A novel deep reinforcement learning enabled agent for pumped storage hydro-wind-solar systems voltage control

Huang, Qin; Hu, Weihao; Zhang, Guozhou; Cao, Di; Liu, Zhou; Huang, Qi; Chen, Zhe

Published in:
IET Renewable Power Generation

DOI (link to publication from Publisher):
10.1049/rpg2.12311

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Received: 3 July 2021 | Revised: 25 September 2021 | Accepted: 27 September 2021

IET Renewable Power Generation

DOI: 10.1049/rpg2.12311

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

A novel deep reinforcement learning enabled agent for pumped storage hydro-wind-solar systems voltage control

Qin Huang[1] | Weihao Hu[1] | Guozhou Zhang[1] | Di Cao[1] | Zhou Liu[2] | Qi Huang[1] | Zhe Chen[2]

[1] School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China

[2] Department of Energy Technology, Aalborg University, Pontoppidanstraede 111, Aalborg, Denmark

Correspondence
Weihao Hu, School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611730, China.
Email: whu@uestc.edu.cn

Funding information
National Key Research and Development Program of China, Grant/Award Number: 2018YFE0127600

Abstract

With the large-scale penetration of wind and solar energies in the power system, the randomness of this renewable energy increases the non-linear characteristics and uncertainty of the system, which causes a mismatch between renewable energy generation and load demand and it will badly affect the bus voltage control of distribution network. In this context, this study applies pumped storage hydroelectric (PSH) which tracks the load variation rapidly, operate flexibly and reliably to balance the power of the system to minimize the bus voltage deviation. Moreover, to obtain the optimal control policy of PSH, a deep-reinforcement-learning algorithm, that is, deep deterministic policy gradient, is utilized to train the agent to address the continuous transformation of the pumped storage hydro-wind-solar (PSHWS) system. The performance of a well-trained agent was evaluated on the IEEE 30-bus power system. Simulation results show that the proposed method achieves an improvement of 21.8% in cumulative deviation per month, which implies that it can keep the system operating in a safe voltage range more effectively.

1 | INTRODUCTION

To address global climate change, the development and utilization of renewable energy sources (RESs) have become a consensus to replace fossil-based energy worldwide. In particular, because of their clean, widely distributed, and large-scale characteristics, wind and solar energies are of general interest [1], and governments around the world have formulated various policies to support their development [2]. Over the past decade, wind power (WP) and solar power capacities increased rapidly. By the end of 2020, the global installed capacities of WP and solar power have surged to 742.69 GW [3] and 716.15 GW [4], respectively.

However, WP and solar power generation are affected by climate change, and the characteristics of uncertainty, randomness, and uncontrollability challenge the stability of the systems through aspects such as frequency stability [5], small signal stability [6], and voltage stability [7]. In particular, the voltage overrun problem is particularly serious. It limits the utilization rate of photovoltaic (PV) power generation, and reduces the power quality and reliability of power systems [8]. Therefore, the voltage problem has become an urgent problem to be solved, currently receiving widespread attention.

In fact, voltage problem was completed by reasonably balancing the reactive power flow of the system. Various model-based optimization strategies have been developed in order to implement reactive power dispatch. In [9], a time-series power flow (TSPF) method was proposed to capture the reactive power dispatch of a doubly fed induction generator (DFIG) wind farm at the worst-case point, which was used to evaluate the impact of WP generation on power system voltage stability. However, the TSPF method depends on the prediction accuracy of the worst-case point model, prediction errors have a significant influence on the evaluation result. In [10], the reactive power capability of the inverter of the dispatching photovoltaic system is used to alleviate voltage fluctuation. In [11], a distributed optimal active and reactive power control (DARPC) strategy based on the alternating direction method of multipliers (ADMM) is applied for wind farms (WFs). In [12], both active and reactive power management methods which known as dynamic voltage

support (DVS) was adopted to improve the short-term voltage stability. Although the reactive power compensation methods mentioned in [9–12] contribute to stabilizing the system voltage. DFIG, PV inverter or WF provide reactive power, which are suitable for large wind farms or photovoltaic power plants, the above methods were model-based optimization, the modelling process is cumbersome and complex with certain errors, and the compensation cost is expensive. However, the compensation equipment need to have the characteristics of high energy storage capability, fast start-up, and low power generation cost. Pumped storage hydroelectric (PSH) power generation not merely meet basic requirement, it can also be used as power generation or load. These advantages were usually applied for peak modulation, frequency regulation, load standby, and accident standby. Hence, this study proposes to solve the voltage problem through the reactive power management of PSH units.

In recent years, some studies successfully focused on improving the voltage control ability of the power system by compensating RE generation through PSH dispatching. In previous studies, heuristic algorithms, including fuzzy evolutionary programming [13], improved differential evolution [14], multi-objective particle swarm optimization [15], and improved cuckoo search algorithm [16], were applied for the dispatching schedule of PSH. In [17], the nuclear density method was adopted to estimate the prediction error distribution of wind and solar energies, and the linear programming (LP) method was utilized to resolve the PSH peak regulation model to minimize the peak-valley difference of residual loads. In [18], a novel reference-point-based non-dominated sorting GA (NSGA-III) algorithm using a constraint violation criterion was introduced to solve the multi-objective scheduling problem under multi-complex constraints. In [19], the optimal operation scheduling of a hybrid hydro-thermal-wind system, peaking, and reduction of carbon dioxide emissions were achieved. In [20], a remote integrated transmission mode between a wind farm (WF) and a pumped storage power station was proposed, and piecewise linear approximation technology was used to transform the model into mixed integer linear programming (MILP) to mitigate the negative impact of WP fluctuations and increase profits. In [21], an ant-lion optimization algorithm was utilized to solve the combined hydro-thermal dispatching problem with multiple reservoirs and multiple targets (such as cost). Various emissions and losses were optimized simultaneously. In [22], the load and WP were predicted by a scenario-based stochastic method, and hydroelectric power scheduling was optimized to minimize the total generation cost and combustion emission in the short term. In [23], in order to reduce the operation cost of the integrated hybrid system, a new energy management method was used to find the operation strategy of the hybrid system. In [24], an adaptive differential evolution algorithm was adopted to solve the optimal power flow. In [25], a modified bacterial foraging algorithm (MBFA) was utilized to obtain the optimal power generation scheme of the hydro-thermal-wind system to maintain the voltage stability of the system. In [26], a multi-time-scale coordination method, that is, stochastic programming (SP), was adopted to control reactive power equipment to eliminate voltage fluctuations and deviations. Other algorithms, including simulated annealing (SA) [27], LP [28], genetic algorithm (GA) [29] are also commonly used for early scheduling optimization. The above algorithms achieved some results, but there are also evident deficiencies. For example, PSO can complete a task in a fixed scene but cannot perform well in a variable environment. The SA algorithm has strong robustness and is usually employed to address non-linear optimization problems, but its convergence is slow and can easily fall into local optimization. GA has a good global searching ability and does not fall into local optimization, but it cannot solve large-scale high-dimensional computing problems. It is a high-dimensional and non-linear optimization process to solve the voltage problem through the reactive power management of PSH units. In conclusion, SA and GA algorithms are not suitable for dynamic scheduling strategy in uncertain environment. Therefore, it is urgent to find a real-time intelligent scheduling method that can adapt to the characteristics of renewable energy power generation. Inspired by behaviourist psychology, the combination of artificial intelligence (AI) and data-driven technology is profoundly affecting and changing the global power and energy industry, and playing a great potential in the smart grid. Artificial intelligence focuses on the cumulative return of individuals in the process of interacting with dynamic random environment. In power system, reward can be expressed as the operation index of the system, such as minimum voltage fluctuation and minimum frequency fluctuation. With the emergence of AlphaGo [30], deep reinforcement learning, as a representative of artificial intelligence, has made gratifying progress in improving training speed. At present, artificial intelligence algorithm is widely used in high-dimensional non-linear optimization problems in smart grid, and has made gratifying progress. Therefore, this study proposes to apply artificial intelligence algorithm to voltage control problem.

Many studies have applied AI to smart dispatching. Deep learning (DL) [31] can extract high-order data features in high-dimensional continuous spaces. Reinforcement learning (RL) [32, 33] is adopted to settle decision problems and repeatedly detecting targets in dynamic uncertain environments. DRL combines the advantages of DL and RL. After successive evolution and renewal, from Q-learning [34] value-based updates, to Deep Q Network (DQN) [35] off policy-based updates, to deep deterministic policy gradient (DDPG) valued-based and policy-based updates. DRL algorithm [36, 37] has brought new ideas for complex control tasks owing to their strong adaptability to dynamic and uncertain environments. Q-learning is a decision algorithm in reinforcement learning. The Q-learning output action is discrete. When there are multiple states, Q-learning lists the Q table in the form of table, so the search and storage need a lot of time and space, which cannot solve high-dimensional continuous state action space in uncertain environment. Although DQN solves the problem of high-dimensional observation space, it can only deal with discrete action space. Deep reinforcement learning uses the powerful representation ability of neural network to fit the Q table or direct fitting strategy to solve the continuous state action space problem when solve the decision-making problem.

It is a non-linear and complex problem to balance the uncertain output of renewable energy and realize voltage stability by dispatching pumped storage hydropower. The dynamic scheduling of pumped storage under the conditions of renewable energy output and load fluctuation is an optimal control problem under uncertain scenarios. Due to its strong decision-making ability, deep reinforcement learning is very suitable to solve this problem. In the solution process of deep reinforcement learning, deep neural network is used to fit some functions. As a representative algorithm of deep reinforcement learning, deep deterministic strategy gradient is widely used in decision-making of non-linear complex problems with its advantage of continuous action state space. Some research has already done. In [38], an optimal voltage control scheme based on security DRL was proposed. In [39], based on a networked microgrid system, an energy storage system was dynamically scheduled online by deep deterministic policy gradient (DDPG) to support the standby dispatching scheme in case of an emergency when the dispatching centre was unavailable or the day-ahead plan was infeasible.

