TYPE Original Research
PUBLISHED 30 June 2023
DOI 10.3389/fenrg.2023.1202701

Coordinated control of multiple converters in model-free AC/DC distribution networks based on reinforcement learning

Qianyu Zhao[1,2], Zhaoyang Han[1,2], Shouxiang Wang[1,2]*, Yichao Dong[3] and Guangchao Qian[3]

[1]Key Laboratory of the Ministry of Education on Smart Power Grids, Tianjin University, Tianjin, China, [2]Tianjin Key Laboratory of Power System Simulation and Control, Tianjin University, Tianjin, China, [3]State Grid Tianjin Electric Power Company, Tianjin, China

Check for updates

OPEN ACCESS

EDITED BY
Xueqian Fu,
China Agricultural University, China

REVIEWED BY
Lin Gong,
Jackson State University, United States
Xiaoyu Duan,
Tsinghua University, China
Kaiqi Sun,
Shandong University, China

*CORRESPONDENCE
Shouxiang Wang,
✉ sxwang@tju.edu.cn

RECEIVED 09 April 2023
ACCEPTED 20 June 2023
PUBLISHED 30 June 2023

CITATION
Zhao Q, Han Z, Wang S, Dong Y and Qian G (2023), Coordinated control of multiple converters in model-free AC/DC distribution networks based on reinforcement learning.
Front. Energy Res. 11:1202701.
doi: 10.3389/fenrg.2023.1202701

COPYRIGHT
© 2023 Zhao, Han, Wang, Dong and Qian. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Taking into account the challenges of obtaining accurate physical parameters and uncertainties arising from the integration of a large number of sources and loads, this paper proposes a real-time voltage control method for AC/DC distribution networks. The method utilizes model-free generation and coordinated control of multiple converters, and employs a combination of agent modeling and multi-agent soft actor critic (MASAC) techniques for modeling and solving the problem. Firstly, a complex nonlinear mapping relationship between bus power and voltage is established by training an power-voltage model, to address the issue of obtaining physical parameters in AC/DC distribution networks. Next, a Markov decision process is established for the voltage control problem, with multiple intelligent agents distributed to control the active and reactive power at each converter, in response to the uncertainties of photovoltaic (PV) and load variations. Using the MASAC method, a centralized training strategy and decentralized execution policy are implemented to achieve distributed control of the converters, with each converter making optimal decisions based on its local observation state. Finally, the proposed method is verified by numerical simulations, demonstrating its sound effectiveness and generalization ability.

KEYWORDS

AC/DC distribution networks, voltage control, multi-agent soft actor critic, centralized training, markov decision process

1 Introduction

In recent years, there has been a growing focus on promoting energy conservation and emission reduction to achieve the goals of “carbon neutrality” and “peak carbon emissions” (Fu and Zhou, 2023). In this context, the integration of DC power sources, represented by photovoltaic (PV), and DC loads, represented by electric vehicles, has become increasingly widespread in distribution networks. However, this integration method can result in network losses due to the need for AC/DC conversion. To address this challenge, researchers have explored the direct integration of DC power sources and loads into DC distribution networks, thereby avoiding the need for a large number of AC/DC conversion steps and reducing network losses. In addition, DC distribution networks offer several advantages, including high voltage quality and large transmission capacity, making them an attractive area of study for researchers (Blaabjerg et al., 2004; Wang et al., 2021; Zhang et al., 2022).

Considering that current distribution networks are primarily based on AC distribution networks, AC/DC hybrid distribution networks are considered a feasible transitional mode. This approach involves converting some AC branches into DC branches and connecting them through converters, leveraging the advantages of DC power sources and loads to reduce network losses. In the development of future power systems, AC/DC hybrid distribution networks have become an important direction (Wei et al., 2022). However, the high proportion of DC power sources, such as PV, being integrated into these networks has presented challenges. Due to the strong uncertainty of PV power output and its susceptibility to environmental factors, issues such as voltage fluctuations or exceeding limits are easily caused. Therefore, it is necessary to reasonably adjust the various controllable resources in the AC/DC distribution networks to achieve effective voltage control.

Currently, traditional voltage regulation methods for distribution networks include on-load tap changers (OLTC) (Jiao et al., 2022), reactive power compensation equipment (Kryonidis et al., 2021), and others. In (Wu et al., 2017), a voltage control method that combines OLTC and feeder control section (FCS) was proposed to minimize voltage deviation and DG output reduction. The authors used a least squares method to fit the objective function curve to achieve optimal multi-objective control. In (Pachanapan et al., 2012), a decentralized voltage control method for multi-time scale distributed generations (DGs) was proposed. The authors also proposed a principle for dividing regions based on the number of DGs, which effectively controlled the voltage of the distribution network (Valverde and Van Cutsem, 2013). Proposed a method that centralized the control of OLTC and DGs to solved the voltage control problem using quadratic programming algorithms. These methods can accurately determine the optimal voltage control strategy under given conditions, however, they require accurate physical models of the distribution network. With the increasingly complex structure of AC/DC distribution networks, it has become more difficult to estimate the model parameters accurately, requiring a large amount of measurement data to obtain line parameters. Traditional model-based voltage control methods are difficult to achieve ideal results in practical operation. Therefore, it is necessary to develop new data-driven voltage control methods that combine existing traditional methods to achieve precise control of the voltage in AC/DC distribution networks.

In solving the voltage control problem in distribution networks, it is necessary to address not only the issue of parameter acquisition, but also the uncertainty caused by the randomness of PV and load factors (Fu et al., 2020; Fu, 2022). Currently, there are two main methods for dealing with uncertainty in the system: stochastic programming (Bizuayehu et al., 2016) and robust optimization (Huang et al., 2022). Stochastic programming models uncertain factors as random variables and uses probability distribution functions to describe their stochastic properties. The optimization problem is then transformed into a stochastic programming problem and solved using corresponding stochastic optimization algorithms. For example, (Wei et al., 2022), established a probability model of distributed generation output and constructed a distribution network economic dispatch model considering the uncertainty of DG, with the goal of achieving economic efficiency while considering opportunity constraints (Nguyen and Crow, 2016). Established a scheduling model that includes renewable energy and energy storage, and used probability constraints to handle the uncertainty of renewable energy and load (Su et al., 2014). Established a stochastic output model of renewable energy, constructed a two-stage stochastic scheduling model with the goal of minimizing network loss. Robust optimization describes the range of changes in uncertain factors using a set of uncertainties. For example, (Xu et al., 2017), established a distributed robust control model of controllable devices in a distribution network with the goal of maximizing social welfare (Li et al., 2021). Evaluated the operating costs and risks of wind power to determine the allowable range of wind power output, and converted a centralized robust scheduling model into multiple sub-scheduling models for solution using the alternating direction method of multipliers. Although these methods can address uncertainty, they also have some limitations. For example, stochastic programming methods require obtaining the probability density function of random variables, but their actual distribution is often difficult to obtain and the use of artificially set methods cannot include all scenarios, thus leading to weak generalization. Robust optimization, although it does not require consideration of the distribution of random variables, needs to consider extreme scenarios, which can make the optimization results more conservative. In future research, it is necessary to consider the uncertainty of DG and loads and the actual situation of distribution networks, and to develop more effective voltage control methods to meet the needs of AC/DC hybrid distribution networks.

