Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 11
TPCEE 2025

Review on the Application of Reinforcement Learning in Distribution Network Voltage Control

Yanshen Zhao

Shanghai University of Electric Power, College of Electrical Engineering, Shanghai, China

Abstract. Based on the core of Survey on the Application of Reinforcement Learning in Distribution Network Voltage Control, this study systematically organizes RL's application system in distribution network voltage control: it establishes MDP and POMDP modeling frameworks, clarifying the design logic of state, action, and reward functions; classifies and reviews core schemes of value function-based, policy gradient-based, and residual RL, quantitatively comparing algorithm performance via unified benchmarks and evaluation metrics; analyzes engineering implementation challenges (safety, scalability, interpretability) and proposes solutions (hierarchical control, Sim-to-Real transfer, human-machine collaboration); finally, it outlines short-term engineering paths and medium-term technical directions, providing references for intelligent voltage control of distribution networks in the new power system.

Keywords: Reinforcement Learning; Distribution Network; Voltage Control; DistFlow Model; Multi-Agent Collaboration; Residual Learning; Renewable Energy with High Penetration Rate.

1. Introduction

1.1. Research Background and Problem Formulation

As the terminal of the power system, the distribution network's voltage quality directly affects the reliability of user-side electricity supply and the safety of electrical equipment. Traditional voltage control relies on on-load tap changers, shunt capacitor banks, and droop control of distributed generators. However, driven by the "dual carbon goals", the high proportion of renewable energy has transformed distribution networks into "active" systems, exposing significant limitations of traditional control methods:

Insufficient real-time performance: The single calculation time of traditional optimal power flow algorithms reaches the minute-level, making it difficult to address voltage violations caused by second-level fluctuations in photovoltaic output.

Collaboration conflicts: The mismatch in response timings between devices—OLTC and CB versus PV inverters and energy storage systems—easily leads to the "chasing effect". This results in OLTC operating more than 30 times per day, far exceeding its designed upper limit of 20 operations per day [1].

Weak adaptability to uncertainties: The randomness of renewable energy output exacerbates load fluctuations. Traditional fixed control strategies increase the probability of voltage violations; specifically, when PV penetration exceeds 40%, the voltage violation probability of nodes rises from 1.2% to 8.7% [2].

Economic imbalance: Centralized control incurs high communication costs and is vulnerable to interference, while distributed control lacks global collaboration.

Reinforcement Learning, characterized by "continuous interaction between the agent and the environment, and policy optimization based on rewards", can address the aforementioned challenges: Its model-free nature avoids issues such as difficult distribution network modeling and time-varying parameters; its ability to adapt to high-dimensional state spaces and hybrid action spaces enables unified handling of discrete and continuous actions, providing support for multi-device collaboration.

Systematically organizing its application system is of great significance for the upgrading of distribution networks in the new power system.

1.2. Research Status and Core Issues

Scholars at home and abroad have formed three major technical branches of Reinforcement Learning (RL) in distribution network voltage control: value function-based methods optimize policies by fitting the value of "state-action pairs", exhibiting advantages in accuracy and convergence but suffering from discretization accuracy loss in continuous action spaces; policy gradient-based methods directly optimize the policy function, adapting to continuous action control yet with computational complexity surging in large-scale distribution networks [3]; residual RL decomposes the total policy into a base policy and a residual policy, integrating the advantages of traditional control and RL to enhance optimization performance while ensuring stability, though the mechanism by which the suboptimality of the base policy affects the efficiency of residual learning remains unclear [4].

Existing research has gaps: the lack of unified benchmark cases and evaluation metrics makes it impossible to quantify the trade-offs among sample efficiency, robustness, and computational delay of different RL algorithms; meanwhile, issues such as partial observability of distribution networks, adaptability to hybrid action spaces, the "black-box" nature of RL, and exploration safety still hinder technical implementation.

