EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

Reinforcement Learning-enhanced Policy-aware Modeling of Smart Grid Efficiency under Carbon Constraints: Integration of SBM DEA and Dynamic Policy Response Simulation

Gemei Shi

Guangzhou University of Software, Guangzhou 510990, China

Abstract

Artificial intelligence-based sensing, forecasting, and decision optimization are being rapidly integrated into smart grid operations. Although reinforcement learning has created new opportunities for dispatch optimization and low-carbon transition under carbon constraints, most existing studies focus primarily on short-term economic or operational objectives and rarely incorporate system-level carbon reduction efficiency benchmarks into the learning process. To address this gap, this study proposes an integrated SBM-DEA and reinforcement learning framework for policy-aware smart-grid dispatch, in which carbon reduction efficiency scores are transformed from static evaluation results into dynamic learning signals for dispatch optimization. Using panel data from 30 provinces in China over the indicator system covering capital input, labor input, electricity service output, and electricity-related carbon dioxide emissions. An SBM-DEA model with undesirable outputs is employed to measure the carbon reduction efficiency of smart grids. The estimated efficiency scores are then embedded into both the state representation and reward reinforcement learning framework, where the agent learns dispatch policies that balance economic performance, carbon constraints, and efficiency improvement. A dynamic policy response simulation environment is further constructed, incorporating a hybrid energy storage system comprising battery storage and pumped hydro storage. The results show that the carbon reduction efficiency of smart grids in China exhibits stage-specific fluctuations, with annual average values ranging from 0.505 to 0.568 and pronounced interprovincial disparities. In the simulation learning agent trained with efficiency-based penalties achieves 7.3% lower operational costs and 8.5% higher average efficiency compared to an economic-only agent. The trained policies also exhibit clear policy-responsive behavior: carbon prices rise, hybrid storage utilization increases and coal-fired generation declines. The main innovation of this study is that it integrates historical efficiency benchmarking with reinforcement learning-based dispatch optimization, policy-aware and efficiency-guided decision-support framework for carbon-constrained smart grids.

Keywords: artificial intelligence; reinforcement learning; smart grid; carbon reduction efficiency; SBM-DEA; dynamic policy response

Received on 03 April 2026, accepted on 25 May 2026, published on 04 June 2026

doi: 10.4108/ew.12473

*Corresponding Author. Email: 609084168@qq.com

1. Introduction

Artificial intelligence-driven sensing, forecasting, and control are being rapidly integrated into the entire operational workflow of power systems. Strategy-learning methods such as reinforcement learning interact with grid operating environments to learn dispatching and coordination policies, thereby improving dispatch efficiency, system resilience, and renewable energy accommodation capability [1-2]. Under the constraints of carbon peaking and carbon neutrality goals, power systems are accelerating their transition from fossil-fuel-dominated supply toward a clean, low-carbon, and more coordinated paradigm [3]. Installed renewable capacity continues to expand, while the structure of electricity demand is also evolving. As a result, power grids require greater operational flexibility, as well as stronger capabilities in state awareness, source-grid-load-storage coordination, and digitalized control. In particular, hybrid energy storage systems combining batteries with pumped hydro or other technologies have emerged as critical assets for balancing renewable variability and enhancing grid stability [4-6]. At the same time, grid expansion, energy storage deployment, and digital infrastructure development are generally characterized by large capital requirements, long construction cycles, and strong system coupling [7,8]. In this context, how to use artificial intelligence to achieve refined state awareness, accurate prediction of multi-source disturbances, and adaptive optimization of dynamic control has become a key research direction for new-type power systems. In particular, under scenarios where carbon constraints and safety constraints coexist, safe reinforcement learning provides technical support for deployable and verifiable policy learning [9,10].

In recent years, artificial intelligence research in power systems has rapidly expanded from conventional machine learning to deep learning, graph learning, Transformers, federated learning, and foundation models [11-13]. Deep learning-based modeling has been widely applied to tasks such as load forecasting, renewable power forecasting, fault diagnosis, state estimation, transient stability analysis, and demand response [14,15]. Compared with methods based on heuristic rules or shallow statistical models, deep learning and its variants demonstrate stronger capabilities in feature extraction and relational modeling when handling the high-dimensional, nonlinear, strongly coupled, and multi-timescale data generated during power system operation [16-18].