Inspired by the above studies, DDPG was applied in this study to solve the problems of bus voltage over limit. Specifically, in the environment of WP, PV variable in real time, the DDPG algorithm is employed to establish a data-driven model in a high-dimensional state space, and the optimal decision is obtained through repeated exploration and training of an agent. Through the experience of interacting with the environment, agent can integrate some aspects of the environment into its internal state, form its own understanding of specific behaviour applications, and make good decisions in response to environmental changes. After training, the agent learns the optimal control policy to control the active and reactive power outputs of the PSH to ensure high-quality operation of the power system.

The contributions of this study are as follows:

1. The voltage management problem of a pumped storage hydro-wind-solar (PSHWS) system is transformed into a Markov decision process (MDP), and the DDPG algorithm is adopted to train an agent to solve the reactive power scheduling of PSH units in an uncertain environment. After the off-policy training process, an optimal dynamic control strategy is obtained.
2. The proposed method is a data-driven method, and historical data are used to train the agent, so that this method not merely less affected by model errors, higher correctness and fast training speed when it solves high-dimensional state space problems.
3. SP and DQN algorithms are introduced as comparison examples, and the results of the three algorithms are analysed to prove the effectiveness of the proposed method.

The remainder of this paper is organized as follows: Section 2 describes the RE power model. Section 3 introduces the problem formulation. Section 4 introduces the optimization algorithm. Section 5 describes a case study. Finally, Section 6 concludes this paper.

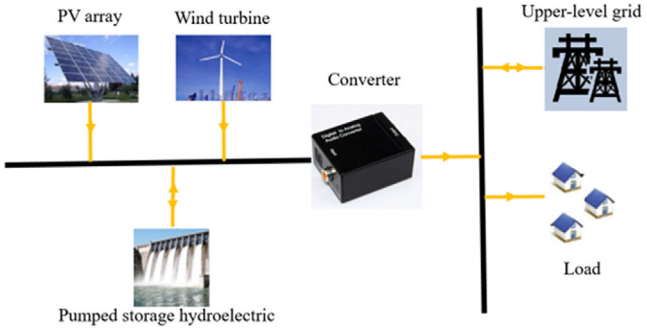

FIGURE 1 Schematic diagram of hybrid system

2 | RE GENERATION MODEL AND LOAD MODEL

This work applies DRL to PSHWS hybrid system, see in Figure 1. Which contains different electrical device. Such as WTs, PVs, and PSHs. With the penetration of renewable energy into the power system, the output of wind power generation and photovoltaic power generation is easily affected by environmental factors such as wind speed, illumination and temperature, so it has strong intermittence and uncertainty. The volatility of the distributed generator output and load are considered. The DRL algorithm is utilized to solve voltage problem and obtain a dynamic scheduling PSH unit policy. Each device of PSHWS is modelled separately according to its characteristics, which is described in detail in the following sections.

2.1 | Wind turbine model

When the wind speed is between the cut-in wind and the rated wind speeds, the output of the wind turbine (WT) is affected by the randomness of the wind speed, air density, wind energy utilization coefficient, and swept area [40], as expressed in the following equation:

$$P_W = 0.5\rho C_P \pi R^2 v^3 \tag{1}$$

where ρ is the air density, C_P is the wind energy utilization coefficient, and R is the turbine rotor.

According to the WT factory regulations, the output power of the WT is related to $v_{c,i}$, v_r and $v_{c,o}$ as expressed in the following equation:

$$P_W = \begin{cases} 0 & 0 \le v < v_{c,i} \\ a + bv^3 & v_{c,i} \le v < v_r \\ p_r & v_r \le v < v_{c,o} \\ 0 & v_{c,o} \le v \end{cases} \quad a = \frac{P_r v_{c,i}^3}{v_{c,i}^3 - v_r^3} \quad b = \frac{P_r}{v_r^3 - v_{c,i}^3} \tag{2}$$

where P_r is the rated power of the WT, and $v_{c,i}$, v_r, $v_{c,o}$ are the cut-in, rated, and cut-out wind speeds, respectively. If the WT is controlled by a constant power factor, the reactive power output

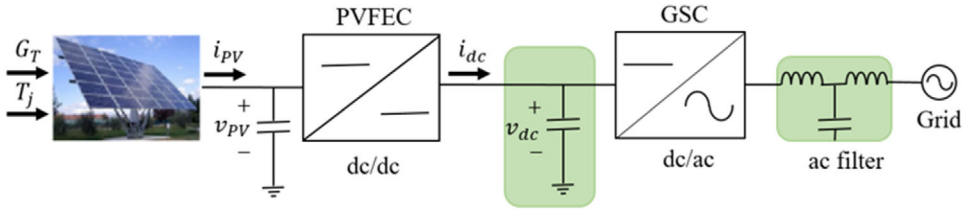

FIGURE 2 PV converter in PSHWS hybrid system

is expressed as follows:

$$Q_W = P_W \tan \lambda_w \tag{3}$$

where Q_w is the reactive power of the WT, and λ_w is the power factor angle of the WT.

2.2 | PV model

PV generation is affected by light intensity and angle. Light intensity has strong correlations with climate, season, and region. PV operates at the maximum power point tracking (MPPT) [41], as expressed in the following equations:

$$P_{PV} = P_{PV,STC}\frac{G_T}{1000}[1-\gamma(T_j-25)] \tag{4}$$

$$T_j = T_a + \frac{G_T}{800}(NOCT-20) \tag{5}$$

where P_{PV} is the output power at MPPT, $P_{PV,STC}$ is the rated power at MPPT under standard test conditions, G_T is the irradiance level under standard test conditions, γ is the temperature coefficient, T_j is the PV unit temperature, T_a is the air temperature, and $NOCT$ is the nominal operating cell temperature.

PV is a DC power supply. IEEE-30 is an AC system, so inverter (see in Figure 2) is required between photovoltaic and power grid [42]. This paper introduces a two-stage grid connected photovoltaic power interface, which includes PV array, photovoltaic front-end converter (PVFEC) for dc/dc conversion, intermediate DC link (IDCL), AC filter and grid side converter (GSC) for dc/ac conversion.

2.3 | Pumped storage hydro model

Pumped storage hydroelectric (PSH) is related to hydroturbine conversion efficiency, water flow rate, and water height. The output of the PSH unit [43] is expressed as follows:

$$P_h = K\eta_j Q_j h_j \tag{6}$$

where P_h is the output of the PSH, K is the hydroturbine conversion efficiency, η_j is the efficiency of the PSH station, Q_j is the water flow rate passing through the turbine j, and h_j is the net water height of the power station.

2.4 | Load model

The ZIP (constant impedance (Z), constant current (I), and constant power (P)) load model describes the relationship between the load and voltage amplitude [44] as follows:

$$P_L = P_0\left(Z_P\left(\frac{V}{V_N}\right)^2 + I_P\left(\frac{V}{V_N}\right) + P_P\right) \tag{7}$$

$$Q_L = Q_0\left(Z_Q\left(\frac{V}{V_N}\right)^2 + I_Q\left(\frac{V}{V_N}\right) + P_Q\right) \tag{8}$$

where P_L and Q_L are the active and reactive powers of the load when the voltage deviates from the rating, respectively; Z_P, I_P, and P_P are the constant impedance, constant current, and constant power coefficient of active power, respectively; P_0 and Q_0 are the active and reactive loads under nominal voltage, respectively; V is the voltage amplitude; and V_N is the nominal voltage.

3 | PROBLEM FORMULATION OF VOLTAGE CONTROL FOR PSHWS SYSTEM

The objective of this study was to minimize the voltage deviation of a PSHWS system when implementing the generation plan of PSH units. This generation plan ensures that the load demand can be met under safe and stable operating conditions of the system. On this basis, a cooperative operation model that considers the uncertainty of WP and PV power generation is established. Let us assume a system with a WP unit, PV power unit, a PSH unit, and B nodes.

$$\begin{bmatrix}\Delta P_1\\ \Delta Q_1\\ \vdots\\ \Delta P_B\\ \Delta Q_B\end{bmatrix} = \begin{bmatrix}\frac{\partial P_1}{\partial \theta_1} & \frac{\partial P_1}{|\partial V_1|} & \cdots & \frac{\partial P_1}{\partial \theta_B} & \frac{\partial P_1}{|\partial V_B|}\\ \frac{\partial Q_1}{\partial \theta_1} & \frac{\partial Q_1}{|\partial V_1|} & \cdots & \frac{\partial Q_1}{\partial \theta_B} & \frac{\partial Q_1}{|\partial V_B|}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \frac{\partial P_B}{\partial \theta_1} & \frac{\partial P_B}{|\partial V_1|} & \cdots & \cdots & \cdots\\ \frac{\partial P_B}{\partial \theta_1} & \frac{\partial Q_B}{|\partial V_1|} & \cdots & \cdots & \cdots\end{bmatrix}\begin{bmatrix}\Delta\theta_1\\ \Delta V_1\\ \vdots\\ \Delta\theta_B\\ \Delta V_B\end{bmatrix} \tag{9}$$

$$|V| = [V_2 \ldots V_B]^T,\ \theta = [\theta_2 \ldots \theta_B]^T \tag{10}$$

$$\text{Minimize } ||V_{B_i} - V_{ref}||_2 \; i = 1, 2, \ldots, B \tag{11}$$

where B is the number of bus nodes. ΔP_i and ΔQ_i represent the active and reactive power imbalances of the i-th node, respectively, and ΔV_i, $\Delta\theta_i$ represent the voltage and angle imbalances of the i-th node, respectively, V_{B_i} is the voltage of bus i, V_{ref} is the voltage reference value.

The equality constraints include the active and reactive power constraints of the power system. These two constraints ensure the stability of the system's frequency and voltage. Inequality constraints include the upper and lower limits of the active and reactive power, voltage, and transformer.

Active power balance means that the output power of all generators is consumed by all loads except the grid loss and plant power load. Reactive power balance means that the reactive power supplied by the generators and compensation equipment afford the reactive power consumed by the load and lost in the switch. The reactive power to high or low can lead to a voltage over limit, and the power balance relations are as follows:

$$P_{G_i} - P_{D_i} - V_i \sum_{j=1}^{N} V_j [G_{ij} cos(\delta_i - \delta_j) + B_{ij} sin(\delta_i - \delta_j)] = 0$$
$$i = 1, 2, \ldots, B \tag{12}$$

$$Q_{G_i} - Q_{D_i} - V_i \sum_{j=1}^{N} V_j [G_{ij} cos(\delta_i - \delta_j) + B_{ij} sin(\delta_i - \delta_j)] = 0$$
$$i = 1, 2, \ldots, B \tag{13}$$

where P_{G_i} and Q_{G_i} are the active power and reactive power of the i-th bus, respectively; P_{D_i} and Q_{D_i} are the active and reactive loads of the i-th bus, respectively; V_i and V_j are the voltages of the i-th and j-th buses, respectively; G_{ij} and B_{ij} are the conductance and susceptance between bus i and j, respectively; and δ_i is the voltage angle of the i-th bus.