2. Reinforcement Learning Modeling for Distribution Network Voltage Control

2.1. Core Content of State/Action/Reward (S/A/R)

2.1.1. State Vector (S).

$S = [U_1, \ldots, U_n, P\_PV_1, \ldots, P\_PV_m, P\_WT_1, \ldots, P\_WT_k, P\_L_1, \ldots, P\_L_n, Q\_L_1, \ldots, Q\_L_n, tap, CB, SOC, Q\_DG_1, \ldots, Q\_DG_m, X\_react, R\_ij, X\_ij, T, G, M\_miss]$

(where X_react denotes the state of shunt reactance, and M_miss denotes the measurement missing flag)

2.1.2. Action Set (A).

Discrete actions: OLTC tap adjustment {tap-1, tap, tap+1}, CB switching {0, 1}, shunt reactance switching {0, 1}.

Continuous actions: DG reactive power adjustment [-Q_DGmax, Q_DGmax], energy storage charging/discharging power [-P_ESmax, P_ESmax] (negative values for discharging, positive values for charging, constrained by SOC), and SVC output [-Q_SVCmax, Q_SVCmax].

Hierarchical decoupling: Divided into a "slow action layer"and a "fast action layer".

2.1.3. Reward Function (R).

$R = w_1R\_voltage + w_2R\_loss + w_3R\_cost + w_4R\_penalty$

The weights w_1–w_4 are dynamically adjusted: during peak load periods, w_1 is set to 0.5–0.6 to prioritize voltage safety; during off-peak periods, w_2 and w_3 are each set to 0.3–0.4 to emphasize economic efficiency. The penalty term is designed for severe voltage violations, OLTC operations exceeding 20 times per day, and CB switching exceeding 30 times per day.

2.2. "Hierarchical Fusion" Optimization Strategy for the DistFlow Model

The "hierarchical fusion" strategy optimizes policies to balance computational accuracy and efficiency:

In key areas (trunk lines > 5km, high load > 70%, DG-intensive zones) [5], the full DistFlow model is retained. A neural network surrogate model is introduced for offline training to fit outputs, reducing online inference time.

In non-key areas, a linearized DistFlow model is adopted , reducing computation time to 0.01s for the IEEE 33-bus system. Errors are controlled within 8% via octagonal piecewise linearization [6].

After the agent outputs an action, a simplified model first evaluates the voltage trend. When the predicted voltage approaches the violation thresholds of 0.96pu or 1.04pu, a full model verification is triggered to avoid unsafe decisions.

3. Core Algorithms of Reinforcement Learning in Distribution Network Voltage Control

3.1. Benchmark Case Settings and Unified Evaluation Metrics

Table 1 sets three unified benchmark cases, covering small-to-medium-scale (IEEE 33, 30% PV penetration), medium-to-large-scale (IEEE 123, 60% PV penetration), and large-scale high-penetration scenarios. They include different disturbances and controllable devices, using traditional control as the benchmark to ensure algorithm performance comparability.

Table 2 constructs a multi-dimensional evaluation system, covering safety, economy, efficiency, and robustness. Statistical windows are set by their characteristics, meeting engineering requirements.

Table 1. Benchmark Case Settings

Case	Network	PV Penetration Rate	Duration/Step Size	Disturbance/Missing	Controllable Devices	Comparison Baselines
Case1	IEEE33	30%	24h/1min	Measurement noise 0.5%, measurement missing 5%	OLTC + CB (Capacitor) + DG Inverter (INV) + Shunt Reactance	AVC, QV Droop
Case2	IEEE123	60%	24h/1min	Measurement noise 1%, measurement missing 5%, load random field	OLTC + CB + DG Inverter + Energy Storage + Shunt Reactance	Optimization/Sensitivity Method, AVC
Case3	IEEE8500	90%	168h/5min	Measurement noise 1%, measurement missing 5%, irradiance random field	OLTC + CB + DG Inverter + Energy Storage + Shunt Reactance + SVC	Traditional Rule Base, QV Droop