As research advances from sensing and forecasting toward decision-making and control, reinforcement learning is emerging as an important technological pathway for smart grid operational optimization. Glavić pointed out that reinforcement learning and deep reinforcement learning have already been applied to power system control and related problems, demonstrating considerable potential for closed-loop control and online updating [19]. Liang et al. addressed the real-time scheduling problem of integrated energy systems by introducing a Soft Actor-Critic framework, formulating the operating process as a Markov decision process, and thereby reducing system operating costs while improving renewable energy utilization [20]. Li et al. proposed a data-driven adaptive control method based on Deep Deterministic Policy Gradient for voltage control in active distribution networks, enabling rapid response to distributed generation fluctuations and suppressing voltage violations [21]. Xiang et al. further developed a topology-aware multi-agent deep reinforcement learning model, allowing distributed energy storage to participate in real-time voltage regulation under topology-changing conditions [22]. Recent advances have also explored reinforcement learning for hybrid energy storage management, demonstrating its effectiveness in coordinating multiple storage technologies to improve grid flexibility and economic performance [23,24]. Despite these advances, existing RL-based approaches typically focus on local control objectives and rarely incorporate system-level efficiency benchmarks into the learning process. Moreover, the ability to respond dynamically to evolving carbon policies remains underexplored.

Accurate battery state estimation is essential for smart grid operation. Recent studies have improved battery state-of-charge and state-of-energy estimation by combining Bayesian optimization, bidirectional long short-term memory networks, improved neural network algorithms, and electrochemical-thermal coupling models. For example, Wang et al. proposed an improved hyperparameter Bayesian optimization-bidirectional long short-term memory model for high-precision battery state-of-charge estimation [25]. Wang et al. reviewed improved neural network algorithms for battery state-of-energy estimation in smart grids and emphasized the importance of estimation accuracy, robustness, computational efficiency, and adaptability [26]. In addition, Wang et al. developed an improved multiple feature-electrochemical thermal coupling model with real-time coefficient correction for lithium-ion batteries under low-temperature operating conditions [27]. These studies show that intelligent energy storage operation increasingly depends on the integration of data-driven estimation, physical modeling, and adaptive decision-making. However, most existing studies focus on battery state estimation or component-level modeling, while the integration of system-level carbon reduction efficiency evaluation with reinforcement learning-based dispatch remains insufficiently explored.

To address these gaps, this study proposes an RL-enhanced policy-aware modeling framework that integrates system-level carbon reduction efficiency measures into the reinforcement learning process. The proposed method consists of four main steps. First, provincial panel data are collected to construct an input-output indicator system for smart-grid carbon reduction efficiency evaluation. Second, an SBM-DEA model with undesirable outputs is used to estimate the carbon reduction efficiency of provincial smart grids. Third, the estimated efficiency scores are incorporated into the Markov decision process as both state information and reward-penalty feedback. Fourth, a PPO agent is trained in a dynamic policy response simulation environment with

hybrid battery and pumped-hydro storage, so that the learned policy can adapt to changing carbon prices and green credit intensities. Compared with existing DEA-based efficiency studies, the proposed framework does not treat carbon reduction efficiency as only an ex post evaluation result. Instead, it further uses the estimated efficiency scores to guide subsequent dispatch learning. Compared with conventional reinforcement learning-based dispatch studies, the proposed framework improves the reward design by incorporating system-level efficiency feedback, so that the learned policy is guided by both economic performance and carbon reduction efficiency. The overall analytical framework of this study is shown in Figure 1, which illustrates the logical connection among carbon reduction efficiency measurement, reinforcement learning-based dispatch optimization, and dynamic policy response simulation.