With the rapid development of AI technology, using Deep Reinforcement Learning (DRL) to solve sequential decision-making problems in distribution networks has become a hot research topic in recent years. DRL has excellent generalization ability and fast online decision-making speed, which can effectively deal with the uncertainty of PV and loads. Currently, DRL has been widely used in distribution network economic dispatch (Shuai et al., 2021), voltage control (Liu and Wu, 2021a; Liu and Wu, 2021b), reactive power optimization (Zhang et al., 2021), network reconfiguration (Oh et al., 2020), and other areas. In (Bai et al., 2023), the QMIX method was used to control DGs, and graph neural network (GNN) and graph attention (GAT) layers were introduced to enhance the information capture ability of agents. In (Xiang et al., 2023), based on the multi-agent deep deterministic policy gradient (MADDPG) method and perception of the network topology, the power of energy storage was controlled to achieve real-time voltage regulation. In (Yang et al., 2022), an advantage actor-critic (A3C)-based energy management strategy for distribution networks was proposed to solve large-scale decision-making problems. Currently, centralized control based on a single intelligent agent requires improved communication facilities, as well as centralized computing and storage centers. However, as the number of devices and network topology in AC/DC distribution networks increases, the amount of data to be collected also increases sharply, and centralized control is susceptible to the impact of local communication failures, thereby affecting control effectiveness.

The paper proposes a data-driven, model-free voltage control method for AC/DC distribution networks based on multi-agent reinforcement learning. Firstly, a power-voltage model is trained to capture the nonlinear mapping between the bus power and voltage using a large amount of historical data. Then, an agent is assigned to each subarea and a multi-agent deep reinforcement learning method (MADRL) is adopted for centralized training. Finally, the trained agents are deployed in a decentralized manner and each agent can make optimal decisions based on local observations only, achieving global optimal performance. This method does not require an accurate physical model of the distribution network and can control the inverters. Each converter can make optimal decisions based on local observations only, and the coordination of multiple converters can minimize the overall voltage deviation. The proposed method provides valuable insights and references for the future development of AC/DC distribution networks.

2 Problem description of AC/DC distribution networks voltage control

The AC/DC distribution networks mainly consists of a voltage source converter (VSC), an AC distribution network, and a DC distribution network. The VSC can control various variables such as AC/DC voltage, active power, and reactive power, and has various control modes. The V-Q control mode is used to control the voltage of the DC branch and adjust the power factor of the AC branch. The P-Q control mode is used to control the power exchange between the AC and DC distribution networks. Since the ratio of resistance to reactance of the distribution network line R/X is relatively large, both active and reactive power have a significant impact on the bus voltage. By adjusting the active and reactive power of the VSC, the voltage of the AC/DC distribution networks can be regulated.

Voltage control in AC/DC distribution networks is a complex nonlinear problem. Although traditional mathematical methods perform well in terms of solution accuracy and control effects, accurate mathematical models of the distribution network and differentiable objective functions are required, etc. The voltage control model is as follows:

$$\begin{cases} \min f(x) = \sum_i \left|V_{\text{ac},i} - V_{\text{ac0}}\right| + \sum_j \left|V_{\text{dc},j} - V_{\text{dc0}}\right| \\ \text{s.t. } g\left(P_{\text{ac},i}, Q_{\text{ac},i}, P_{\text{load},i}, Q_{\text{load},i}, P_{\text{PV},i}, P_{\text{ac,VSC}}, Q_{\text{ac,VSC}}\right) = 0 \\ g\left(P_{\text{dc},j}, P_{\text{load},j}, P_{\text{PV},j}, P_{\text{dc,VSC}}\right) = 0 \\ g\left(P_{\text{ac,VSC}}, P_{\text{dc,VSC}}\right) = 0 \\ -P_{\text{VSC, max}} \le P_{\text{ac,VSC}} \le P_{\text{VSC, max}} \\ -Q_{\text{VSC, max}} \le Q_{\text{ac,VSC}} \le Q_{\text{VSC, max}} \\ V_{\text{ac, min}} \le V_{\text{ac},i} \le V_{\text{ac, max}} \\ V_{\text{dc, min}} \le V_{\text{dc},j} \le V_{\text{dc, max}} \end{cases} \tag{1}$$

Where, in the objective function is to minimize the voltage deviation in the AC/DC distribution networks. $V_{\text{ac},i}$ represents the voltage of AC bus i; V_{ac0} represents the reference voltage of AC bus; $V_{\text{dc},i}$ represents the voltage of DC bus i; V_{dc0} represents the reference voltage of DC bus; g (·) denotes the load flow equation; $P_{\text{ac},i}$ and $Q_{\text{ac},i}$ are the active and reactive power injected at AC bus j, respectively; $P_{\text{load},i}$, $Q_{\text{load},i}$, and $P_{\text{PV},i}$ represent the active power, reactive power, and PV output of the load at AC bus i, respectively; $P_{\text{dc},i}$ denotes the active power injected at DC bus j; $V_{\text{ac,j}}$ represents the voltage of AC bus j; $V_{\text{dc},j}$ represents the voltage of DC bus j; $P_{\text{load},j}$ and $P_{\text{PV},j}$ are the active power and PV output of the load at DC bus j, respectively; $V_{\text{ac,max}}$ and $V_{\text{ac,min}}$ are the upper and lower limits of AC bus voltage, respectively; $V_{\text{dc,max}}$ and $V_{\text{dc,min}}$ are the upper and lower limits of DC bus voltage, respectively. $P_{\text{ac,VSC}}$ and $Q_{\text{ac,VSC}}$ represent the active and reactive power on the AC side of the VSC, respectively; $P_{\text{VSC,max}}$ and $Q_{\text{VSC,max}}$ are the upper limits of VSC active power and reactive power, respectively. Figure 1.

In reality, the following problems exist in the process of realizing the above optimization problems: 1) There are more types of devices in AC/DC distribution networks and the models are complicated, so it is difficult to obtain accurate physical models; 2) For distribution networks with large structures, the amount of information to be observed is large and the information transfer time is long, so it is difficult to make the optimal control decision quickly based on the centralized control method.

In summary, this paper proposes a model-free voltage control method for AC/DC distribution networks based on multi-agent soft actor critic (MASAC), which solves the problem of difficult to obtain model parameters by constructing a power-voltage model; by setting multi-agent to control multiple VSCs, each VSC can make optimal decisions based on the local observation state and realize distributed control.

3 Model-free coordinated voltage control framework for multiple converters

The model-free voltage control framework for multi-converter collaboration is shown in Figure 2. The steps are as follows: 1) Train the power-voltage model using the operation data of AC/DC distribution networks and establish the nonlinear mapping relationship between node power and node voltage through neural network. 2) Construct a Markov decision process for the voltage control problem and set an agent for each VSC. 3) Train the agents by MASAC method to solve the Markov decision problem. 4) Deploy the agents in a distributed manner to achieve real-time voltage control.

3.1 Power-voltage model

3.1.1 Model construction

There is a complex nonlinear mapping relationship between node power and voltage in the AC/DC distribution networks. Since neural networks have strong capability in fitting nonlinear models, Multilayer Perceptron (MLP) is used to solve this problem. The power-voltage model takes a six-layer structure, including one input layer, four hidden layers, and one output layer. Among them, the input layer includes the power of each node and the power of the converter, and the output layer is the voltage amplitude of each node, which can be expressed as:

$$[V_{\text{ac}}, V_{\text{dc}}] = \langle W^s, g^s\left(P_{\text{load},i}, Q_{\text{load},i}, P_{\text{PV},i}, P_{\text{VSC}}, Q_{\text{VSC}}\right)\rangle + b^s \tag{2}$$

Where, <·,·> is internal product; W^s is the weights; b^s is the bias; g^s (·) is the activation function.