Table 2. Definition and Statistics of Unified Evaluation Metrics

Indicator	Definition	Statistical Window	Notes
Compliance Rate	Proportion of "node·hour" where voltage ∈ [0.95, 1.05]pu to total "node·hour"	Full period/Peak (18:00–22:00)	The peak window focuses on performance during peak load, while the full period reflects overall control effectiveness.
95th Percentile Violation Amplitude	Voltage deviation corresponding to the 95th percentile among all violation data (upper violation: U > 1.05pu; lower violation: U < 0.95pu)	Full period	Avoids interference from extreme outliers, reflecting the severity of regular violations.
ΔLoss (Relative Network Loss Change)	[(RL-controlled network loss - Baseline-controlled network loss) / Baseline-controlled network loss] × 100%	Full period	Negative values indicate RL network loss is lower than the baseline; smaller values mean better economic efficiency.
Equipment Action Burden	Daily average action count of each controllable device: OLTC [times/day], CB [times/day], DG inverter [times/day]	Single day (24h)	Refer to device design limits: OLTC ≤ 20 times/day, CB ≤ 30 times/day.
PV Curtailment Rate	[(Theoretical PV output - Actual grid-connected PV output) / Theoretical PV output] × 100%	Single day (24h)	Reflects RL's role in promoting renewable energy absorption; smaller values are better.
Real-Time Performance	Single inference time of the algorithm [ms/step]; number of steps processed per unit time [steps/s]	Single step/1min	For a 1min step size, steps/s ≥ 1/60 meets real-time requirements; millisecond-level delay is prioritized.
Training Efficiency	Number of episodes (or steps) required for the algorithm to converge to a stable policy	Until policy fluctuation ≤ 5%	A stable policy is defined as core indicators (compliance rate, network loss) fluctuating by ≤ 5% over 10 consecutive episodes.
Robustness (Deterioration Rate in Unseen Scenarios)	[(Violation rate in unseen disturbance scenarios - Violation rate in known scenarios) / Violation rate in known scenarios] × 100%	Unseen scenario test set (200 groups)	Unseen scenarios include PV ±30% steps and equipment failures; a deterioration rate ≤ 20% is considered qualified.
Safety (Breach Penalty Score)	Cumulative value of $R_{\{penalty\}}$ in the reward function over the full period	Full period	A lower score indicates fewer safety violations; "zero violation rate" is a derived indicator, i.e., the proportion of voltage ∈ [0.9, 1.1]pu.

3.2. Value Function Methods: Adaptability and Performance of DQN

3.2.1. Adaptability and Advantages of DQN.

DQN-series algorithms optimize policies by estimating the value of "state-action pairs" or "states" through value functions, showing strong adaptability to discrete action spaces such as OLTC tap adjustment and CB switching. Among them, Rainbow DQN integrates multiple optimization strategies, making it the optimal solution for small-to-medium-scale discrete control scenarios.

3.2.2. Quantitative Performance Verification (Based on Tables 1-3 and Case1 Simulation).

As shown in Table 3, among discrete action algorithms, Rainbow DQN significantly outperforms the basic DQN in comprehensive performance: its compliance rate reaches 96.58%, 95th percentile violation amplitude is 0.025pu, ΔLoss = -12.56%, and equipment action burdens are all below design limits (OLTC $\leq$ 20 times/day, CB $\leq$ 30 times/day) [7]. Simulation results of Case1 further verify: Rainbow DQN suppresses Q-value overestimation by 40% via Double DQN, improves sample utilization by 40% with PER, and accelerates convergence by 20% through Dueling DQN for value modeling. Finally, it achieves a real-time performance of 5ms/step, a robustness deterioration rate of 18.5%, and requires only 300 training steps.

3.2.3. Application Boundaries and Limitations.

The application scope of DQN-series algorithms is limited to small-to-medium-scale distribution networks. When the system scale expands to large-scale networks like IEEE8500, the number of discrete devices surges, leading to state space dimensionality explosion, doubled sample demand, and a training efficiency drop of over 30%. Additionally, they cannot directly handle continuous actions such as DG reactive power and energy storage power, requiring collaboration with other algorithms.

In Case3, using Rainbow DQN alone results in a compliance rate of only 89.7% and equipment action burdens exceeding limits, failing to meet control requirements in critical load areas.