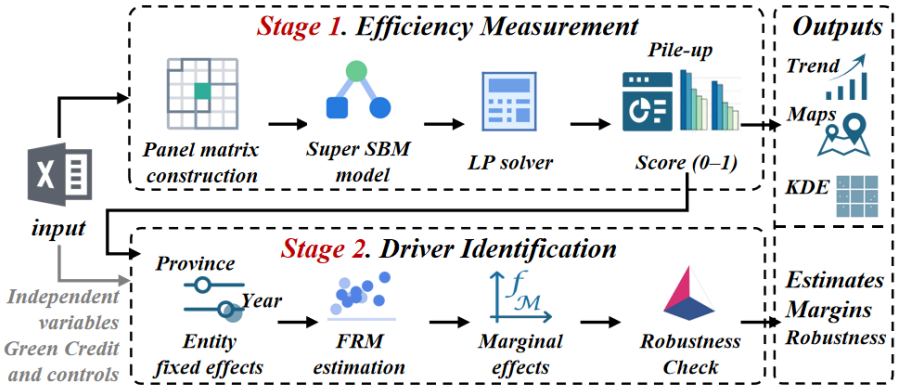

Figure 1. Analytical framework for smart-grid carbon reduction efficiency measurement and reinforcement learning-based dispatch optimization

2. Data and SBM-DEA Efficiency Measurement

2.1 Data Sources and Variable Definitions

This study uses China's provincial-level regions as the research sample over the period 2011–2022. Tibet is excluded due to missing data, and Hong Kong, Macao, and Taiwan are not included in the statistical scope, resulting in a panel dataset of 30 provinces. The sample period is chosen primarily because key indicators such as industry-level interest expenses are available with good continuity during these years [28]. In addition, the CEADs provincial CO_2 emission inventories can be matched over time to ensure consistent measurement definitions [29].

Data are mainly obtained from authoritative public statistical sources. Industry-level interest expenses are drawn from the China Industry Statistical Yearbook [28]. The undesirable output is measured using the CEADs provincial CO_2 emission inventories, and sector-consistent emissions are extracted for the "electricity and heat production and supply" industry [29]. Control variables, including GDP per capita, industrial structure, and urbanization, are collected from the China Statistical Yearbook and related official statistical materials [30]. Power-sector fixed-asset investment, employment, and electricity consumption are compiled from the China Electric Power Statistical Yearbook and the China Energy Statistical Yearbook [31,32].

To enhance robustness and intertemporal comparability, all raw variables are pre-processed in a uniform manner. Monetary variables are deflated to constant 2011 prices using province-specific price indices. Continuous variables are winsorized at the 1st and 99th percentiles to mitigate the influence of extreme values [33]. Ratio variables are constrained to their feasible domain, with values above 1.0 truncated to 1.0. A small number of missing observations are imputed using linear interpolation or adjacent-period filling [34].

2.2 Indicator Construction for Efficiency Measurement

In the efficiency measurement phase, the input side selects fixed asset investment and employment in the power industry to represent capital and labor inputs, respectively; the desirable output utilizes total electricity consumption to characterize the scale of power supply services; and the undesirable output adopts CO_2 emissions from the electric power and heat production and supply industry based on CEADs sectoral accounts. A summary of variable definitions is provided in Table

Table 1. Summary of variable definitions for efficiency measurement

Type	Variable	Symbol	Measure
Input	Power investment & labor	L	Power FAI; power employment
Good output	Energy services	Y_g	Electricity use; GDP
Bad output	CO_2 emissions	Y_b	Power-related CO_2 (CEADs: power & heat)

2.3 Measuring Carbon Reduction Efficiency: SBM-DEA with Undesirable Outputs

Traditional radial DEA models imply proportional contraction or expansion of inputs or outputs, making it difficult to simultaneously characterize non-radial features such as input redundancy, desirable output shortfalls, and undesirable output excess. Considering that smart grid carbon reduction efficiency includes the three-dimensional requirements of resource input, power supply services, and carbon emission constraints, this paper adopts the Slacks Based Measure (SBM) model [35] and incorporates carbon emissions as an undesirable output into a unified planning framework.