The main training process consists of the following steps.

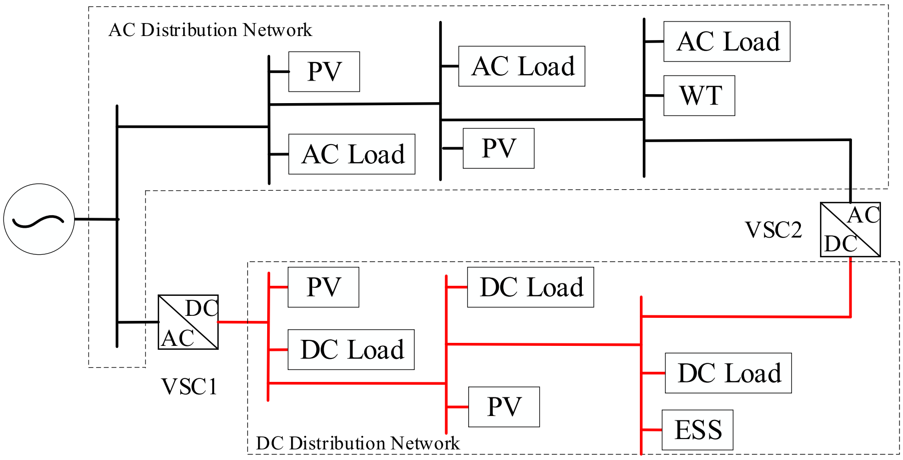

FIGURE 1
Typical structure of AC/DC distribution networks.

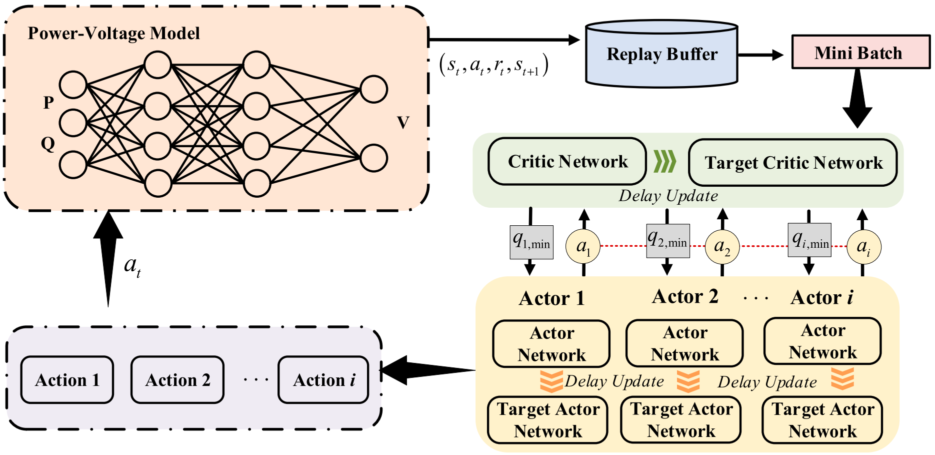

FIGURE 2
Model-free coordinated voltage control framework.

1) Initialization: random initialization of the weights and biases of the MLP model;
2) Forward propagation: $(P_{\text{load},i}, Q_{\text{load},i}, P_{\text{PV},i}, P_{\text{VSC}}, Q_{\text{VSC}})$ is passed to the input layer of the model and enters the output layer through the hidden layer, where the activation function is chosen as the ReLU function, and the analytic formula of the function is:

$$\text{ReLU}(\text{x}) = \max(0, x) \tag{3}$$

3) Calculated loss: The error between the voltage amplitude obtained by forward propagation and the actual voltage amplitude is calculated using the MSE function as:

$$L(y, \hat{y}) = \frac{1}{n}\sum (y_i - \hat{y}_i)^2 \tag{4}$$

Where, y is the predicted voltage value of the power-voltage model, $\hat{y}$ is the true voltage value.

4) Back propagation: Calculate the gradient of the loss function with respect to each weight and bias.
5) Updating weights and biases: With the objective of minimizing the loss function, the Stochastic Gradient Descent (SGD) algorithm is used to update the weights and biases:

$$\theta \leftarrow \theta - \eta \cdot \nabla L(\theta; x, y) \tag{5}$$

6) Iterative training: Repeat steps 2)-5) to train the power-voltage model.

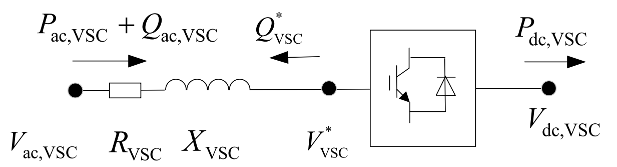

FIGURE 3
VSC model.

3.1.2 Sample generation

The training of the power-voltage model requires a large amount of operational data of the AC/DC distribution networks and converter stations. To verify the accuracy of the proxy model, it is based on real AC/DC distribution networks load data and accurate physical model of the distribution network. The voltage of the distribution network is obtained by power flow, and then the power-voltage model is trained and the accuracy of it is verified by test sets.

(1) AC Distribution Network Model

The model for the AC side of the AC/DC hybrid distribution network is:

$$P_{\mathrm{ac},j} = \sum_{k\in\pi^{jk}} P_{\mathrm{ac},jk} - \sum_{i\in\pi^{ij}} \left(P_{\mathrm{ac},ij} - I^2_{\mathrm{ac},ij} R_{\mathrm{ac},ij}\right) \tag{6}$$

$$Q_{\mathrm{ac},j} = \sum_{k\in\pi^{jk}} Q_{\mathrm{ac},jk} - \sum_{i\in\pi^{ij}} \left(Q_{\mathrm{ac},ij} - I^2_{\mathrm{ac},ij} X_{\mathrm{ac},ij}\right) \tag{7}$$

$$V^2_{\mathrm{ac},j} = V^2_{\mathrm{ac},i} - 2\left(P_{\mathrm{ac},ij} R_{\mathrm{ac},ij} + Q_{\mathrm{ac},ij} X_{\mathrm{ac},ij}\right) + I^2_{\mathrm{ac},ij}\left(R^2_{\mathrm{ac},ij} + X^2_{\mathrm{ac},ij}\right) \tag{8}$$

$$I^2_{\mathrm{ac},ij} V^2_{\mathrm{ac},ij} = P^2_{\mathrm{ac},ij} + Q^2_{\mathrm{ac},ij} \tag{9}$$

$$P_{\mathrm{ac},i} = P_{\mathrm{PV},i} - P_{\mathrm{load},i} \tag{10}$$

$$Q_{\mathrm{ac},i} = -Q_{\mathrm{load},i} \tag{11}$$

In the formula, $P_{\mathrm{ac,ij}}$ and $Q_{\mathrm{ac,ij}}$ are the active power and reactive power of AC branch *ij*, respectively; $R_{\mathrm{ac},ij}$ and $X_{\mathrm{ac},ij}$ are the resistance and reactance of AC branch *ij*, respectively; $I_{\mathrm{ac},ij}$ is the current of AC branch *ij*; π^{ij}, and π^{jk} are the sets of branches with AC bus *j* as the terminal node and AC bus *j* as the starting node, respectively.