The inequality constraints are shown in Equations (14)–(19) below, including the upper and lower limits of the PSH output, the output power, the voltage, and the transformer, which are the preconditions for safe operation of the PSHWS system.

- PSH constraints:

$$P_{b_i}^{\min} \le P_{b_i} \le P_{b_i}^{\max} \; i = 1, 2, \ldots, H \tag{14}$$

where H is the quantity of the pumped storage hydro generator, and P_{b_i} is limited to values between the lower limit ($P_{b_i}^{\min}$) and upper limit ($P_{b_i}^{\max}$).

- Power constraints:

$$P_{G_i}^{\min} \le P_{G_i} \le P_{G_i}^{\max} \; i = 1, 2, \ldots, G \tag{15}$$

$$Q_{G_i}^{\min} \le Q_{G_i} \le Q_{G_i}^{\max} \; i = 1, 2, \ldots, G \tag{16}$$

where G is the quantity of the generator, and P_{G_i} and Q_{G_i} are limited to values between lower limits ($P_{G_i}^{\min}$, $Q_{G_i}^{\min}$) and upper limits ($P_{G_i}^{\max}$, $Q_{G_i}^{\max}$).

- Voltage constraints:

$$V_{G_i}^{\min} \le V_{G_i} \le V_{G_i}^{\max} \; i = 1, 2, \ldots, G \tag{17}$$

$$V_{B_i}^{\min} \le V_{B_i} \le V_{B_i}^{\max} \; i = 1, 2, \ldots, B \tag{18}$$

where V_G and V_B are limited to values between the lower limits ($V_{G_i}^{\min}$, $V_{B_i}^{\min}$) and upper limits ($V_{G_i}^{\max}$, $V_{B_i}^{\max}$).

- Transformer constraint:

$$T_{k_i}^{\min} \le T_{k_i} \le T_{k_i}^{\max} \; i = 1, 2, \ldots, T \tag{19}$$

where T is the number of transformers, and T_{k_i} is limited to values between the lower limit ($T_{k_i}^{\min}$) and upper limit ($T_{k_i}^{\max}$).

4 | ALGORITHM

In fact, the system voltage control problem is modelled as a discrete horizon MDP, it is an optimal decision-making problem for the reactive power management of PSH units under stochastic environment. Then, the DDPG algorithm is employed to train an agent and solve the MDP.

4.1 | Formulation of the voltage control problem as a Markov decision process

The above PSH dispatching process to solve the voltage management problem can be established as a discrete-time MDP with a continuous state and an action space. MDP is an effective approach to establish a DRL model, indicating that the next state is only related to the current state and current actions (see in Figure 5). The environment is described by four tuples: $<S, A, P, R>$. Among them, S is an all-state set, A is a set with all actions, transfer function P is defined as $S \times A \times S \rightarrow [0,1]$, and R is the reward function, defined as $S \times A \times S \rightarrow R$.

- S is the state set; the state ($s_t \in S$) is defined as $(P_{W-1}(t), \ldots, P_{W-N}(t); P_{s-1}(t), \ldots, P_{s-M}(t); \; P_{Load-1}(t), \ldots, P_{Load-L}(t))$; N, M, L represent the number of WT, PV, and load, respectively; $P_{W-i}(t)$, $P_{s-i}(t)$, $P_{Load-i}(t)$ represent the power of WT, PV, and load, respectively.
- A is the action set of control variables; the action ($a_t \in A$) is defined as $(Q_{H-1}(t), \ldots, Q_{H-K}(t))$; the neural network selects action a_t by policy π; K represents the number of hydroturbine; $Q_{H-i}(t)$ represent the active power and reactive power of hydroturbine, respectively.

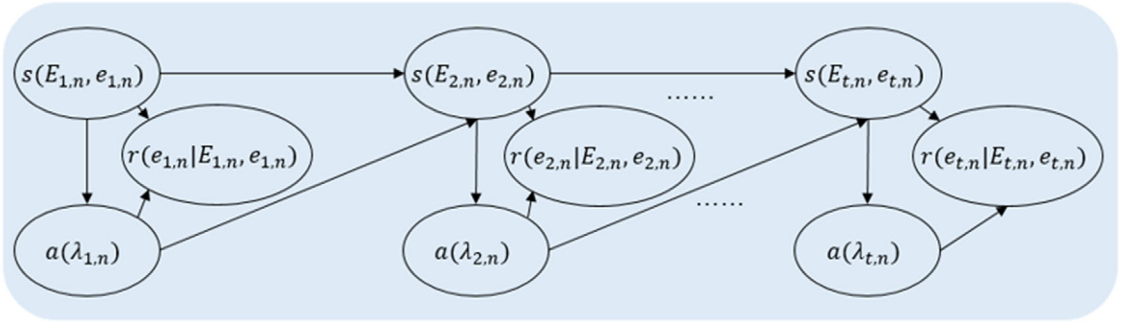

FIGURE 3 MDP processing of PSHWS system

- P is the state transition probability, $(s_{t+1} \approx P(s_t, a_t))$ represent the state s_{t+1} that is obtained by the state s_t that executes actiona_t.
- R is the reward set; the reward $(r_t \approx R(s_t, a_t))$ is returned by the external environment after the state s_t executes action a_t.

The key components contained in this MDP, discrete time t, action a $(\lambda_{t,n})$, state s $(E_{t,n}, e_{t,n})$, and reward r $(r(e_{t,n}|E_{t,n}, \lambda_{t,n}))$, see in Figure 3.

- $\lambda_{t,n}$ is a reactive power of PSH units that P choose at time t.
- $E_{t,n}$ represents the last state before it receives the reactive power of PSH units signal from P, and $e_{t,n}$ represents the voltage deviation after it receives the reactive power of PSH units signal from P.

One episode of the MDP is formed as: 1, $E_{1,n}$, $\lambda_{1,n}$, $e_{1,n}$, $r(e_{1,n}|E_{1,n}, \lambda_{1,n})$; 2, $E_{2,n}$, $\lambda_{2,n}$, $e_{2,n}$, $r(e_{2,n}|E_{2,n}, \lambda_{2,n})$; t, $E_{t,n}$, $\lambda_{t,n}$, $e_{t,n}$, $r(e_{t,n}|E_{t,n}, \lambda_{t,n})$; T, $E_{T,n}$, $\lambda_{T,n}$, $e_{T,n}$, $r(e_{T,n}|E_{T,n}, \lambda_{T,n})$.

According to the transition process of the uncertain environment, the transition to the next state s_{t+1} is not only related to the previous state s_t, but also to previous states s_{t-1}, s_{t-2} etc. The resulting transition model is complex and difficult to model. Therefore, the model is simplified by assuming the Markov property of state transition, such that the probability of transition to the next state s_{t+1} is only related to the previous state s_t, as follows:

$$P_{s_t s_{t+1}} = E(S_{t+1} = s_{t+1} | S_t = s_t, A_t = a_t) \tag{20}$$

The value-based function also depends only on the current state, so the cumulative reward value [45] from time step t is calculated as follows:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} = \sum_{k=0}^{S} r_{t+1+k} \tag{21}$$

where γ is the discounting rate ($\gamma \sim [0, 1]$), which reflects the impact of the reward returned in the future on the current action.

The above value-based function does not consider the impact of the action. To this end, the action-value function following policy π, which maps the state-action pair to the reward, is introduced. It can also be described by the Bellman expectation equation [45]:

$$Q^{\pi}(s_t, a_t) = \mathbb{E}^{\pi}[R_{t+1} + \gamma E_{a_{t+1}}[Q^{\pi}(s_{t+1}, a_{t+1})]] \tag{22}$$

The Bellman equation indicates that the value of a state is composed of the reward of that state and the subsequent value according to a certain attenuation ratio.

4.2 | Adoption of DDPG to solve MDP

The DDPG adopts an actor-critic architecture, and two neural networks are used to approximate them. The actor network is responsible for selecting the current action a according to the current state s, and interacting with the environment to generate the next state s' and reward value R. The critic network is responsible for calculating the current $Q(s_t, a_t|\theta^w)$ value and target Q value y_t.

Concerning the critic network, the objective is to minimize the difference between the Q value calculated in the current state and the target Q value, which can be updated by the loss function as follows [45]:

$$y_t = r(s_t, a_t) + \gamma Q\left(s_{t+1}, u\left(s_{t+1}\right)|\theta^{\omega}\right) \tag{23}$$

$$L(\theta^{\omega}) = \mathbb{E}_{\omega'}\left[\left(Q(s_t, a_t|\theta^{\omega}) - y_t\right)^2\right] \tag{24}$$

$$\omega_{t+1} = \omega_t + n_c \nabla_{\mu} L(\theta^{\omega}) \tag{25}$$

where n_c is the learning rate of the critic network. In fact, the meaning of critic network training is to minimize the difference between y_t and $Q(s_t, a_t|\theta^w)$.

Regarding the actor network, the function J [45] is utilized to output a deterministic value through a deterministic strategy

gradient, as expressed in the following equations:

$$\nabla_{\theta^\mu} J = \mathbb{E}_{s_t \approx \rho^\mu}\left[\nabla_{\theta^\mu} Q(s, a|\theta^\omega)\Big|_{s=s_i, a=\mu(s_i)}\right] = \mathbb{E}_{s_t \approx \rho^\mu}\left[\nabla_a Q(s, a|\theta^\omega)\Big|_{s=s_i, a=\mu(s_i)} \cdot \nabla_{\theta^\mu}\mu(s|\theta^\omega)\Big|_{s=s_i}\right] \tag{26}$$

$$\mu_{t+1} = \mu_t + n_a \nabla_{\theta^w} J \tag{27}$$

where n_a is the learning rate of the actor network. The purpose of actor network training is to maximize $Q(s, a)$.