Suppose there are n decision making units (DMUs) in year t, where each DMU uses m inputs to produce s_1 desirable outputs accompanied by s_2 undesirable outputs. For the evaluated unit o, the SBM model with undesirable outputs is expressed as:

$$\rho = \frac{1-\frac{1}{m}\Sigma_i\left(\frac{s_i^-}{x_{i0}}\right)}{1+\frac{1}{s_1+s_2}\left(\Sigma_r\frac{s_r^{g+}}{y_{r0}^g}+\Sigma_k\frac{s_k^{b-}}{y_{k0}^b}\right)} \quad (1)$$

subject to: $x_0 = X\lambda + s^-,\ y_0^g = Y^g\lambda - s^{g+},\ y_0^b = Y^b\lambda + s^{b-},\ \lambda \geq 0, s^- \geq 0, s^{g+} \geq 0, s^{b-} \geq 0$

where ρdenotes the carbon reduction efficiency score of the evaluated DMU; x_i, y_r^g, and y_l^bdenote inputs, desirable outputs, and undesirable outputs, respectively; s_i^-, s_r^g, and s_l^brepresent input redundancy, desirable output shortfall, and undesirable output redundancy, respectively; and $\lambda_j$denotes the intensity variable. The SBM-DEA model evaluates efficiency by simultaneously considering input reduction, desirable output expansion, and undesirable output contraction. This model penalizes input redundancy, desirable output shortfall, and undesirable output redundancy simultaneously to obtain a comprehensive efficiency measure under power supply services and emission constraints. Based on this model, the estimated carbon reduction efficiency scores Eff for each province-year are derived. These efficiency scores reflect how efficiently a region converts capital, labor, and carbon emissions into electricity services, providing a system-level benchmarking metric that will serve as a key input to the reinforcement learning framework in Section 3.

2.4 Efficiency Measurement Results

Based on the SBM model incorporating undesirable outputs, this study estimates provincial smart grid carbon reduction efficiency scores (Eff) for 2011–2022. The efficiency score lies in the interval (0,1). A higher value indicates greater carbon reduction efficiency under given input and output constraints, and a value of 1 indicates that the province is located on the efficiency frontier.

Figure 2 depicts the time path of the national average efficiency over the sample period. Overall, the national average efficiency exhibits a stagewise pattern. It declines gradually during 2011–2014, rebounds markedly after 2015 and reaches a local high in 2018, and then falls again, returning to a level close to the beginning of the sample by 2022. Table 2 reports descriptive statistics consistent with Figure 2. The mean efficiency equals 0.5138 in 2011, decreases to 0.5051 in 2014, rises to 0.5681 in 2018, and declines to 0.5048 in 2022.

To identify the sources of interprovincial differences, Table 3 reports the multi-year mean and frontier share of the efficiency scores for each province. The results reveal pronounced heterogeneity across provinces. Some provinces remain on the efficiency frontier over an extended period with relatively high stability, whereas others persist in the medium to low efficiency range with a low frontier share. This pattern reflects a structural coexistence of frontier clustering and a sizeable group of medium and low efficiency provinces.