(2) DC Distribution Network Model

The model for the DC side is:

$$P_{\mathrm{dc},j} = \sum_{k\in\Omega^{jk}} P_{\mathrm{dc},jk} - \sum_{i\in\Omega^{ij}} \left(P_{\mathrm{dc},ij} - I^2_{\mathrm{dc},ij} R_{\mathrm{dc},ij}\right) \tag{12}$$

$$V^2_{\mathrm{dc},j} = V^2_{\mathrm{dc},i} - 2P_{\mathrm{dc},ij} R_{\mathrm{dc},ij} + I^2_{\mathrm{dc},ij} R^2_{\mathrm{dc},ij} \tag{13}$$

$$I^2_{\mathrm{dc},ij} V^2_{\mathrm{dc},ij} = P^2_{\mathrm{dc},ij} \tag{14}$$

$$P_{\mathrm{dc},i} = P_{\mathrm{PV},i} - P_{\mathrm{load},i} \tag{15}$$

Where $P_{\mathrm{dc},ij}$ is the active power of DC branch *ij*; $R_{\mathrm{dc},ij}$ is the resistance of DC branch *ij*; $I_{\mathrm{dc},ij}$ is the current of DC branch *ij*; Ω^{ij}, and Ω^{jk} are the sets of branches with DC bus *j* as the terminal node and DC bus *j* as the starting node, respectively.

(3) VSC Model

The model of the VSC is shown in Figure 3. In the figure, Q^*_{VSC} represents the reactive power output of the VSC; $V_{\mathrm{ac,VSC}}$ is the AC side voltage magnitude, $V_{\mathrm{dc,VSC}}$ is the DC side voltage magnitude; $V^*_{\mathrm{VSC}}$is the equivalent internal potential magnitude, and R_{VSC} and X_{VSC} are the equivalent resistance and reactance, respectively.

The VSC model can be expressed as:

$$P_{\mathrm{ac,VSC}} - I^2_{\mathrm{VSC}} R_{\mathrm{VSC}} = P_{\mathrm{dc,VSC}} \tag{16}$$

$$Q_{\mathrm{ac,VSC}} - I^2_{\mathrm{VSC}} X_{\mathrm{VSC}} = -Q^*_{\mathrm{VSC}} \tag{17}$$

$$V^*_{\mathrm{VSC}} = \frac{\sqrt{3}}{3} \mu M_{\mathrm{VSC}} V_{\mathrm{dc,VSC}} \tag{18}$$

Where μ is the DC voltage utilization factor; 0≤μ ≤ 1, when the modulation mode is SPWM, it is set to 0.866; M_{VSC} is the modulation index of the VSC, $0\le M_{\mathrm{VSC}}\le 1$.

3.2 Markov decision process construction

In the whole AC/DC distribution system, there are multiple converters, and multiple converters cooperate cooperatively to achieve the minimum overall voltage deviation of the whole distribution system. The synergistic problem of multiple converters is modeled as Markov Decision Process (MDP), which mainly contains the following components:

3.2.1 Agent

In MDP, each VSC that needs to be controlled is modeled as an Agent.

3.2.2 Environment

The agent needs to get the power information of the nodes and the output of PV from the distribution network, and then make an action based on the local observed state, and then the power-voltage model gives the node voltage based on the power information and the action of the agents, which gives the agents a reward. Therefore, environment contains two parts: the AC/DC distribution networks and the power-voltage model.

3.2.3 State

The local observed state at moment t of agent *i* contains the PV output and load in region *i*, defined as $s_{t,i} = (P_{\mathrm{PV},t,i}, P_{\mathrm{load},t,i})$. The local observed states of all agents constitute the set of states $S_t = \{s_{t,1}, s_{t,2}.\cdots s_{t,i}\}$.

3.2.4 Action

The action at moment *t* of agent *i* contains the active power and reactive power of VSC *i*, defined as $a_{t,i} = (P_{\mathrm{VSC},t,i}, Q_{\mathrm{VSC},t,i})$. The action of all agents constitute the set of actions $A_t = \{a_{t,1}, a_{t,2}.\cdots a_{t,i}\}$.

3.2.5 Reward

In this paper, the agents are fully cooperative and the goal is to minimize the voltage deviation of the distribution network, so all agents use the same reward function. And the objective function in reinforcement learning is the maximum cumulative reward, so it is transformed into:

$$r_t = -\left(\sum_i \left| V_{\mathrm{ac},i,t} - V_{\mathrm{ac0}} \right| + \sum_j \left| V_{\mathrm{dc},i,t} - V_{\mathrm{dc0}} \right| \right) \tag{19}$$

3.3 MASAC-based markov decision process solving

Unlike standard reinforcement learning, SAC encourages the agents to explore a wider action space when performing tasks by incorporating maximizing policy entropy in policy optimization. This maximum entropy-based exploration method can make SAC more robust and flexible for PV and AC/DC distribution networks with high load uncertainty. The objective of the training process is:

$$\max J(\pi_i) = \max E_{(s_{t,i}, a_{t,i}) \sim \rho_{\pi_i}} \left\{ \sum_t \left[r(s_{t,i}, a_{t,i}) + \alpha_i \mathcal{H}\left(\pi_i\left(\cdot \middle| s_{t,i}\right)\right) \right] \right\} \tag{20}$$

Where, $J(\pi_i)$ is the expected cumulative return of the strategy; α_i is the temperature coefficient of entropy; $\mathcal{H}(\pi_i(\cdot|s_i))$ is the strategy entropy, which can be specifically expressed as:

$$\mathcal{H}\left(\pi_i\left(\cdot \middle| s_{t,i}\right)\right) = -\sum \pi_i\left(a_{t,i} \middle| s_{t,i}\right) \log \pi_i\left(a_{t,i} \middle| s_{t,i}\right) \tag{21}$$

To realize the training of multiple agents, a centralized training method with decentralized execution is adopted. The global critic of each intelligent body consists of two sets of action-value networks: the global twin soft Q network and the global target Q network, respectively. During training, MASAC uses the global critic to collect the observed states and actions of all the agents to obtain global information for evaluation. Among them, the Critic network is updated by minimizing the loss function, which can be expressed as:

$$L(\theta_i) = \left(y_t - Q_{\pi_i}(S_t, A_t)\right)^2 \tag{22}$$

$$y_t = r_t + \gamma E_{a_{t+1,i} \sim \pi_i'} \left[Q_{\pi_i'}'(S_{t+1}, A_{t+1}) - \alpha_i \log\left(\pi_i'\left(a_{t+1,i} \middle| s_{t+1,i}\right)\right) \right] \tag{23}$$

Where π'_i and Q_{π_i}' are target actor network and target Q network, respectively. θ_i is the parameter of Q_{π_i}, γ is discount factor.

Each agent has its own actor network and still makes actions through local observation states during execution without interacting with other agents. The actor network of each agent is updated by gradient ascent, and the gradient is calculated by the formula:

$$\nabla J(\phi_i) = E_{s_{t,i} \sim D, a_{t,i} \sim \pi_i} \left[\nabla \log\left(\pi_i\left(a_{t,i} \middle| s_{t,i}\right)\right) \left(-\alpha_i \log\left(\pi_i\left(a_{t,i} \middle| s_{t,i}\right)\right) + B(S_t, a_{t,\backslash i})\right) \right] \tag{24}$$

$$B(S_t, a_{t,\backslash i})) = Q_{\pi_i}(S_t, A_t) - b(S_t, a_{t,\backslash i}) \tag{25}$$

$$b(S_t, a_{t,\backslash i}) = \sum_{a_{t,i} \sim \pi_i} \pi_i\left(a_{t,i} \middle| s_{t,i}\right) Q_{\pi_i}(S_t, (a_{t,i}, a_{t,\backslash i})) \tag{26}$$

Where ϕ_i is the parameter of π_i; B (S_t, $a_{t,\backslash i}$) is a function to determine whether the increase in reward is attributed to other agents; b (S_t, $a_{t,\backslash i}$) is the base line of multi-agent, which represents the average value of all actions of the agents in the current state. D is replay buffer, which is used to store experience and make the training process of neural network more stable.