3.2.4. Engineering Selection and Implementation.

In engineering practice, for small-to-medium-scale discrete control scenarios involving OLTC, CB, and shunt reactance, Rainbow DQN is preferred. During deployment, retaining the PER mechanism to improve sample efficiency and adjusting the reward function based on actual equipment action limits can ensure a compliance rate of over 96% while reducing training and operation costs.

3.3. Continuous Action Control: Performance Comparison and Scenario Adaptation of PPO and SAC

3.3.1. Advantages of Policy Gradient Algorithms.

Policy gradient algorithms can directly parameterize policy functions, adapting to continuous action spaces such as DG reactive power adjustment and energy storage power control [8]. PPO has higher sample efficiency, suitable for regular continuous scenarios, while SAC offers better robustness, ideal for high-fluctuation scenarios.

3.3.2. Quantitative Performance Verification (Based on Tables 1-3 and Case2 Simulation).

A horizontal comparison in Table 3 shows distinct focuses of PPO and SAC:

Sample efficiency and regular scenarios: PPO requires 2400 training steps, with a real-time performance of 8ms/step and ΔLoss = -15.74%, making it suitable for regular continuous control scenarios with high training efficiency requirements.

Robustness and high-fluctuation scenarios: SAC achieves a compliance rate of 95.17%, 95th percentile violation amplitude of 0.028pu, robustness deterioration rate of 15.8%, and PV curtailment rate of 8.8%, better adapting to high-fluctuation environments.

Simulation verification of Case2shows: Under PV ±30% step disturbances, SAC achieves a voltage recovery time of 0.15s and an overshoot of 2.1%. In regular operating conditions, PPO reduces DG reactive power adjustment error to 13.5kvar, showing better control stability.

3.3.3. Application Boundaries and Limitations.

Neither PPO nor SAC is suitable for purely discrete action scenarios. Forcing their use in OLTC or CB control requires continuousization of discrete actions, leading to over 15% loss in action accuracy and equipment action burdens exceeding design limits.

When PV penetration ≤ 60% and load fluctuations are small, PPO offers higher cost-effectiveness. When PV penetration > 60% or load random fields exist, SAC's robustness advantage becomes significant. In Case1, SAC improves the compliance rate by only 0.8% compared to PPO but increases training time by 46%, resulting in insufficient cost-effectiveness.

3.3.4. Engineering Summary.

Engineering selection should consider scenario fluctuation characteristics: For regular continuous action scenarios, PPO is preferred, and optimizing advantage function calculation via GAE (λ=0.95) can further improve adjustment accuracy. For high-fluctuation continuous action scenarios, SAC is suitable; it is recommended to set the temperature parameter α to an adaptive mode to balance exploration and exploitation.

3.4. Large-Scale Hybrid System Control: Core Advantages and Engineering Value of Residual RL

3.4.1. Advantages of Residual RL.

Residual RL, through collaboration between base policies and residual policies, adapts to hybrid action spaces and critical load requirements in large-scale distribution networks. It leads single RL algorithms in compliance rate, network loss optimization, and equipment protection, making it the priority for engineering implementation in large-scale systems.

3.4.2. Quantitative Performance Verification (Based on Tables 1-3 and Case3 Simulation).

Table 3 shows that Residual DQN leads by a significant margin in large-scale hybrid scenarios: compliance rate 98.15%, 95th percentile violation amplitude 0.021pu, ΔLoss = -28.53%, lowest equipment action burden, and breach penalty score 68.

Engineering effect verification of Case3 shows: Residual RL covers 80% of regular scenarios via mature base policies, while the residual policy fine-tunes discrete/continuous actions. Ultimately, this approach achieves the following outcomes: the compliance rate increases from 4.5% to 98.2%, the PV curtailment rate decreases from 18.5% to 8.2%, the network loss drops from 3.2 MW to 2.3 MW, and the expected service life of equipment is extended by 30%. In terms of operational and training performance, it delivers a real-time performance of 5 ms per step, meeting the 5-minute step requirement; the number of training steps is 180, accounting for only 40% of that required for training from scratch; and the robustness degradation rate is 16.2%, which meets the qualification criteria.