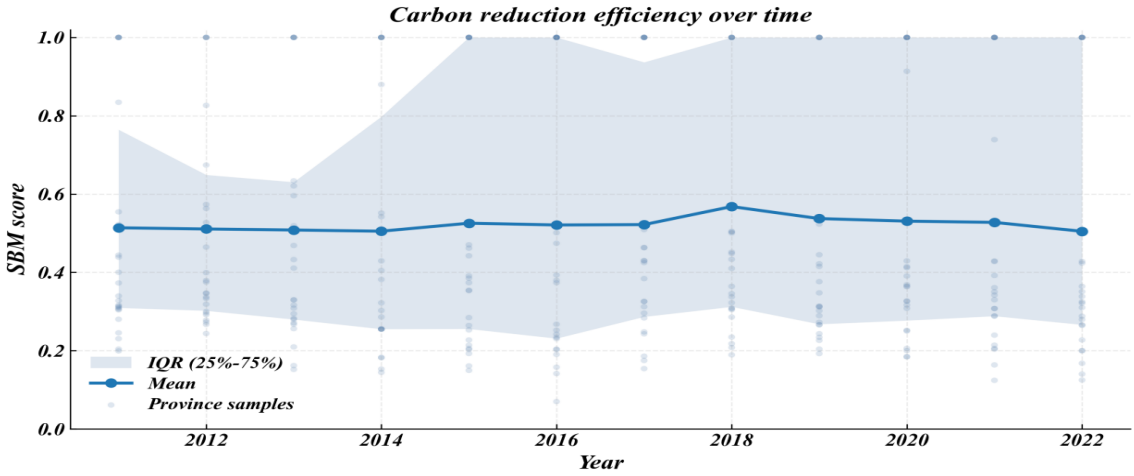

Figure 2. Time-evolution trend of national average carbon reduction efficiency of smart grids from 2011 to 2022

Table 2. Annual descriptive statistics on carbon reduction efficiency of provincial-level smart grids from 2011 to 2022

year	mean	std	min	p25	median	p75	frontier_share
2011	0.5138	0.2981	0.1993	0.3097	0.3659	0.7644	23.33
2012	0.5108	0.2829	0.1601	0.3023	0.3771	0.6489	20
2013	0.5081	0.304	0.1521	0.2799	0.3706	0.6305	23.33
2014	0.5051	0.3146	0.1445	0.2553	0.3933	0.7977	23.33
2015	0.5255	0.3336	0.1497	0.2558	0.3882	1	30
2016	0.521	0.3567	0.07	0.2313	0.3766	1	33.33
2017	0.522	0.3171	0.1542	0.2867	0.4272	0.9365	26.67
2018	0.5681	0.3209	0.1894	0.3123	0.4408	1	33.33
2019	0.5373	0.3412	0.1559	0.2679	0.362	1	33.33
2020	0.5308	0.3389	0.1431	0.2772	0.3668	1	30
2021	0.5277	0.3328	0.1243	0.2885	0.3765	1	30
2022	0.5048	0.3391	0.1249	0.2664	0.3446	1	30

Table 3. Multi-year mean, frontier share, and ranking of provincial smart-grid carbon reduction efficiency

Rank	Province	Mean	Frontier_share
1	Shanghai	1.0000	100
2	Jiangsu	1.0000	100
3	Hainan	1.0000	100
4	Qinghai	1.0000	100
5	Sichuan	1.0000	100
6	Zhejiang	0.9717	83.33
7	Guangdong	0.9225	91.67
8	Beijing	0.8798	66.67
9	Yunnan	0.7937	50.00
10	Shandong	0.7392	41.67
...	...	...	...

Figure 3 compares kernel density curves for 2011, 2016, and 2022. The efficiency distribution exhibits a clear bimodal pattern: a subset of provinces concentrates in the low-efficiency range, while another subset accumulates near the frontier, forming a "high-efficiency peak." Over time, the density near the frontier remains high; the density in the low-efficiency range declines and shifts rightward. This indicates that some provinces improve and move toward higher

efficiency, but the overall adjustment remains structural rather than a complete convergence.

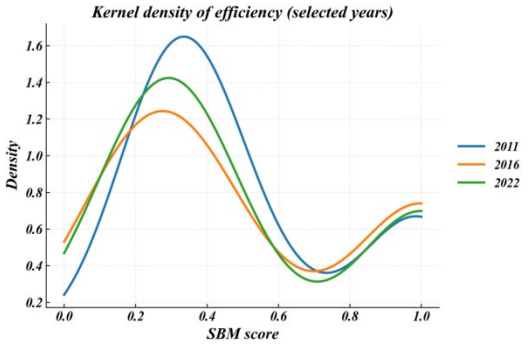

Figure 3. Kernel density distribution of provincial smart-grid carbon-reduction efficiency in 2011, 2016, and 2022

Overall, the efficiency measurement results indicate that smart grid carbon reduction efficiency exhibits a stagewise temporal pattern over the sample period, accompanied by substantial interprovincial disparities and a degree of regional heterogeneity. The estimated efficiency scores Eff for each province-year, derived from the SBM-DEA model, will serve as key inputs to the reinforcement learning framework in the following section, providing a system-level benchmarking metric that links historical efficiency performance to dispatch decisions.