During each parameter update, the loss of α can be calculated according to the formula for the automatic entropy tuning mechanism:

$$L(\alpha_i) = -\mathbb{E}_{s_{t,i} \sim D, a_{t,i} \sim \pi_i} \left[\alpha_i \log \pi_i(a_{t,i} \mid s_{t,i}) - \alpha \log \frac{1}{|\mathcal{A}_i|} \right] \tag{27}$$

Where, $|\mathcal{A}_i|$ is the dimension of the action space.

The value of α_i is updated using the Adam optimizer:

$$\alpha_i \leftarrow \exp\left(\log \alpha_i - \lambda \nabla_{\alpha_i} L(\alpha_i)\right) \tag{28}$$

Where λ is the learning rate.

The algorithm of supposed method is shown in Algorithm 1.

```
Initialize parameters of power-voltage model $\theta_S$. Randomly generate power data of VSC based on historical load data of AC/DC distribution networks
Calculate the power flow to obtain voltage data of AC/DC distribution networks
Train the power-voltage model and update the parameters $\theta_S$ of the agent model
Initialize parameters $\phi_i$ of Actor of SAC agent $i$. Initialize parameters $\theta_{i,1}$ and $\theta_{i,2}$ of global Critic network
Initialize the replay buffer $[\mathcal{D}_i]_{i=1}^n$
Initialize parameters of target networks: $\phi_i' \leftarrow \phi_i$, $\theta_{i,1}' \leftarrow \theta_{i,1}$, $\theta_{i,2}' \leftarrow \theta_{i,2}$
for each epoch do
  for each time do
      observe the state $[s_i]_{i=1}^n$ of environment
    select action $a_i \sim \phi_i$
    feed action $[a_{t,i}]_{i=1}^n$ to environment (power-voltage model), get reward $r_t$ and next state $[s_{t+1,i}]_{i=1}^n$
      $\mathcal{D}_i \sim \mathcal{D}_i \cup (s_{t,i}, a_{t,i}, r_t, s_{t+1,i})$
    end
  for each update step of agent do
    Sample a batch $(s_{t,i}, a_{t,i}, r_t, s_{t+1,i})$ from $\mathcal{D}_i$
    update parameters $\phi_i$ of Actor. update parameters $\theta_{i,1}$ and $\theta_{i,2}$ of global Critic network
    If update then
      $\phi_i' \leftarrow \phi_i + (1-\tau)\phi_i'$, $\theta_{i,1}' \leftarrow \theta_{i,1} + (1-\tau)\theta_{i,1}'$, $\theta_{i,2}' \leftarrow \theta_{i,2} + (1-\tau)\theta_{i,2}'$
    end
end
```

Algorithm 1. Model-Free MASAC Voltage Control.

First, the parameters θ_s of the power-voltage model are randomly initialized. Based on the generated AC/DC distribution networks operation data, the power-voltage model is trained and the accuracy of the model is tested, and the trained model is embedded in a reinforcement learning environment to participate in the training of the agents.

Then, each agent is trained based on the strategy of centralized training. The parameters ϕ_i of actor network, the parameters $\theta_{i,1}$ and $\theta_{i,2}$ of critic network are initialized and copied to target actor network and target critic network. The agents are trained in M rounds, each round contains T hours, and at moment t, the agent i makes an action $a_{t,i}$ to control the active and reactive power of the VSC according to

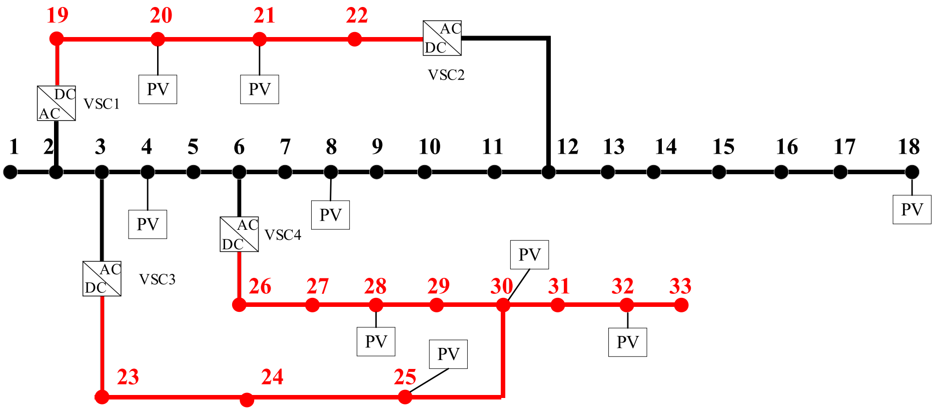

FIGURE 4
Improved IEEE 33-bus AC/DC distribution network.

the locally observed state $s_{t,i}$. The actions of all the agents form the action set A_t at moment t and acting on the power-voltage model, which gives the voltage deviation and reward r_t. Finally the environment enters the next state $s_{t+1,i}$, and $(s_{t,i}, a_{t,i}, r_t, s_{t+1,i})$ is stored in the replay buffer D_i for subsequent updates.

At the beginning of training, the actions of each agent are randomly generated, and when the total number of rounds reaches the set value M_{set}, each agent starts to select actions according to the actor network and the neural network starts to be updated. At each update, the agent draws a batch of the history from the replay buffer for network update. In this case, the parameter minimization loss function of the critic network is updated, which can be expressed as:

$$L(\theta_{i,n}) = \frac{1}{N_{\text{batch}}} \sum_{m=1}^{N_{\text{batch}}} (y_m - Q_{\pi_i,n}(S_m, A_m))^2 \tag{29}$$

$$\theta_{i,n} \leftarrow \theta_{i,n} + \lambda \nabla L(\theta_{i,n}) \tag{30}$$

Where N_{batch} is the batch size; λ is the learning rate.

The actor network is updated by maximizing the policy gradient, which can be expressed as:

$$\begin{aligned} \nabla J(\phi_i) = E_{s_{t,i}\sim D, a_{t,i}\sim \pi_i} \Big[\nabla \log\big(\pi_i(a_{t,i}|s_{t,i})\big)\big(-\alpha \log\big(\pi_i(a_{t,i}|s_{t,i})\big) \\ + B(S_t, a_{t,\backslash i})\big)\Big] \end{aligned} \tag{31}$$

$$\phi_i \leftarrow \phi_i + \lambda \nabla J(\phi_i) \tag{32}$$

The target actor network and target critic network update formulas are as follows:

$$\phi_i{}' \leftarrow \phi_i + (1-\tau)\phi_i{}' \tag{33}$$

$$\theta'_{i,1} \leftarrow \theta_{i,1} + (1-\tau)\theta'_{i,1} \tag{34}$$

$$\theta'_{i,2} \leftarrow \theta_{i,2} + (1-\tau)\theta'_{i,2} \tag{35}$$

Where τ is the parameter that controls the update weight.