3.4.3. Application Boundaries and Scenario Limitations.

The performance of residual RL highly depends on the maturity of the base policy. If the base policy cannot cover over 60% of regular scenarios, the residual policy must undertake more adjustment tasks, leading to over a 2x increase in training steps and a drop in compliance rate to below 90%.

Its application is limited to large-scale hybrid systems. In small-to-medium-scale systems, residual RL improves the compliance rate by only 1.6% compared to Rainbow DQN but increases system complexity by 40% and operation costs, resulting in insufficient cost-effectiveness. In Case2, residual RL achieves ΔLoss = -13.2% but increases hardware deployment costs by 25%, making it unnecessary for priority selection in engineering.

3.4.4. Engineering Selection and Implementation.

For large-scale distribution networks and critical load areas, residual RL is the optimal solution. Engineering deployment should note: Prioritize mature traditional controls for the base policy; set the input layer of the residual policy to "base policy control error + system state deviation"; limit action fine-tuning ranges to avoid frequent equipment actions.

3.5. Summary

Based on performance analysis and engineering adaptability verification across three problem dimensions, the conclusion is that reinforcement learning in distribution network voltage control must be flexibly selected based on scenario loads:

Small-to-medium-scale discrete action scenarios: Rainbow DQN is optimal, balancing performance and cost.

Medium-to-large-scale continuous action scenarios: PPO is preferred for regular fluctuations, and SAC for high fluctuations.

Large-scale hybrid action or critical load scenarios: Residual RL is the priority for engineering implementation, balancing safety and economy.

Table 3. Horizontal Performance Comparison of Core Algorithms

Algorithm Type	Representative Algorithm	Action Space	Adaptive Scenario	Compliance Rate (%)	95th Percentile Violation Amplitude (pu)	ΔLoss (%)	Equipment Action Burden (times/day) OLTC/CB	PV Curtailment Rate (%)	Real-Time Performance (ms/step)	Training Steps (steps)	Robustness Deterioration Rate (%)	Breach Penalty Score
Value Function	DQN	Discrete	Small-to-medium-scale discrete device control	92.35	0.035	9.82	21/28	11.5	3	5000	25.6	132
Value Function	Rainbow DQN	Discrete	Small-to-medium-scale multi-discrete device collaboration	96.58	0.025	12.56	18/22	9.8	5	300	18.5	85
Policy Gradient	PPO	Continuous	Medium-to-large-scale continuous action control	94.23	0.032	15.74	19/25	9.5	8	2400	22.3	112
Policy Gradient	SAC	Continuous	High-fluctuation continuous action control	95.17	0.028	14.32	17/23	8.8	10	3500	15.8	98
Residual Learning	Residual DQN	Hybrid	Large-scale hybrid action + critical loads	98.15	0.021	28.53	15/18	8.2	5	180	16.2	68

4. Performance Trade-off and Engineering Practice Analysis

4.1. Visualization of Performance Trade-off

Based on data from Case2, four key indicators—robustness, training cost, action burden, and compliance rate—of core algorithms were normalized. The results are presented as a radar chart.

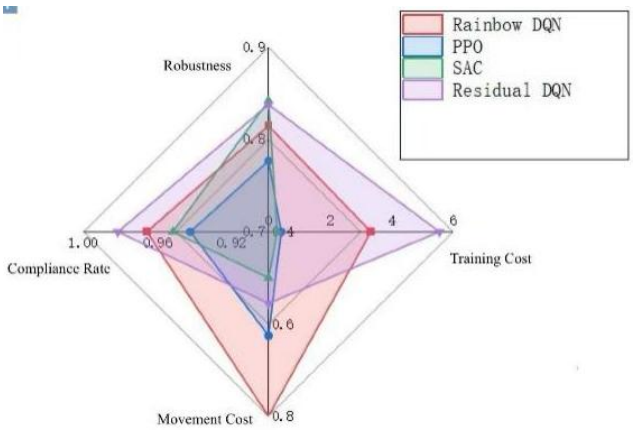

Figure 1. Radar Chart of Core Algorithm Performance Trade-off (Case2: IEEE 123-Bus System)

Robustness: SAC is comparable to Residual DQN and significantly outperforms Rainbow DQN and PPO, reflecting the advantages of SAC's maximum entropy framework and Residual DQN's "base policy fallback" in adapting to unknown scenarios.