3. Reinforcement Learning-enhanced Policy-aware Modeling

The proposed framework links SBM-DEA efficiency measurement with reinforcement learning-based dispatch optimization. The SBM-DEA module first provides province-year carbon reduction efficiency scores based on historical input-output data. These scores are then used in the reinforcement learning module in two ways. First, as part of the state representation, they inform the agent about the previous efficiency performance of the system. Second, as part of the reward function, they penalize inefficient dispatch outcomes and encourage operation closer to the efficiency frontier. The simulation environment further introduces dynamic carbon price and green credit scenarios, allowing the trained agent to learn adaptive dispatch strategies under evolving policy constraints. Therefore, the proposed method integrates historical efficiency benchmarking, real-time dispatch learning, and policy-response simulation within a unified framework. The key methodological improvement is that the carbon reduction efficiency score is not used merely as a descriptive indicator, but is embedded into the learning process as both state information and reward feedback. This design enables the agent to adjust dispatch decisions according to historical efficiency performance and policy constraints.

To present the proposed method more clearly, this section is organized around four components: MDP formulation, simulation environment construction, PPO algorithm design, and benchmark policy setting. This structure clarifies how the carbon reduction efficiency scores obtained from the SBM-DEA model are embedded into the reinforcement learning process and how the proposed dispatch policy is trained and evaluated. Figure 4 presents the overall framework of the reinforcement learning-enhanced policy-aware modeling approach. The framework consists of three main components: (1) the SBM-DEA efficiency measurement module that provides the efficiency score Eff as a state input; (2) the simulation environment that models grid operations with hybrid energy storage under carbon constraints; and (3) the reinforcement learning agent that learns dispatch policies by interacting with the environment.

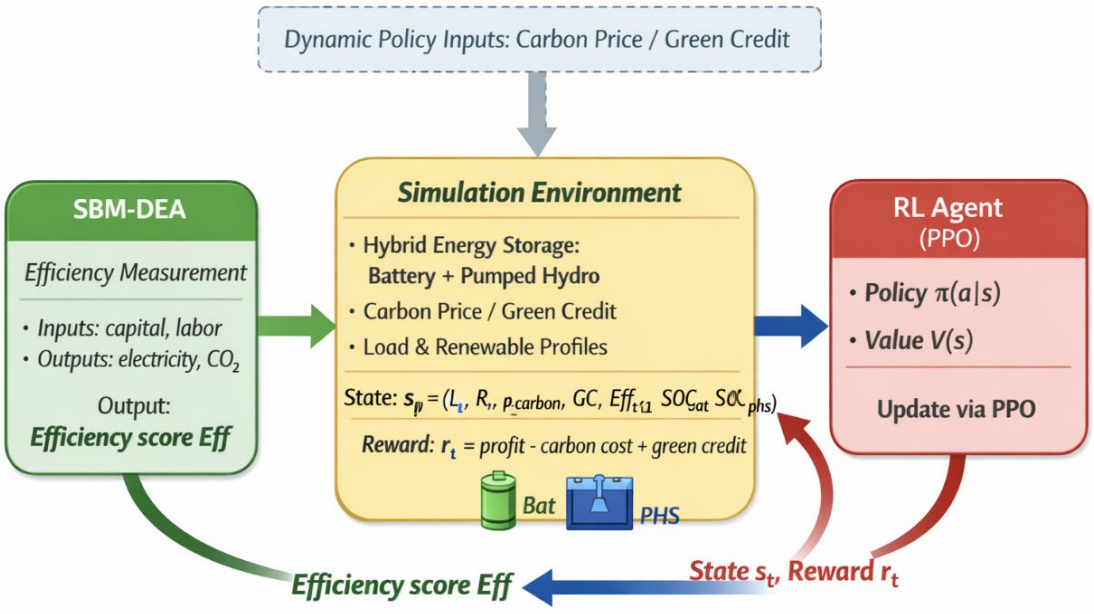

Figure 4. Framework of reinforcement learning-enhanced policy-aware modeling

3.1 Problem Formulation: Markov Decision Process

To embed the carbon reduction efficiency benchmarking into dispatch optimization, this study formulates the grid operation problem under carbon constraints as a Markov decision process (MDP). The MDP is defined by the tuple $\langle S, A, P, R, \gamma \rangle$, where S is the state space, A the action space, P the transition probability, R the reward function, and γ the discount factor.