3.4 Distributed execution of agents

After the training of each agent is completed, the parameters of the actor network are fixed and put into distributed coordination control of the VSCs. Each agent can make actions based on the local observed state only and does not depend on the communication between VSCs. Since the critic network incorporates the global state and the actions of all the agents during training, the actions made by each agent based on local observations are well coordinated during distributed control, and the global voltage deviation can be minimized.

4 Example analysis

4.1 Improved IEEE 33-bus AC/DC distribution network example

To validate the effectiveness of the proposed method, the improved IEEE 33-node AC/DC hybrid distribution network is selected in this paper, as shown in Figure 4. The black solid lines represent AC lines, the red solid lines represent DC lines, and the AC/DC hybrid lines are connected through VSCs. The rated voltage of the AC distribution network is 12.66 kV, and the rated voltage of the DC distribution network is 15 kV. VSC1 and VSC3 use constant voltage and constant reactive power control on the DC side and AC side, respectively. VSC2 and VSC4 use constant active power and constant reactive power control, with an active power capacity of 1.5 MW and a reactive power capacity of 0.4 Mvar for each VSC. There are a total of nine PV arrays connected to the system, with three arrays connected on the AC side with an installed capacity of 0.1 MW each, and six arrays connected on the DC side with an installed capacity of 0.2 MW each.

4.2 Performance testing of the model-free for AC/DC distribution networks

In this paper, 26,200 sets of active and reactive power data for VSC2 and VSC4 were randomly generated. Combined with historical load data of each bus, the voltage data for 33 nodes over 26,200 h were obtained through power flow calculation. The first 25,000 h of data were used as the training set, and the data from 25,001 to 26,200 h were used as the test set to test the performance of the model-free AC/DC

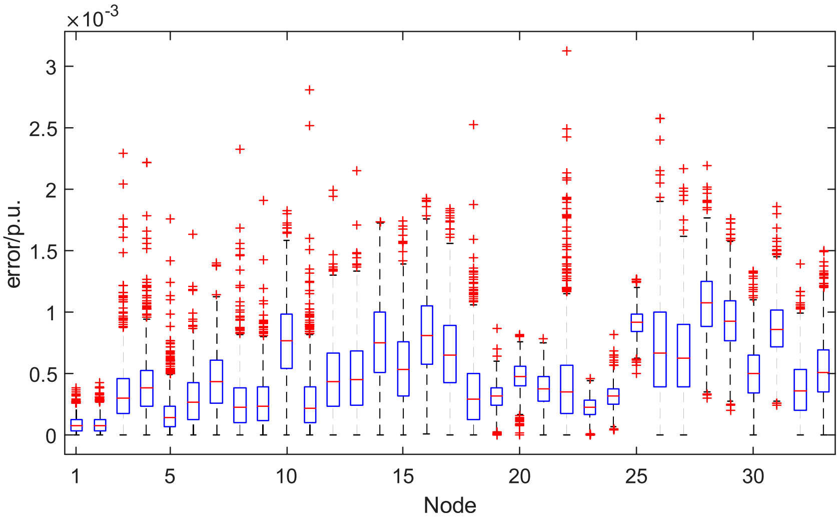

FIGURE 5
Error distribution of the model-free for AC/DC distribution networks.

distribution networks. The number of neurons in each layer was set to 55, 512, 1,024, 2,048, and 33, with a batch size of 128, a learning rate of 1e-5, and the mean squared error (MSE) as the loss function.

The error distribution between the predicted voltage values and the actual values for each bus is shown in Figure 5. As can be seen from the figure, 90% of the error data is distributed in the interval [4.06e-05, 1.16e-3], the maximum error is 3.125e-3 p.u., and the average error is 4.927e-4 p.u. From the results, it can be concluded that the error between the predicted voltage amplitude and the actual voltage amplitude is small, which verifies the effectiveness of the proposed method.

TABLE 1 The example settings.

Case	The example settings
Case 1	model-free, MASAC
Case 2	model-free, SAC
Case 3	accurate model, MASAC
Case 4	accurate model, SAC
Case 5	accurate model, PSO

4.3 Convergence and control performance analysis of the algorithm

To evaluate the effectiveness of the proposed method, this paper sets up five test cases and compares the control performance of different methods, as shown in Table 1. Among them, Cases 1–4 use DRL-based methods for control, while Case 5 uses traditional evolutionary algorithms for control. Cases 1 and 3 use multi-agent training methods, and each agent makes decisions based on local observation states in their respective regions. Cases 2 and 4 use centralized control, with a single agent making decisions based on global observation states. Cases 1 and 2 use the model-free for AC/DC distribution networks for training, while Cases 3 and 4 use the accurate model for training. The discount factor, smoothing factor, learning rate, and batch size are set to 0.99, 0.005, 0.0003, and 256, respectively. The simulation platform is configured with an AMD 3970 × 3.7 GHz CPU, NVIDIA 3090 GPU, and 64 GB memory computer. The algorithm is implemented using the Python language and PyTorch framework.

Except for the particle swarm optimization algorithm, all other methods were trained for 2,000 iterations, with each iteration consisting of a control period of 4,380 h and a time step of 1 h. The convergence curves of different algorithms are shown in Figure 6, and their performance metrics are presented in Table 2.

In terms of convergence time, the convergence curves of the four cases are roughly the same since the environment settings, random seeds, and agent parameters are the same, all of which reach convergence at around 1,200 rounds. However, the required time is different. Cases 3 and 4, which use the accurate model for training, have the longest convergence time, taking more than a day. In contrast, cases 1 and 2, which use the model-free for AC/DC distribution networks based on the historical data, can reduce the training time to a few hours. The convergence time of Cases 1 and 3, which use multi-agent algorithms, is longer than that of single-agent algorithms because more neural network parameters are updated each time.

In terms of control performance, examples 2 and 4, which adopt single-agent centralized control, make decisions based on the global state when applied online, while in the multi-agent setting, each agent makes decisions based on the local observation state, resulting in slightly worse control performance than single-agent centralized control. Examples 3 and 4, which were trained using accurate models, have smaller voltage deviations and better control performance than examples 1 and 2, which were trained using the proposed data-driven approach. However, the difference is small. It can be seen that examples 1–4 outperform the particle swarm algorithm in control performance, as the particle swarm algorithm is prone to getting stuck in local optimal solutions, while the SAC algorithm requires exploration of actions more extensively to maximize the

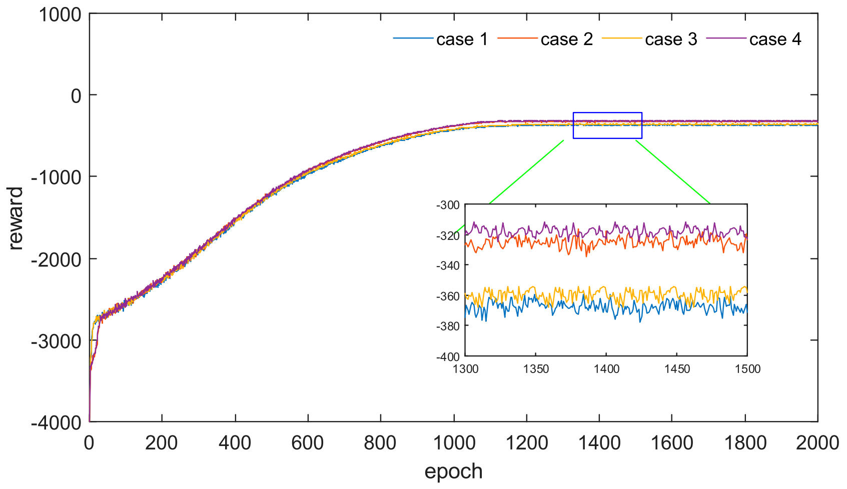

FIGURE 6
Cumulative reward.