To make the training process more stable and reliable, DDPG additionally uses two target actor-critic networks (the number of neurons in the target and actor-critic networks is the same). The parameters of the target network are updated by a soft update process, as defined in the following expression:

$$\begin{cases} \theta^{\omega'} \leftarrow \tau\theta^\omega + (1-\tau)\theta^{\omega'} \\ \theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'} \end{cases} \tag{28}$$

where the target actor-critic network is parameterized by μ' and ω', respectively, and τ is the update rate.

The introduced registers, in accordance with the experience replay, $e = (s_t, a_t, r_t, s_{t+1})$, for each step are deposited in an M-size experience replay buffer $D = (e_1, e_2, \ldots, e_M)$. The experience replay is similar to a brain to store memory in registers. When the capacity of D is exceeded, the new experience overwrites the old one. In each step, a mini-batch of experiences is sampled to calculate the gradients. Subsequently, the networks are updated by the gradients. Suppose the extraction of a mini-batch of experience $\{e_1, e_2, \ldots, e_N\}$, $e = (s_t, a_t, r_t, s_{t+1})$, $i = 1,2, \cdots, N$ to calculate the minimum loss as follows:

$$L(\omega) = \frac{1}{N}\sum_{i=1}^{N}\left[y_i - Q(s_i, a_i|\theta^\omega)\right]^2 \tag{29}$$

$$\omega_{i+1} = \omega_i + n_c \nabla_\mu L(\theta^\omega) \tag{30}$$

where $y_i = r_i + \gamma Q^{w'}(s_{i+1}, \mu(s_t'))$ is the target action value obtained by the target critic network. The parameter is updated by the policy gradient, and the determined value is output by the deterministic strategy gradient, as expressed in the following equations:

$$\nabla_{\theta^\mu} J = \frac{1}{N}\sum_{i=1}^{N}\left[\nabla_{\theta^\mu} Q(s, a|\theta^\omega)\Big|_{s=s_i, a=\mu(s_i)}\right] = \frac{1}{N}\sum_{i=1}^{N}\left[\nabla_a Q(s, a|\theta^\omega)\Big|_{s=s_i, a=\mu(s_i)} \cdot \nabla_{\theta^\mu}\mu(s|\theta^\omega)\Big|_{s=s_i}\right] \tag{31}$$

ALGORITHM 1 DRL-based optimization method

Input: State of the generation system:
$P_{W-1}(t), P_{W-2}(t); P_{s-1}(t); P_{Load-1}(t) \ldots, P_{Load-21}(t)$

Output: State of the PSH unit: $Q_{H-1}(t), Q_{H-2}(t)$

1: Randomly initialize critic network w and actor network μ
2: Initialize target network w' and μ' with weight $w' \leftarrow w$, $\mu' \leftarrow \mu$
3: Initialize memory replay D
4: for episode = 1 to max episode, M do
5: Initialize a random process N for action exploration
6: Receive initial observation state s_1
7: for step = 1 to max step, T do
8: Select action $a_t = \mu(s_t) + N_t$ according to the current policy and exploration noise
9: Equations (9)–(13), execute power flow calculation of hybrid system
10: Equations (14)–(19), inequality constraints are applied to the power flow calculation results
11: Execute action a_t and obtain reward γ_t (calculated by Equation (21)) and new state s_{t+1}
12: Set s_{t+1}=s_t, a_t, γ_t
13: Store transition $(s_t, a_t, \gamma_t, s_{t+1})$ in D
14: If the D is full
15: Sample random mini batch of N transitions $(s_t, a_t, \gamma_t, s_{t+1})$ from D
16: Update critic network according to Equation (30)
17: Update the actor policy according to Equation (32)
18: Update the target networks according to Equation (28)
19: end for
20: end for

$$\mu_{i+1} = \mu_i + n_a \nabla_{\theta^w} J \tag{32}$$

Taking advantage of DQN, the parameters in the actor-critic network are stored back to the target actor-critic network through the gradient back propagation update. During training, the target network synchronizes the weight of the critic network at an update rate τ in each training iteration. Here, $Q^{w'}(s, a)$ is utilized to denote the target critic network, and w' is the weight; $Q^{\mu'}(s, a)$ is utilized to denote the target actor network, and μ' is the weight. The DDPG algorithm flow is shown in Algorithm 1. The flowchart of DDPG to train an agent see in Figure 4.

The overall scheme of DDPG is shown in Figure 5. Specifically, during the training process, the current state of the WT, PV, and load obtained from the environment is computed by the actor–critic network. Then, some state-action pairs are extracted from the memory bank to calculate the Q value, and actions are finally output with the reactive PSH to the environment through the policy gradient. The parameters of actor DNN, critic DNN, target actor DNN, target critic DNN are μ, w, μ' and w', respectively. t represents current episode; s represents the state, a represents the action; Q and Q' are the output value of critic DNN and target critic DNN, respectively; θ is current optimal

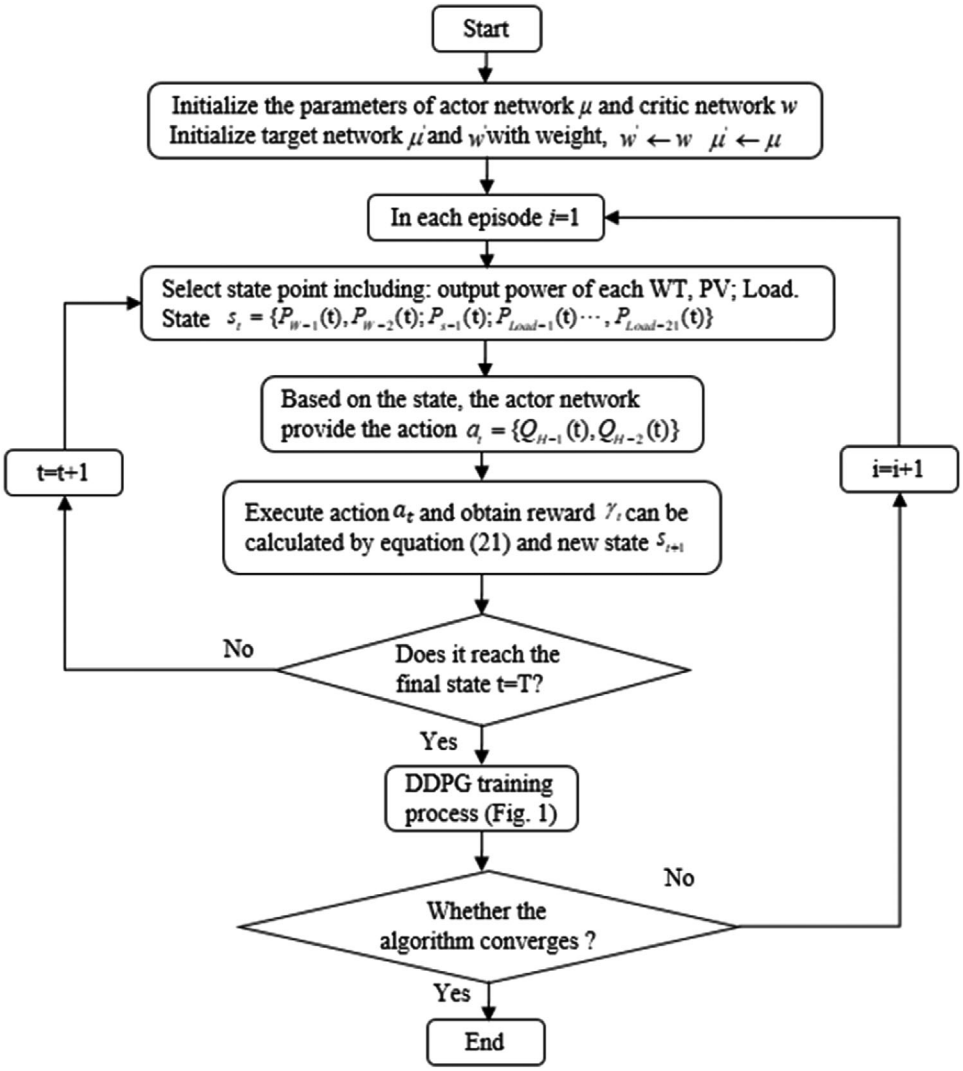

FIGURE 4 Flowchart of DDPG to train an agent

network parameters. The simulation is realized on MATLAB 2018b and Python 3.6 on a 64-bit Windows-based computer with 4 GB of RAM and Intel Core i5 processor clocking at 2.7 GHz.

5 | CASE STUDY

5.1 | IEEE 30-bus power system

For further investigation of the effectiveness of the proposed method, nodes 1 are slack-bus. The rest of the bus nodes, loads, and other parameters are described in Table 1.

Specifically, there are six power nodes in IEEE 30 system. How to connect the three energy sources of WT, PV and PSH is tested according to the configuration in Table 2.

After comparing the three renewable energy access modes. According to Figure 6, the configuration 3 is finally selected to access the IEEE30 system and simulated. the improved IEEE 30-bus-based combined RE was used as a test system (see Figure 7).

TABLE 1 IEEE 30-bus system

Unit	Quantity	Details
Bus	30	[14]
Slack	1	Bus 1(G1)
Load	21	[14]
AC line	41	[14]
Bus voltage	—	[0.95–1.05] p u

Information on RE connected to the PSHWS system is shown in Table 3. This includes two WFs, one including 30 WTs with a total capacity of 2.5 MW, and the other containing 25 WTs. The PV power plants have a capacity of 25 MW, and the two PSH units have a capacity of 160 MW.