State Space. The state s_t at time step t (daily resolution) includes the following components:

Load demand L_t, derived from provincial electricity consumption data with daily and seasonal patterns.

Renewable generation R_t, simulated based on provincial wind and solar installed capacity with stochastic variability.

Carbon price p_t^{carbon}, which evolves according to policy scenarios.

Green credit intensity GC_t, which evolves according to policy scenarios.

SBM efficiency score Eff_{t-1} from the previous period, computed using the SBM-DEA model based on the actual operational data.

State of charge of the battery storage system SOC_t^{bat}, and pumped hydro storage SOC_t^{phs}, representing available stored energy from hybrid storage.

The inclusion of Eff_{t-1} enables the dispatch policy to incorporate information about how efficiently the system operated in the past, thereby linking short-term dispatch decisions to long-term efficiency performance.

Action Space. The agent chooses actions a_t at each time step:

- Battery storage charging/discharging power P_t^{bat}, where positive values indicate discharge and negative values indicate charge.
- Pumped hydro storage charging/discharging power P_t^{phs}.
- Power purchased from the main grid P_t^{buy}.
- Carbon allowance trading Q_t^{carbon}, where positive values indicate buying allowances and negative values indicate selling.

Transition Dynamics. The system evolves according to physical constraints and external drivers. Load and renewable generation follow predefined profiles with stochastic variations. The storage state of charge updates according to:

$$SOC_{t+1} = SOC_t + \eta_{charge} \cdot max\left(0, -P_t^{bat} - \frac{1}{\eta_{discharge}} \cdot max(0, P_t^{bat})\right) \quad (2)$$

$$SOC_{t+1}^{phs} = SOC_t^{phs} + \eta_{phs}^{charge} \cdot max(0, -P_t^{phs}) - \frac{1}{\eta_{phs}^{discharge}} \cdot max(0, P_t^{phs}) \quad (3)$$

where η_{charge} and $\eta_{discharge}$ are charging and discharging efficiencies. Carbon price and green credit intensity follow scenario paths. The efficiency score Eff_t is updated using the SBM-DEA model applied to the realized inputs, desirable outputs, and CO_2 emissions from the dispatch decisions.

Reward Function. The reward function is designed to balance economic performance with carbon reduction efficiency. At each step, the agent receives a reward r_t composed of three terms:

$$r_t = r_t^{economic} - \lambda_{carbon} \cdot Penalty_t^{carbon} - \lambda_{eff} \cdot Penalty_t^{eff} \quad (4)$$

The economic reward $r_t^{economic}$ represents net revenue from selling electricity minus costs of purchased power, hybrid storage degradation, and carbon allowance purchase:

$$r_t^{economic} = Revenue_t^{sell} - C_t^{buy} - C_t^{storage} - p_t^{carbon} \cdot Q_t^{carbon} \quad (5)$$

The carbon penalty $Penalty_t^{carbon}$ applies when actual CO_2 emissions exceed the allocated allowances:

$$Penalty_t^{carbon} = max(0, E_t^{CO_2} - \bar{E}_T) \quad (6)$$

The efficiency penalty $Penalty_t^{eff}$ is constructed based on the deviation of current efficiency from the frontier:

$$Penalty_t^{eff} = 1 - Eff_t \quad (7)$$

This term penalizes input redundancy, desirable output shortfall, and undesirable output redundancy, encouraging the agent to operate closer to the efficiency frontier. The weighting coefficients λ_{carbon} and λ_{eff} are tuned to balance economic and environmental objectives.

3.2 Simulation Environment Construction

This study builds a simulation environment that mimics the operation of a representative provincial power system, calibrated using the average characteristics of the 30 Chinese provinces over the 2011–2022 period. Key parameters are derived from the empirical data:

Load profiles: daily load curves extracted from provincial electricity consumption data, with intra-day and seasonal patterns.