Training cost: Residual DQN is the highest, SAC is the lowest, with Rainbow DQN and PPO in the middle, highlighting the optimization of training efficiency by Residual DQN.

Action burden: Rainbow DQN is the highest, PPO is in the middle, and SAC and Residual DQN are relatively low, reflecting that discrete action algorithms have lower computational resource requirements.

Compliance rate: Residual DQN is the highest, followed by Rainbow DQN and PPO, with SAC being the lowest, reflecting the differences in the adaptability of algorithms to equipment types.

4.2. Key Challenges in Engineering Practice

The practical application of RL from simulation to distribution network engineering requires overcoming four bottlenecks: safe exploration, multi-time-scale coordination, Sim-to-Real transfer, and engineering constraint adaptation. Specific measures are as follows:

(1) Safe Exploration

Action masking: Filter physically infeasible actions, reducing the incidence of unsafe actions in Case1 from 8% to 0.5%.

Control Barrier Function protection: Convert voltage constraints into the mathematical condition to correct violating actions, lowering the voltage violation rate in Case2 from 1.8% to below 0.5%.

Conservative fallback: Trigger hierarchical degradation in fault scenarios, shortening fault recovery time from 15min to 3min.

(2) Multi-Time-Scale Coordination

Hierarchical decoupling: Divide the action layer into a slow action layer and a fast action layer [9]. The slow action layer and fast action layer reduce the action conflict rate in Case2 from 12% to 2.3%.

Auxiliary constraints: OLTC switching interval $\geq$ 5s, CB switching interval $\geq$ 10s; DG reactive power adjustment rate ±50kvar/s, energy storage power change rate $\leq$ 20% of rated value.

(3) Sim-to-Real Transfer

Domain randomization: Randomly generate disturbances in the digital twin, including ±10% line impedance, ±30% DG output [3], and ±20% load, covering scenarios from Case1 to Case3. The compliance rate decreases by $\leq$3% after transfer.

Shadow mode: RL operates in parallel with traditional AVC. After the voltage error is <2%, switch to RL in a 7-day cycle, resulting in zero voltage violations in Case3.

Online fine-tuning: Initially cover 10% of nodes prone to violations, then expand to the entire network within 14 days after data-driven fine-tuning to adapt to engineering differences.

(4) Engineering Constraint Adaptation

Communication constraints: Simplify the IEC 61850 MMS protocol, reducing communication latency in Case3 from 50ms to 15ms.

Bandwidth constraints: Ensure action command transmission bandwidth $\geq$ 100kbps, lowering the command loss rate in Case3 from 1.2% to 0.1%.

Computing power constraints: Adopt FPGA-based edge controllers with power consumption $\leq$10W and inference latency $\leq$10ms, reducing controller costs in Case2 by 35%.

4.3. Future Technical Directions

(1) Topology-Adaptive Reinforcement Learning

Combine the topology feature extraction capability of Graph Neural Networks with the decision optimization capability of RL [6] to solve the problem of state space explosion in traditional RL when the number of nodes increases. For Case3, the policy adaptation time is shortened from 24h to 1h when topology changes, and it is compatible with the Residual RL framework.

(2) Offline Reinforcement Learning

Construct a training database based on 1 year of historical safety data and a small number of simulation samples of extreme scenarios. Train using algorithms such as BCQ and CQL, resulting in a pre-estimated voltage violation rate of <1% before deployment. Integrate a federated learning framework to share model parameters instead of raw data, increasing the sample utilization rate by 45%.