A simulation was conducted on the PSHWS system. Table 4 shows the bus-voltage simulation results. The bold bus indicates

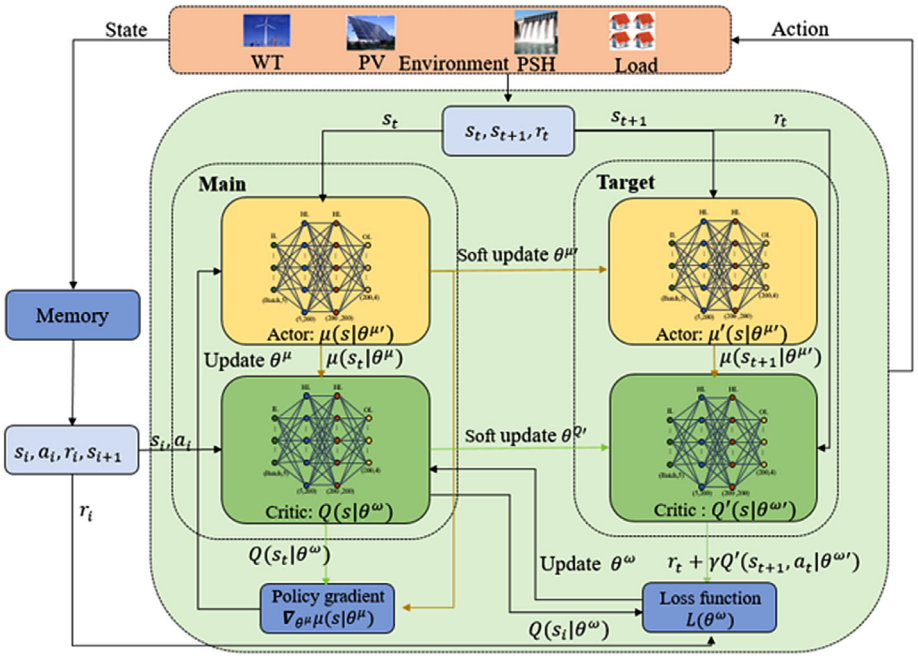

FIGURE 5 DDPG-based agent training diagram

TABLE 2 Renewable energy distribution

Unit	Configuration 1	Configuration 2	Configuration 3
WF	Bus 2(G2), Bus 8(G4)	Bus 11(G5), Bus 13(G6)	Bus 5(G3), Bus 11(G5)
PV	Bus 5(G3)	Bus 2(G2)	Bus 13(G6)
PSH	Bus 11(G5), Bus 13(G6)	Bus 5(G3), Bus 8(G4)	Bus 2(G2), Bus 8(G4)

that the absolute value of the bus-voltage deviation is greater than 5%, and the buses reach voltages over limit. To solve this problem, the DDPG introduced in Section 4.2 is applied to the voltage control problem.

The main hyperparameters that affect the final training results are discount factor (γ) and soft update coefficient (τ). The larger the discount factor, the more difficult it is for the agent to learn, so two test values of 0.9 and 0.95 are set for the discount factor. Then, the larger the soft update coefficient, the more unstable the agent is. Two test values are set for the soft update coefficient, which are 0.001 and 0.0001 respectively. The parameters of the DDPG algorithm are listed in Table 5. The different hyperparameters are selected to train three agent. The test results confirm that the bigger the discount factor γ, the harder

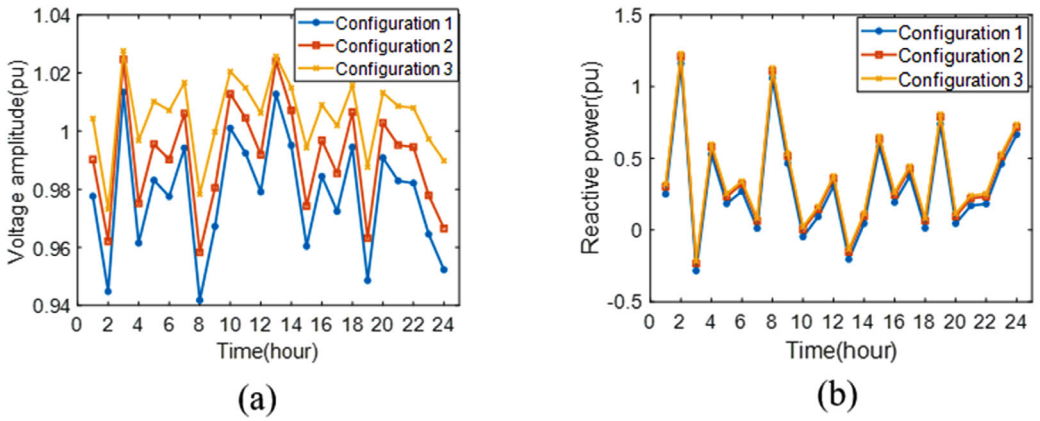

FIGURE 6 Three RE access configurations. (a) Voltage amplitude; (b) Reactive power of two generators

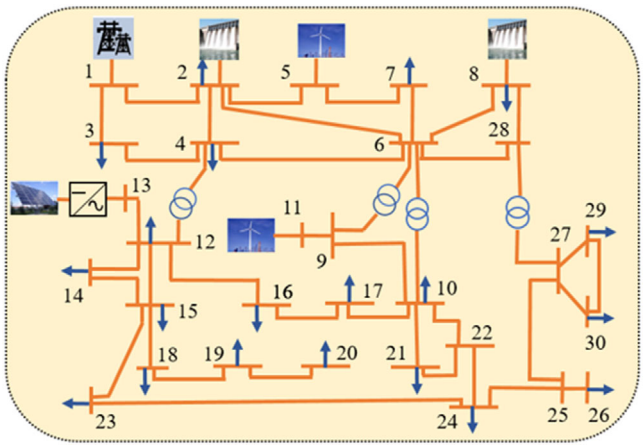

FIGURE 7 PSHWS system based on IEEE 30-bus

TABLE 3 Parameter setting for the generators

Unit	Value	Connected
WT	25*2.5 MW	Node 5
	30*2.5 MW	Node 11
PV	25 MW	Node 13
PSHP	160 MW, 74 MVAR	Node 2
	160 MW, 74 MVAR	Node 8

the agent is to learn (see in Figure 8, Agent 2), after 100 points, there is an obvious learning effect on Agent 1 and Agent 3, while agent 2 has about 500 points. Otherwise, the bigger the soft update coefficient τ, the more unstable the learning effect of the agent (see in Figure 8, Agent 1). Finally, the convergence of Agent 2 and Agent 3 are very stable, and the convergence result

TABLE 4 Optimized previous voltage amplitudes

Bus name	Voltage amplitude (p u)	Bus name	Voltage amplitude (p u)
1	1.00000	**16**	**0.93642**
2	0.97603	**17**	**0.92390**
3	0.96883	**18**	**0.91659**
4	0.96141	**19**	**0.91102**
5	**0.94055**	**20**	**0.91417**
6	0.95411	**21**	**0.91468**
7	**0.94857**	**22**	**0.91572**
8	0.95513	**23**	**0.91780**
9	**0.94010**	**24**	**0.90848**
10	**0.92733**	**25**	**0.92587**
11	0.98500	**26**	**0.90633**
12	0.95809	**27**	**0.95115**
13	0.99075	28	0.95891
14	**0.94011**	**29**	**0.94657**
15	**0.93208**	**30**	**0.93213**

TABLE 5 Parameters of the DDPG algorithm

DDPG	Agent 1	Agent 2	Agent 3
Active power action value	[0, 1.6]	[0, 1.6]	[0, 1.6]
Reactive power action value	[0, 0.74]	[0, 0.74]	[0, 0.74]
Experience replay memory capacity	8000	8000	8000
Step size of each episode	10	10	10
Mini-batch size	40	40	40
Learning rate for actor network (n_a)	0.001	0.001	0.001
Learning rate for critic network (n_c)	0.002	0.002	0.002
Discount factor (γ)	0.9	0.95	0.9
Soft update coefficient (τ)	0.001	0.0001	0.0001

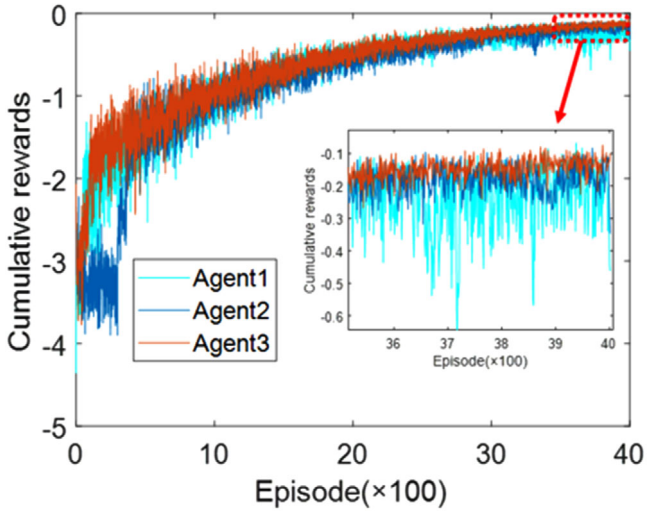

FIGURE 8 Cumulative reward for iterative process

of agent 1 fluctuates more than that of the other two agents. Therefore, after comparison, the well-trained agent 3 is finally selected for testing.

5.2 | The training process of deep deterministic policy gradient

In this study, the DDPG algorithm described in Section 4 was applied to train an agent to solve the voltage control problem. The training process is shown in Figure 8. The curve represents the trend of cumulative rewards in the training process. Note that, initially, the cumulative reward of the agent (see Equation (9)) is small. This is because the agent cannot perform good action at this initial stage. With the increase in training episodes, the agent interacts with the environment to learn experiences and obtain a larger cumulative reward. After approximately 2000 episodes, the cumulative reward converges to a satisfactory range, which means that the agent learns the optimal or near-optimal control policy, and it can carry out optimal actions.

