy, G. K., Sharma, R. K., Gautam, P. K., and Ahmadi, A. (2016). Dynamic Energy Management System for a Smart Microgrid. *IEEE Trans. Neural Netw. Learn. Syst.* 27 (8), 1643–1656. doi:10.1109/TNNLS.2016.2514358

Wang, B., Li, Y., Ming, W., and Wang, S. (2020). Deep Reinforcement Learning Method for Demand Response Management of Interruptible Load. *IEEE Trans. Smart Grid* 11 (4), 3146–3155. doi:10.1109/TSG.2020.2967430

Wen, Z., O'Neill, D., and Maei, H. (2015). Optimal Demand Response Using Device-Based Reinforcement Learning. *IEEE Trans. Smart Grid* 6 (5), 2312–2324. doi:10.1109/TSG.2015.2396993

Xi, L., Zhou, L., Xu, Y., and Chen, X. (2021). A Multi-step Unified Reinforcement Learning Method for Automatic Generation Control in Multi-Area Interconnected Power Grid. *IEEE Trans. Sustain. Energy* 12 (2), 1406–1415. doi:10.1109/TSTE.2020.3047137

Yamashita, K., and Taniguchi, T. (1986). Optimal Observer Design for Load-Frequency Control. *Int. J. Electr. Power & Energy Syst.* 8 (2), 93–100. doi:10.1016/0142-0615(86)90003-7

Yan, W., Zhao, R., Zhao, X., Li, Y., Yu, J., and Li, Z. (2012). Dynamic Optimization Model of AGC Strategy under CPS for Interconnected Power System[J]. *Int. Rev. Electr. Eng.* 7 (5PT.B), 5733–5743.

Yu, T., Zhou, B., Chan, K. W., Chen, L., and Yang, B. (2011). Stochastic Optimal Relaxed Automatic Generation Control in Non-markov Environment Based on Multi-step Q(λ) Learning. *IEEE Trans. Power Syst.* 26 (3), 1272–1282. doi:10.1109/TPWRS.2010.2102372

Zeynelgil, H. L., Demiroren, A., and Sengor, N. S. (2002). The Application of ANN Technique to Automatic Generation Control for Multi-Area Power System. *Int. J. Electr. Power & Energy Syst.* 24 (5), 345–354. doi:10.1016/s0142-0615(01)00049-7

Zhang, X. S., Li, Q., Yu, T., and Yang, B. (2016). Consensus Transfer Q-Learning for Decentralized Generation Command Dispatch Based on Virtual Generation Tribe. *IEEE Trans. Smart Grid* 9 (3), 1. doi:10.1109/TSG.2016.2607801

Zhang, X. S., Yu, T., Pan, Z. N., Yang, B., and Bao, T. (2018). Lifelong Learning for Complementary Generation Control of Interconnected Power Grids with High-Penetration Renewables and EVs. *IEEE Trans. Power Syst.* 33 (4), 4097–4110. doi:10.1109/TPWRS.2017.2767318

Zhang, X., Tan, T., Zhou, B., Yu, T., Yang, B., and Huang, X. (2021). Adaptive Distributed Auction-Based Algorithm for Optimal Mileage Based AGC Dispatch with High Participation of Renewable Energy. *Int. J. Electr. Power & Energy Syst.* 124, 106371. doi:10.1016/j.ijepes.2020.106371

Zhang, X., Xu, Z., Yu, T., Yang, B., and Wang, H. (2020). Optimal Mileage Based AGC Dispatch of a GenCo. *IEEE Trans. Power Syst.* 35 (4), 2516–2526. doi:10.1109/TPWRS.2020.2966509

Zhang, X., Yu, T., Yang, B., and Jiang, L. (2021). A Random Forest-Assisted Fast Distributed Auction-Based Algorithm for Hierarchical Coordinated Power Control in a Large-Scale PV Power Plant. *IEEE Trans. Sustain. Energy* 12 (4), 2471–2481. doi:10.1109/TSTE.2021.3101520

Zhang, X., Yu, T., Yang, B., and Li, L. (2016). Virtual Generation Tribe Based Robust Collaborative Consensus Algorithm for Dynamic Generation Command Dispatch Optimization of Smart Grid. *Energy* 101, 34–51. doi:10.1016/j.energy.2016.02.009

Zhao, X., Ye, X., Yang, L., Zhang, R., and Yan, W. (2018). Chance Constrained Dynamic Optimisation Method for AGC Units Dispatch Considering Uncertainties of the Offshore Wind Farm[J]. *J. Eng.* 2019 (16), 2112. doi:10.1049/joe.2018.8558

Zhou, Y., Zhang, B., Xu, C., Lan, T., Diao, R., Shi, D., et al. (2020). A Data-Driven Method for Fast AC Optimal Power Flow Solutions via Deep Reinforcement Learning. *J. Mod. Power Syst. Clean Energy* 8 (6), 1128–1139. doi:10.35833/MPCE.2020.000522

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.