TABLE 2 Comparison results of the case studies.

Case	Convergence time (s)	Online decision-making time (s)	Mean voltage deviation (%)	Maximum voltage (p.u.)	Minimum voltage (p.u.)	Whether it depends on an accurate physical model	Can distributed control be achieved?
Case 1	9,526	0.019	0.26	1.0125	0.9774	×	√
Case 2	5,908	0.009	0.23	1.0124	0.9783	√	×
Case 3	95,440	0.026	0.25	1.0156	0.9805	×	√
Case 4	90,500	0.019	0.22	1.0156	0.9760	√	×
Case 5	—	5,760	0.36	1.0156	0.9632	√	×

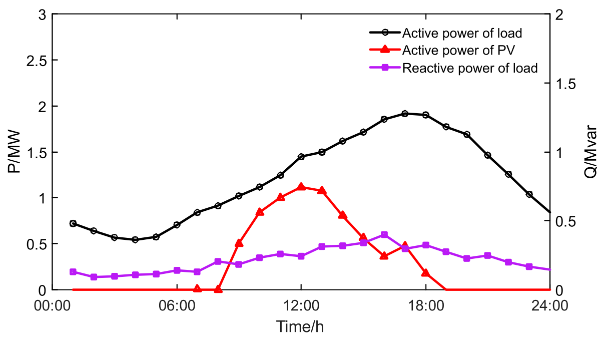

FIGURE 7
The overall load and PV variation.

entropy while maximizing cumulative rewards, thus having a stronger ability to find optimal solutions during training.

In terms of decision-making time, DRL-based algorithms have good generalization performance and can adapt to the fluctuations of PV and load. They can also generate decisions within milliseconds when facing different scenarios, which can meet the real-time voltage control needs. On the other hand, traditional evolutionary algorithms have weaker generalization performance. They need to be re-solved when facing different scenarios and cannot achieve real-time control of voltage.

Overall, the DRL-based methods outperform traditional evolutionary algorithms in terms of real-time decision-making based on observed states, without the need to solve optimization models again. The proposed model-free multi-agent control method can achieve similar performance as the accurate model-based method, without requiring accurate physical models of the AC/DC distribution system, greatly reducing training time, and allowing for distributed control of the VSCs. The trained agents can make decisions based on local state observations, achieving good control performance.

4.4 Validation of example results

To further verify the generalization ability and control effectiveness of the proposed method, the load and PV data of

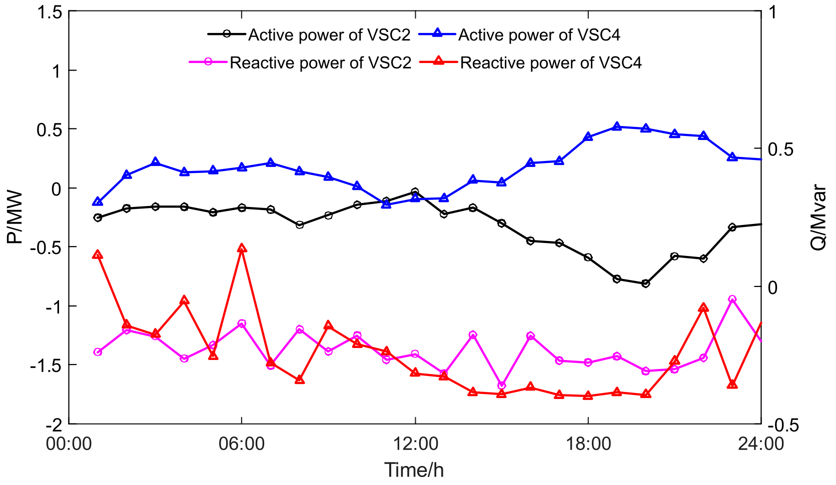

FIGURE 8
The active power and reactive power of the VSC.

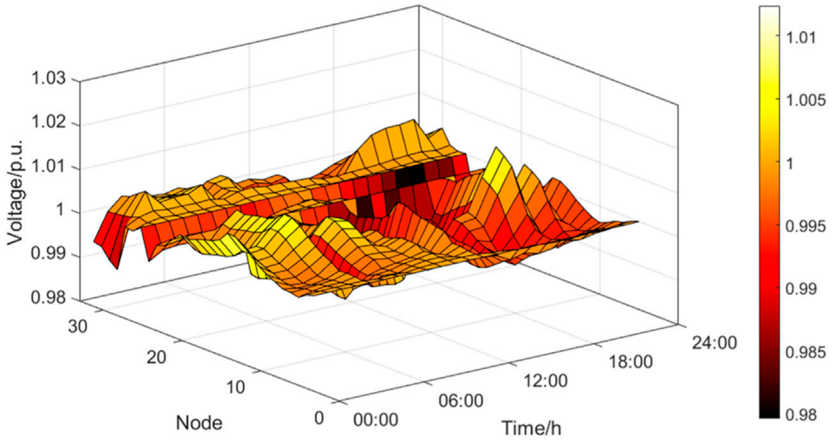

FIGURE 9
Voltage distribution.

1 day outside the training period were selected as the test set, and the trained intelligent agent was tested online. The overall load and PV variation are shown in Figure 7.

The agent generates actions to control the active and reactive power of VSC2 and VSC4 based on the real-time observed state information, as shown in Figure 7. Figure 8 It can be seen that the trend of the active power of VSC2 is roughly opposite to that of the net load, with negative power, i.e., in the inverter state. The active power of VSC4 is roughly the same as the trend of the net load, with positive power, i.e., in the rectifier state. By controlling the active and reactive power of VSC, voltage regulation is achieved, and the voltage distribution of each node in different periods is shown in Figure 9. The average voltage deviation is 0.316%, the highest voltage is 1.012 p.u., and the lowest voltage is 0.978 p.u. It can be seen that the policy generated by the agent has good control effect. At the same time, generating 24 sets of control actions only takes 0.019 s, and the average time consumed for generating one set of actions is 7.916e-04 s, indicating a fast speed. Through online testing of the agent, the fast decision-making and good generalization ability of the proposed method are verified.

5 Conclusion

To achieve real-time voltage control in AC/DC distribution systems, this paper proposes a model-free voltage control method based on multi-agent reinforcement learning, which has the following advantages.

(1) The proposed method uses a agent model to reflect the nonlinear mapping relationship between node power and voltage in distribution systems, without requiring accurate physical models of the system, thus solving the problem of difficult parameter acquisition.

(2) The proposed method trains agents using the MASAC algorithm to solve the model-free voltage control problem, which has strong generalization ability. When applied online, the agents can adapt to the fluctuation of photovoltaic and load in milliseconds and make decisions in real-time, achieving real-time control.
(3) The proposed method uses a "centralized training, decentralized execution" strategy, which can achieve distributed cooperative control of the converters. Each converter only needs to make the optimal decision based on local observation, minimizing the global voltage deviation.

Through simulation experiments, the proposed method has advantages over traditional evolutionary algorithms and particle swarm optimization algorithms in terms of control performance, decision time, and generalization ability. The online testing results on the test set further validate the generalization ability and real-time performance of the proposed method, indicating that it can achieve effective voltage control. Future research can further apply the proposed method to actual AC/DC distribution systems and combine it with actual data to verify and improve the algorithm.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

QZ wrote the manuscript. ZH and SW organized case studies. YD and GQ contributed to the theoretical research of this manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the State Grid Corporation of China (5400-202199280A-0-0-00).