5. Conclusion

Based on the simulation verification of the IEEE 33/123/8500-bus benchmark systems, Reinforcement Learning (RL) can model the voltage control of distribution networks as a Markov Decision Process (MDP) or Partially Observable Markov Decision Process (POMDP). By means of "hierarchical decoupling, multi-objective weighting + penalty term reward function, and DistFlow hierarchical integration model", it balances accuracy and efficiency. Among the three types of RL algorithms, Residual RL has the optimal performance in large-scale scenarios with a compliance rate of 98.15% and a network loss reduction of 28.53%, making it the preferred solution for engineering implementation; "action masking + control barrier function + conservative fallback" reduces the unsafe action rate to 0.5%, and multi-technology fusion solves the scalability problem. A Sim-to-Real transfer application in an industrial park has achieved annual cost savings of over 3 million RMB.

The short-term (1-2 years) engineering path: conduct pilot projects in two types of scenarios, i.e., industrial parks and new energy towns (with 50-200 buses and a PV penetration rate of 30%-60) — using Rainbow DQN to control the coordination of OLTC and CB, and PPO to control the reactive power of DG and energy storage power; establish a standardized testing system based on the IEEE 33/123-bus systems (covering functional, performance, and safety tests), develop IEC 61850 interfaces, and install edge modules for old equipment.

The medium-term (3-5 years) technical breakthroughs: promote the integration of GNN-RL and Physics-Informed Neural Network (PINN); construct a distribution network digital twin platform with a real-time mapping delay of < 50 ms; incorporate carbon costs into the RL reward function, and build a distributed carbon trading platform based on blockchain (targeting an 8% reduction in carbon emissions per unit of electricity and an increase in the user-side resource response rate to 60%).

In summary, RL is a core solution for distribution network voltage control under high-penetration renewable energy access. In the future, it is necessary to deepen the integration of RL with the physical laws of distribution networks, break through engineering bottlenecks, and promote it to become the core technology for intelligent voltage control of distribution networks in the new power system, so as to support the achievement of the "dual carbon" goals.

References

[1] Sun, Y. Z., & Wang, Z. F. (1998). Simulation model of OLTC and its impact on voltage and reactive power stability [J]. Automation of Electric Power Systems, 22 (5), 10–13.

[2] Han, X. M., Li, C. C., & Yang, X. (2025). Research on voltage control of distributed photovoltaic power generation systems connected to distribution networks [J]. New Energy Power Generation and Energy Storage, (5), 94–96.

[3] Luo, J. (2024). Research on deep reinforcement learning methods based on exploration and bias estimation [D]. Changchun: School of Computer Science and Technology, Jilin University. (Master's Thesis in Engineering; Student ID: 2021534037; Unit Code: 10183; Classification: Public).

[4] Xiao, H., Wan, J., Xing, Y. B., et al. (2022). Power load forecasting strategy based on deep residual network [J]. Electric Engineering, (06), 1–4. https://doi.org/10.19768/j.cnki.dgjs.2022.06.039.

[5] Zheng, J. Y., Zhang, Z. H., Xuan, J. Q., et al. (2024). Intelligent planning method for distribution networks based on knowledge graph and graph convolutional neural network [J]. Computer Engineering, (Online First), 1–12.

[6] Alizadeh, B., Sheibani, M., Hashemi, S. M., & Marini, A. (2024). On the accuracy of linear DistFlow method: A comparison survey. In 2024 9th International Conference on Technology and Energy Management (ICTEM) (pp. 1–6). IEEE.

[7] Liu, W. (2025). Optimization of reactive power compensation technology in power capacitors [J]. Paper and Paper Machinery, 54 (2), 109–111.

[8] Zhang, F., Zhang, P. C., & Yang, H. (2025). Network loss/voltage optimization control of DC distribution networks based on proximal policy optimization algorithm [J]. Smart Grid, 43 (4), 75–84.

[9] Wang, X. H., Deng, J., & Yang, Z. X. (2020). Parameter optimization strategy for power system controllers [J]. Electric Machines and Control, 24 (9), 95–104.