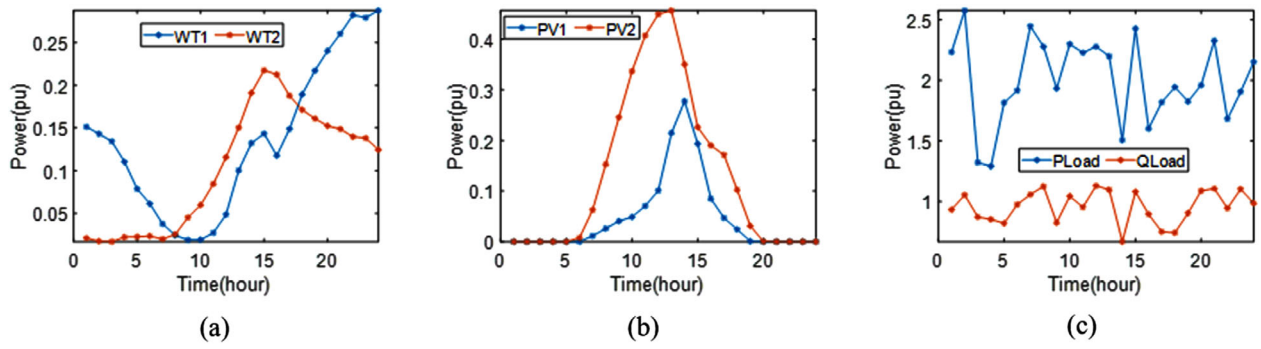

FIGURE 9 Data collected throughout 24 h for testing: (a) WF output power; (b) PV output power; (c) Total load

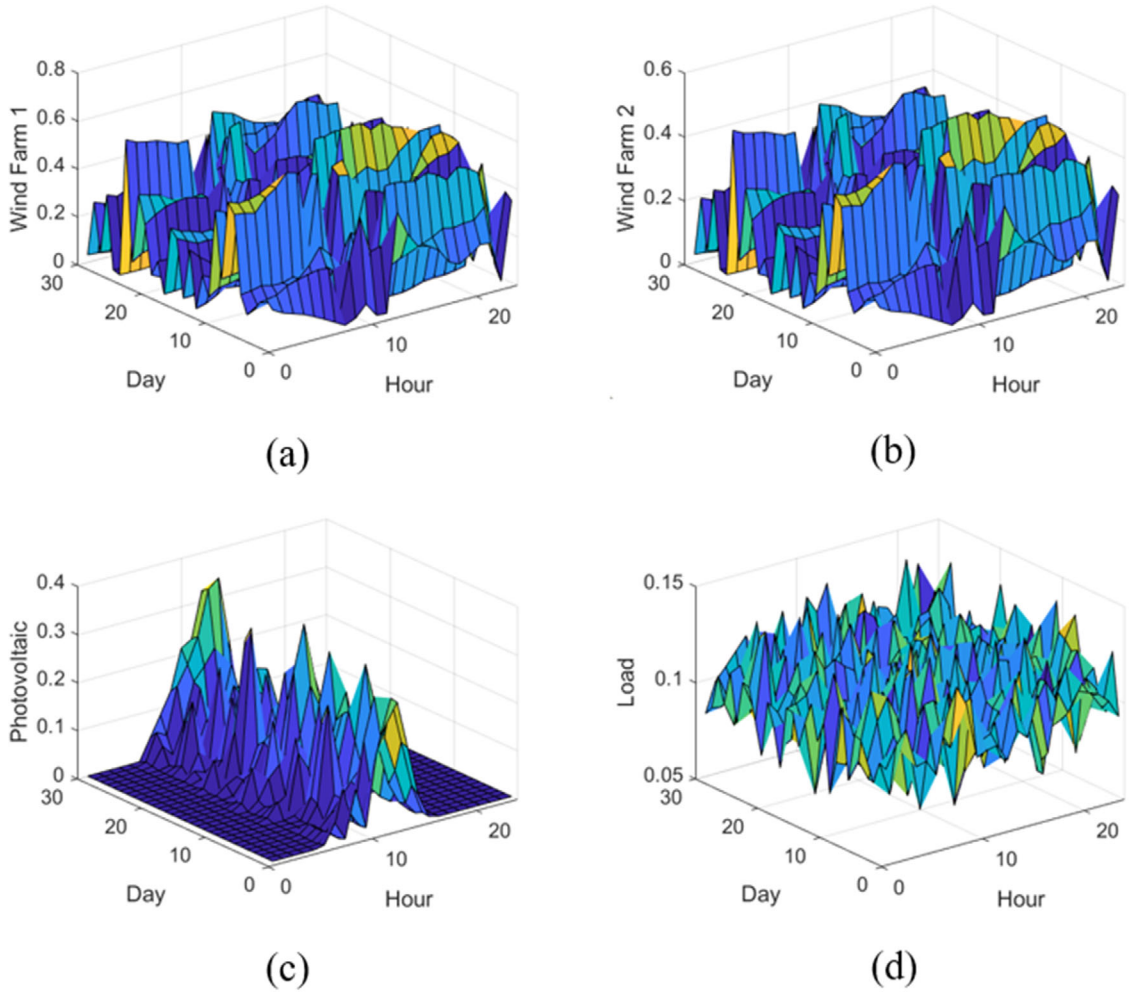

FIGURE 10 Data collected throughout 30 day for testing: (a) WF1 output power; (b) WF2 output power; (c) PV output power; (d) Total load

5.3 | Testing the well-trained agent

Then, the above well-trained agent would provide reactive power dynamic dispatch schemes for the PSH units to compensate for the voltage problem. To verify the effectiveness of the dispatch scheme provided by the agent, data collected throughout 24-h were selected as test data, as shown in Figure 9. And a month simulation data were also selected as test data (see in Figure 10). The output power of WFs and PV power plants adopts the actual data [46]; the load is collected by a Gaussian process [47].

First, to verify the effectiveness of the reactive power dynamic dispatch scheme provided by the agent, we compared it with a traditional reactive power control scheme [48]. The

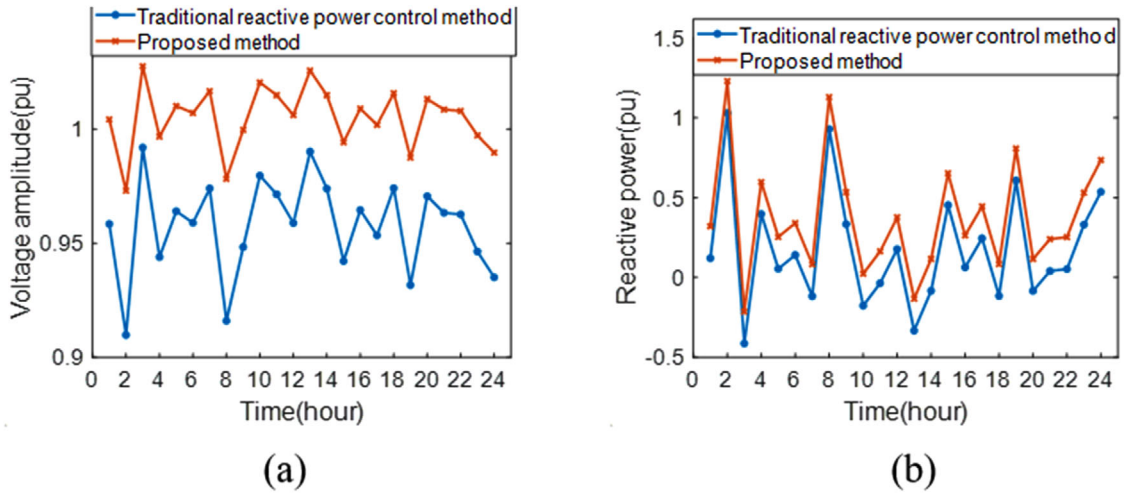

FIGURE 11 24-h simulation comparison: (a) Voltage amplitude; (b) Reactive power of two generators

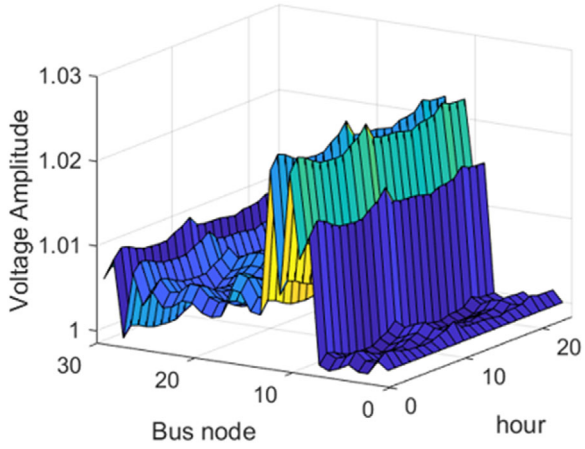

FIGURE 12 Voltage amplitude of all bus node

simulation results are shown in Figure 11a. To compare the voltage control effect of these two reactive power schemes, the variation in the Bus-6 voltage amplitude with different reactive power schemes is shown in Figure 11b. Note that the proposed reactive power control schemes can make the bus voltage closer to the rated voltage than the traditional reactive power control scheme. This means that the proposed method can show a better voltage control effect in comparison with the traditional method.

Then, the voltage amplitude of all bus nodes within 24-h which optimized by DDPG algorithm is seen in Figure 12. It shows that the voltage amplitude are limited in [0.97, 1.03].

5.4 | Testing the robustness of DDPG algorithm

Then, three well-trained agents by DDPG algorithm, and testing in the same wind, PV and load data, see in Figure 13. In the Bus-6, the performance of different agents are variant. But the voltage amplitude are also limited in [0.97, 1.03].