Acknowledgments

The authors would like to thank the referees and the editor of this journal for valuable comments.

Conflict of interest

YD and GQ were employed by State Grid Tianjin Electric Power Company.

The remaining 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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fenrg.2023.1202701/full#supplementary-material

References

Bai, Y., Chen, S., Zhang, J., Xu, J., Gao, T., Wang, X., et al. (2023). An adaptive active power rolling dispatch strategy for high proportion of renewable energy based on distributed deep reinforcement learning. *Appl. Energy* 330, 120294. doi:10.1016/j.apenergy.2022.120294

Bizuayehu, A. W., Sanchez de la Nieta, A. A., Contreras, J., and Catalao, J. P. S. (2016). Impacts of stochastic wind power and storage participation on economic dispatch in distribution systems. *IEEE Trans. Sustain. Energy* 7, 1336–1345. doi:10.1109/TSTE.2016.2546279

Blaabjerg, F., Chen, Z., and Kjaer, S. B. (2004). Power electronics as efficient interface in dispersed power generation systems. *IEEE Trans. Power Electron.* 19, 1184–1194. doi:10.1109/TPEL.2004.833453

Fu, X., Guo, Q., and Sun, H. (2020). Statistical machine learning model for stochastic optimal planning of distribution networks considering a dynamic correlation and dimension reduction. *IEEE Trans. Smart Grid* 11, 2904–2917. doi:10.1109/TSG.2020.2974021

Fu, X. (2022). Statistical machine learning model for capacitor planning considering uncertainties in photovoltaic power. *Prot. Control Mod. Power Syst.* 7, 5. doi:10.1186/s41601-022-00228-z

Fu, X., and Zhou, Y. (2023). Collaborative optimization of PV greenhouses and clean energy systems in rural areas. *IEEE Trans. Sustain. Energy* 14, 642–656. doi:10.1109/TSTE.2022.3223684

Huang, B., Li, Y., Zhan, F., Sun, Q., and Zhang, H. (2022). A distributed robust economic dispatch strategy for integrated energy system considering cyber-attacks. *IEEE Trans. Ind. Inf.* 18, 880–890. doi:10.1109/TII.2021.3077509

Jiao, W., Chen, J., Wu, Q., Li, C., Zhou, B., and Huang, S. (2022). Distributed coordinated voltage control for distribution networks with DG and OLTC based on MPC and gradient projection. *IEEE Trans. Power Syst.* 37, 680–690. doi:10.1109/TPWRS.2021.3095523

Kryonidis, G. C., Malamaki, K. N. D., Gkavanoudis, S. I., Oureilidis, K. O., Kontis, E. O., Mauricio, J. M., et al. (2021). Distributed reactive power control scheme for the voltage regulation of unbalanced LV grids. *IEEE Trans. Sustain. Energy* 12, 1301–1310. doi:10.1109/TSTE.2020.3042855

Li, P., Yang, M., Tang, Y., Yu, Y., and Li, M. (2021). Robust decentralized coordination of transmission and active distribution networks. *IEEE Trans. Ind. Appl.* 57, 1987–1994. doi:10.1109/TIA.2021.3057342

Liu, H., and Wu, W. (2021a). Online multi-agent reinforcement learning for decentralized inverter-based volt-VAR control. *IEEE Trans. Smart Grid* 12, 2980–2990. doi:10.1109/TSG.2021.3060027

Liu, H., and Wu, W. (2021b). Two-stage deep reinforcement learning for inverter-based volt-VAR control in active distribution networks. *IEEE Trans. Smart Grid* 12, 2037–2047. doi:10.1109/TSG.2020.3041620

Nguyen, T. A., and Crow, M. L. (2016). Stochastic optimization of renewable-based microgrid operation incorporating battery operating cost. *IEEE Trans. Power Syst.* 31, 2289–2296. doi:10.1109/TPWRS.2015.2455491

Oh, S. H., Yoon, Y. T., and Kim, S. W. (2020). Online reconfiguration scheme of self-sufficient distribution network based on A reinforcement learning approach. *Appl. Energy* 280, 115900. doi:10.1016/j.apenergy.2020.115900

Pachanapan, P., Anaya-Lara, O., Dysko, A., and Lo, K. L. (2012). Adaptive zone identification for voltage level control in distribution networks with DG. *IEEE Trans. Smart Grid* 3, 1594–1602. doi:10.1109/TSG.2012.2205715

Shuai, H., Li, F., Pulgar-Painemal, H., and Xue, Y. (2021). Branching dueling Q-network-based online scheduling of a microgrid with distributed energy storage systems. *IEEE Trans. Smart Grid* 12, 5479–5482. doi:10.1109/TSG.2021.3103405

Su, W., Wang, J., and Roh, J. (2014). Stochastic energy scheduling in microgrids with intermittent renewable energy resources. *IEEE Trans. Smart Grid* 5, 1876–1883. doi:10.1109/TSG.2013.2280645

Valverde, G., and Van Cutsem, T. (2013). Model predictive control of voltages in active distribution networks. *IEEE Trans. Smart Grid* 4, 2152–2161. doi:10.1109/TSG.2013.2246199

Wang, X., Gu, L., and Liang, D. (2021). Decentralized and multi-objective coordinated optimization of hybrid AC/DC flexible distribution networks. *Front. Energy Res.* 9, 762423. doi:10.3389/fenrg.2021.762423

Wei, W., Hao, T., and Xu, T. (2022). Day-ahead economic dispatch of AC/DC hybrid distribution network based on cell-distributed management mode. *Front. Energy Res.* 10, 832243. doi:10.3389/fenrg.2022.832243

Wu, H., Huang, C., Ding, M., Zhao, B., and Li, P. (2017). Distributed cooperative voltage control based on curve-fitting in active distribution networks. *J. Mod. Power Syst. Clean. Energy* 5, 777–786. doi:10.1007/s40565-016-0236-1

Xiang, Y., Lu, Y., and Liu, J. (2023). Deep reinforcement learning based topology-aware voltage regulation of distribution networks with distributed energy storage. *Appl. Energy* 332, 120510. doi:10.1016/j.apenergy.2022.120510

Xu, Y., Yang, Z., Gu, W., Li, M., and Deng, Z. (2017). Robust real-time distributed optimal control based energy management in a smart Grid. *IEEE Trans. Smart Grid* 8, 1568–1579. doi:10.1109/TSG.2015.2491923

Yang, Y., Li, H., Shen, B., Pei, W., and Peng, D. (2022). Microgrid energy management strategy base on UCB-A3C learning. *Front. Energy Res.* 10, 858895. doi:10.3389/fenrg.2022.858895

Zhang, L., Yu, S., Cai, Y., Tang, W., and Cheng, H. (2022). Reliability evaluation method of AC/DC hybrid distribution network considering voltage source converter restoration capability and network reconfiguration. *Front. Energy Res.* 10, 899985. doi:10.3389/fenrg.2022.899985

Zhang, X., Liu, Y., Duan, J., Qiu, G., Liu, T., and Liu, J. (2021). DDPG-based multi-agent framework for SVC tuning in urban power Grid with renewable energy resources. *IEEE Trans. Power Syst.* 36, 5465–5475. doi:10.1109/TPWRS.2021.3081159