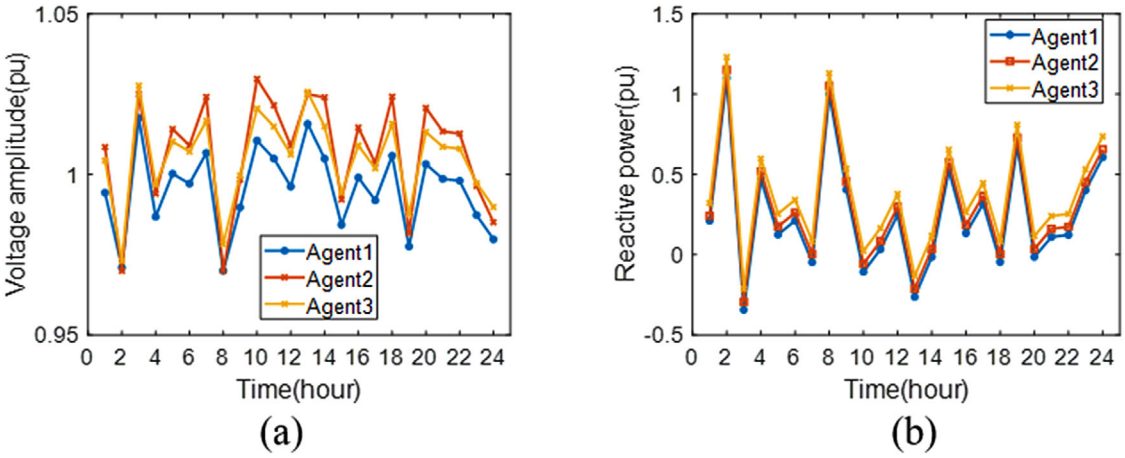

FIGURE 13 24-h simulation compare in three agents. (a) Voltage amplitude; (b) Reactive power of two generators

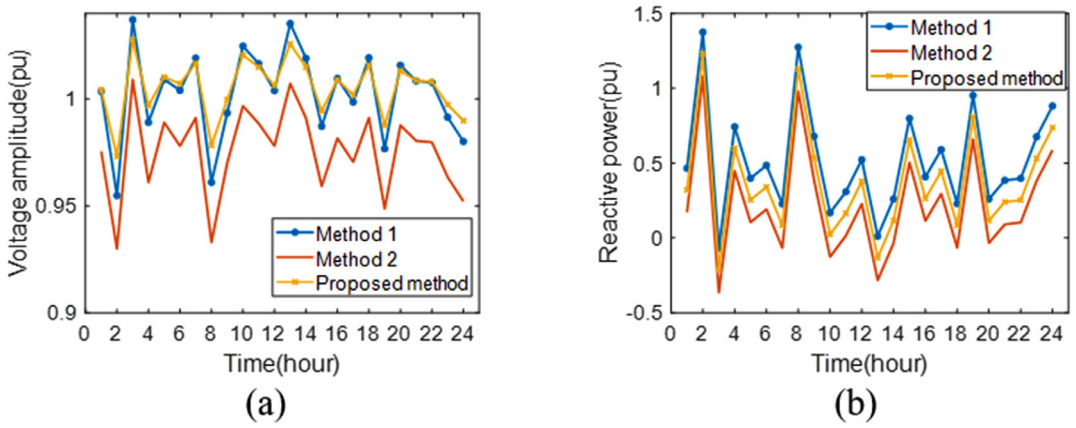

FIGURE 14 24-h simulation compare with previous methods. (a) Voltage amplitude; (b) Reactive power of two generators

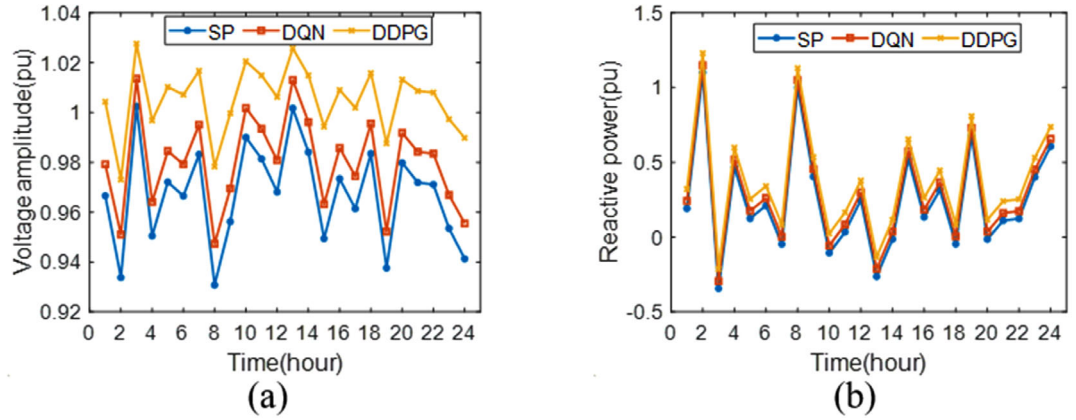

FIGURE 15 24-h comparison in SP, DQN, and DDPG: (a) Voltage amplitude; (b) Reactive power of two generators

5.5 | Comparison with two previous methods

Then, two previous methods [22, 23] are selected for comparison. According to Figure 14, the method 1 obtain a fine effect, and the method 2 obtain an unsatisfactory results. But the result of the proposed method is better than the two previous method. Otherwise, the proposed method uses one year of WT, PV and load data training, so the well-trained neural network can give the optimal decision immediately in the face of any test data. The previous method can only give one day's data in the training process, and the training results can only be used for the decision-making of the current training data. Each training will take up too much time and space.

5.6 | Comparison with SP and DQN

For further investigation of the advantages of the proposed method, both the SP and DQN methods were also employed as comparison examples. The detailed settings of these two methods are listed in Table 6 [49]. Figure 15 shows the comparisons

TABLE 6 Parameters of SP and DQN algorithms

Algorithm	Parameter	Value
SP	Active power action value	[0, 1.6]
	Reactive power action value	[0, 0.74]
	Iterations	2000
DQN	Active power action value	[0, 0.3, 0.6, 0.9, 1.2, 1.5]
	Reactive power action value	[0, 0.148, 0.296, 0.444, 0.592, 0.74]
	Discount factor (γ)	0.9
	Soft update coefficient (τ)	0.0001
	Experience replay memory capacity	8000
	Step size of each episode	10
	Learning rate for actor network (n_a)	0.001
	Learning rate for critic network (n_c)	0.002
	Mini-batch size	40

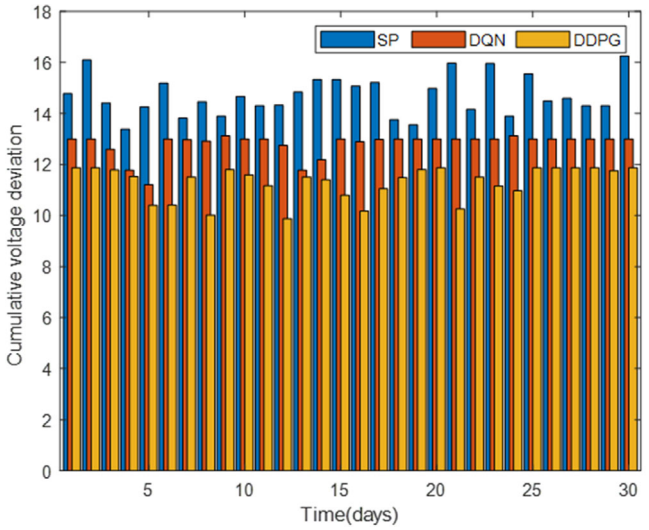

FIGURE 16 One-month simulation comparison in SP, DQN, and DDPG

TABLE 7 Comparison of the three methods for a month

Method	Total deviation (p u)	Improvement		Simulation time (h)
SP	428.9	0	0	3.55
DQN	378.58	−50.35	11.7%	3.03
DDPG	335.25	−93.65	21.8%	2.95

of the voltage amplitudes and reactive power in SP, DQN, and DDPG. Note that DDPG optimizes the voltage amplitude to [0.97, 1.03] p u, while SP is [0.93, 1.02] p u and DQN is [0.94, 1.02] p u. This means that the proposed method achieves a better voltage control effect than the other two methods.

Moreover, one month's data were also selected for further comparison. The change in total deviation (cumulative voltage deviation) with SP, DQN, and DDPG is shown in Figure 16. The proposed method clearly reduces the obtained total deviation for each day. For quantitative comparison of the different methods, the sum of the total deviation for one month was calculated; it is listed in Table 7. Note from Table 7 that the total deviation of DQN (11.7%) is less than that of SP, and the value of DDPG (21.8%) is less than that of SP. The above analysis shows that DDPG achieves a better performance than SP and DQN.

6 | CONCLUSION

In this study, the reactive power of pumped storage hydroelectric units was employed as a dynamic dispatch to compensate for voltage fluctuations. The problem of voltage fluctuation was solved by utilizing an AI algorithm, that is, deep deterministic policy gradient. After renewable energy is connected, the uncertain fluctuation of WT, PV and load makes the system voltage fluctuate between [0.9, 1.1] p u. Through testing on an IEEE 30-bus power system, deep deterministic policy gradient solves the problem of voltage deviation and controls the voltage of 30 nodes within a permitted range. At the same time, through the robustness experiment, three agents are training to optimize the target, it shows brilliant stability of deep deterministic policy gradient, and the optimization effect is satisfactory. In the same environment, stochastic programming and deep Q network were introduced to perform a comparative analysis. The results of deep Q network in voltage deviation control are [0.94, 1.02] p u, whereas the stochastic programming optimization results are [0.93, 1.02] p u. Moreover, the cumulative deviation of deep deterministic policy gradient per month is 21.8% less than that of stochastic programming. In conclusion, for complex problems with high dimensionality, the optimization effect of deep deterministic policy gradient is clearly better than that of deep Q network and stochastic programming, and the proposed solution to address voltage fluctuations was successfully demonstrated.

ACKNOWLEDGEMENT

This work was supported by the National Key Research and Development Program of China (2018YFE0127600).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ABBREVIATIONS

AI	Artificial intelligence
DARPC	Distributed optimal active and reactive power control
DDPG	Deep deterministic policy gradient
DFIG	Doubly fed induction generator
DL	Deep learning
DNN	Dense neural network
DQN	Deep Q network
DRL	Deep reinforcement learning
DVS	Dynamic voltage support
GA	Genetic algorithm
GSC	Grid side converter
IDCL	Intermediate DC link
LP	Linear programming
MBFA	Modified bacterial foraging algorithm
MDP	Markov decision process
MILP	Mixed integer linear programming
MPPT	Maximum power point tracking
NSGA	Non-dominated sorting genetic algorithm
PSH	Pumped storage hydroelectric
PSHWS	Pumped storage hydro-wind-solar
PSO	Particle swarm optimization
PV	Photovoltaic
PVFEC	Photovoltaic front-end converter
RESs	Renewable energy sources
RL	Reinforcement learning
SA	Simulated annealing
SP	Stochastic programming
TSPF	Time-series power flow
WFs	Wind farms

WP Wind power
WT Wind turbine

ORCID

Qin Huang https://orcid.org/0000-0003-1955-9173
Di Cao https://orcid.org/0000-0002-7019-7289
Zhe Chen https://orcid.org/0000-0002-2919-4481

REFERENCES

1. Agyekum, E.B.: Energy poverty in energy rich Ghana: A SWOT analytical approach for the development of Ghana's RE. Sustainable Energy Technol. Assess 40, 100760 (2020)
2. Zhang, S., Wei, J., Chen, X., Zhao, Y.: China in global wind power development: Role, status and impact. Renewable Sustainable Energy Rev. 127, 109881 (2020)
3. Global Offshore Wind Report 2020. Global Wind Energy Council. https://gwec.net/global-offshore-wind-report-2020/#report-overview (2020). Accessed 5 August, 2020
4. Renewable energy statistics 2021, International Renewable Energy Agency. https://iarena.org/solar (2021). Accessed August
5. Ren, H.-.P., Gao, Y., Huo, L., Song, J.-H., Grebogi, C.: "Frequency stability in modern power network from complex network viewpoint. Phys. A 545, 123558 (2020)
6. Yao, W., Jiang, L., Wen, J., Wu, Cheng, S.: Wide-area damping controller of FACTS devices for inter-area oscillations considering communication time delays. IEEE Trans. Power Syst. 29(1), 318–329 (2014)
7. Kawabe, K., Ota, Y., Yokoyama, A., et al.: Novel dynamic voltage support capability of photovoltaic systems for improvement of short-term voltage stability in power systems. IEEE Trans. Power Syst. 32(3), 1796–1804 (2016)
8. Molina-García, Á., Mastromauro, R.A., García-Sánchez, T., et al.: Reactive power flow control for PV inverters voltage support in LV distribution networks. IEEE Trans. Smart Grid 8(1), 447–456 (2016)
9. Vittal, E., O'Malley, M., Keane, A.: A steady-state voltage stability analysis of power systems with high penetrations of wind. IEEE Trans. Power Syst. 25(1), 433-442 (2010)
10. Kolhe, M.L., Rasul, M.: 3-Phase grid-connected building integrated photovoltaic system with reactive power control capability. Renewable Energy 154, 1065–1075 (2020)
11. Liao, W., Li, P., Wu, Q., et al.: Distributed optimal active and reactive power control for wind farms based on ADMM. Int. J. Electr. Power Energy Syst. 129(2), 106799 (2021)
12. Kawabe, K., Ota, Y., Yokoyama, A., Tanaka, K.: Novel dynamic voltage support capability of photovoltaic systems for improvement of short-term voltage stability in power systems. IEEE Trans. Power Syst. 32(3), 1796–1804 (2017)
13. Dong, H., Dong, Y., Zhou, C., Yin, G., Hou, W.: A fuzzy clustering algorithm based on evolutionary programming. Expert Syst. Appl. 36(9), 11792–11800 (2009)
14. Alsac, O., Stott, B.: Optimal load flow with steady-state security. IEEE Trans. Power Appar. Syst. PAS-93(3), 745–751 (1974)
15. Xu, X., Hu, W., Cao, D., Huang, Q., Chen, C., Chen, Z.: Optimized sizing of a standalone PV-wind-PSH station with pumped-storage installation hybrid energy system. Renewable Energy 147(P1), 1418–431 (2020)
16. Meng, X., Chang, J., Wang, X., Wang, Y.: Multi-objective PSH station operation using an improved cuckoo search algorithm. Energy (Oxford) 168, 425-439 (2019).
17. Liu, B., Lund, J.R., Liao, S., Jin, X., Liu, L., Cheng, C.: Optimal power peak shaving using PSH to complement wind and solar power uncertainty. Energy Convers. Manage. 209, 112628 (2020)
18. Yuan, X., Tian, H., Yuan, Y., Huang, Y., Ikram, R.M.: An extended NSGA-III for solution multi-objective hydro-thermal-wind scheduling considering wind power cost. Energy Convers. Manage. 96, 568–578 (2015)
19. Wang, Y., Zhao, M., Chang, J., Wang, X., Tian, Y.: Study on the combined operation of a hydro-thermal-wind hybrid power system based on hydro-wind power compensating principles. Energy Convers. Manage. 194, 94–111 (2019)
20. Su, C., Cheng, C., Wang, P., Shen, J., Wu, X.: Optimization model for long-distance integrated transmission of wind farms and pumped-storage PSH plants. Appl. Energy 242, 285–293 (2019)
21. Dubey, H.M., Pandit, M., Panigrahi, B.K.: Hydro-thermal-wind scheduling employing novel ant lion optimization technique with composite ranking index. Renewable Energy 99, 18–34 (2016)
22. Bahmani-Firouzi, B., Farjah, E., Azizipanah-Abarghooee, R.: An efficient scenario-based and fuzzy self-adaptive learning particle swarm optimization approach for dynamic economic emission dispatch considering load and wind power uncertainties. Energy 50(1), 232–244 (2013)
23. Sanaye, S., Sarrafi, A.: A novel energy management method based on Deep Q Network algorithm for low operating cost of an integrated hybrid system. Energy Rep. 7(3), 2647–2663 (2021)
24. Li, S., Gong, W., Wang, L., Yan, X., Hu, C.: Optimal power flow by means of improved adaptive differential evolution. Energy 198, 117314 (2020)
25. Panda, A., Tripathy, M., Barisal, A.K., Prakash, T.: A modified bacteria foraging based optimal power flow framework for hydro-thermal-wind generation system in the presence of STATCOM. Energy 124, 720–740 (2017)
26. Xu, Y., Dong, Z.Y., Zhang, R., Hill, D.J.: Multi-timescale coordinated voltage/var control of high renewable-penetrated distribution systems. IEEE Trans. Power Syst. 32(6), 4398–4408 (2017)
27. Wong, K.P., Fung, C.C.: Simulated annealing based economic-dispatch algorithm. Iee Proc.-C Gener. Transm. Distrib. 140(6), 509–515 (1993)
28. Moazeni, F., Khazaei, J.: Optimal operation of water-energy microgrids; a mixed integer linear programming formulation. J. Cleaner Prod. 275, 122776 (2020)
29. Baskar, S., Subbaraj, P., Rao, M.V.C.: Hybrid real coded genetic algorithm solution to economic dispatch problem. Comput. Electr. Eng. 29(3), 407–419 (2003)
30. Silver, D., et al.: Mastering the game of Go with deep neural networks and tree search. Nature 529.7587, 484–489 (2016)
31. Kather, J.N., Calderaro, J.: Development of AI-based pathology biomarkers in gastrointestinal and liver cancer. Nat. Rev. Gastroenterol. Hepatol. 17.10, 591–592 (2020)
32. Wang, Z., Hong, T.: Reinforcement learning for building controls: The opportunities and challenges. Appl. Energy 269, 115036 (2020)
33. Cao, D., Hu, W., Zhao, J., et al.: Reinforcement learning and its applications in modern power and energy systems: A review. J. Mod. Power Syst. Clean Energy 8(6), 1029–1042 (2020)
34. Watkins, C.J.C.H: Learning from Delayed Reward. Ph.D. Thesis. College University of Cambridge (1989)
35. Mnih, V., et al.: Playing atari with deep reinforcement learning arXiv preprint arXiv:1312.5602 (2013)
36. Cao, D., Hu, W., Zhao, J.B., Huang, Q., Chen, Z., Blaabjerg, F.: A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters. IEEE Trans. Power Syst. 35(5), 4120–4123 (2020)
37. Zhang, G., Hu, W., Cao, D., Huang, Q., Chen, Z., Blaabjerg, F.: A novel deep reinforcement learning enabled sparsity promoting adaptive control method to improve the stability of power systems with wind energy penetration. Renewable Energy 178, 363–376 (2021)
38. Kou, P., Liang, D., Wang, C., Wu, Z., Gao, L.: Safe deep reinforcement learning-based constrained optimal control scheme for active distribution networks. Appl. Energy 264, 114772 (2020)
39. Hao, R., Lu, T., Ai, Q., Wang, Z., Wang, X.: Distributed online learning and dynamic robust standby dispatch for networked microgrids. Appl. Energy 274, 115256 (2020)
40. Hou, P., Enevoldsen, P., Hu, W., Chen, C., Chen, Z.: Offshore wind farm repowering optimization. Appl. Energy 208, 834–844 (2017)
41. Gandhi, O., Kumar, D.S., Rodríguez-Gallegos, C.D., Srinivasan, D.: Review of power system impacts at high PV penetration part I: Factors limiting PV penetration. Solar Energy 210, 181–201 (2020)

42. Xiao, W., Torchyan, K., El Moursi, M.S., et al.: Online supervisory voltage control for grid interface of utility-level PV plants. IEEE Trans. Sustainable Energy 5(3), 843–853 (2014)
43. He, Z., Zhou, J., Qin, H., Jia, B., He, F., Liu, G., Feng, K.: A fast water level optimal control method based on two stage analysis for long term power generation scheduling of PSH station. Energy (Oxford) 210, 118531 (2020)
44. Asres, M.W., Girmay, A.A., Camarda, C., Tesfamariam, G.T.: Non-intrusive load composition estimation from aggregate ZIP load models using machine learning. Int. J. Electr. Power Energy Syst. 105, 191–200 (2019)
45. Li, A.J., et al.: Deep reinforcement learning for pedestrian collision avoidance and human-machine cooperative driving. Information Sciences 532, 110–124 (2020)
46. Zhang, G., Hu, W., Cao, D., et al.: Data-driven optimal energy management for a wind-solar-diesel-battery-reverse osmosis hybrid energy system using a deep reinforcement learning approach. Energy Convers. Manage. 227, 113608 (2021)
47. Shepero, M., Van Der Meer, D., Munkhammar, J., et al.: Residential probabilistic load forecasting: A method using Gaussian process designed for electric load data. Appl. Energy 218, 159–172 (2018)
48. Martinez-Rojas, M., et al.: Reactive power dispatch in wind farms using particle swarm optimization technique and feasible solutions search. Appl. Energy 88(12), 4678–4686 (2011)
49. Zhang, G., Hu, W., Cao, D., et al.: Deep reinforcement learning based optimization strategy for hydro-governor PID parameters to suppress ULFO. In: 2020 5th International Conference on Power and Renewable Energy (ICPRE). Shanghai, China (2020)

How to cite this article: Huang, Q., et al.: A novel deep reinforcement learning enabled agent for pumped storage hydro-wind-solar systems voltage control. IET Renew. Power Gener. 15, 3941–3956 (2021). https://doi.org/10.1049/rpg2.12311