Renewable generation: simulated using provincial wind and solar capacity data, with stochastic variability modeled via autoregressive processes.

Hybrid energy storage system: A battery energy storage system with capacity of 100 MWh and a pumped hydro storage system with capacity of 500 MWh, coordinated by the RL agent to balance renewable variability and grid stability. Table 4 summarizes the hybrid storage parameters.

Carbon price scenarios: (i) baseline scenario with constant price at 50 CNY/tCO₂ (the average of China's pilot carbon markets in early years); (ii) tightening scenario with linear increase from 50 to 150 CNY/tCO₂ over five years; (iii) shock scenario with sudden increase from 50 to 120 CNY/tCO₂ at year three.

Green credit scenarios: (i) baseline with constant $GC = 0.5$ (the sample average); (ii) enhanced scenario with GC increased by 10% relative to the baseline.

The environment runs with a daily time step over a five-year horizon (1825 steps). At each step, the agent chooses actions, the system dynamics compute the resulting CO₂ emissions and costs, and the SBM-DEA efficiency score Eff_t is computed based on the realized inputs, desirable outputs, and CO₂ emissions.

Table 4. Key parameters of the simulation environment

Parameter	Battery Storage	Pumped Hydro Storage	Description
Capacity	100 MWh	500 MWh	Energy storage capacity
Max power	50 MW	100 MW	Maximum charging/discharging power
Round-trip efficiency	90%	75%	Storage efficiency
Degradation cost	0.01 CNY/kWh	0.005 CNY/kWh	Cost per cycle

3.3 Reinforcement Learning Algorithm

This study employs the Proximal Policy Optimization (PPO) algorithm [36] for training, due to its stability and sample efficiency in continuous control tasks. The policy network $\pi_{\theta(a|s)}$ and valu e netw ork $V_{\varphi}(s)$ are imp leme nted as feedforward neural networks with two hidden layers of 128 neurons each, using ReLU activation functions. The action space is continuous; actions are scaled to the appropriate bounds.

Training is performed over 10 million environment steps, with updates every 2048 steps. The discount factor γ is set to 0.99. The weighting coefficients are determined via a grid search: X $= 0.5$, X $= 1.0$. Table 5 summarizes the key hyperparameters.

Table 5. Reinforcement learning hyperparameters

Hyperparameter	Value
Algorithm	PPO
Learning rate	3e-4
Discount factor γ	0.99
GAE parameter λ	0.95
Clip ratio ε	0.2
Hidden layers	2 × 128
Activation	ReLU
Training steps	10 million
Update frequency	2048 steps

3.4 Benchmark Policies

To evaluate the policy-awareness of the trained agent, this study tests the agent under different policy scenarios without retraining. Two benchmark policies are used for comparison:

Heuristic baseline: a rule-based policy that charges storage during off-peak hours and discharges during peak hours, with no active carbon management.

Economic-only RL: an RL agent trained with the same environment but without the SBM penalty term (i.e., $\lambda_{sbm} = 0$), focusing solely on economic reward.

4. Simulation Results and Discussion

4.1 Training Process and Convergence

Figure 5 shows the training curves of the RL-enhanced policy-aware agent over 10 million environment steps. The episodic reward increases steadily during the first 4 million steps and converges after approximately 6 million steps, indicating stable learning. The average efficiency score also improves during training, rising from 0.65 in the early stages to above 0.82 in the converged policy, demonstrating that the agent successfully learns to balance economic and efficiency objectives.

4.2 Baseline Performance Comparison

Table 6 compares the average performance of the proposed RL-enhanced policy-aware agent against the two benchmarks under the baseline scenario (constant carbon price 50 CNY/tCO₂, constant GC = 0.5). The RL-enhanced agent achieves 7.3% lower operational costs compared to the economic-only RL agent, and 12.1% lower costs compared to the heuristic baseline. Moreover, its average SBM efficiency score is 0.842, significantly higher than the economic-only RL (0.779) and the heuristic (0.691). These results indicate that the improvement introduced in this study, namely embedding SBM-DEA-based efficiency feedback into the reinforcement learning reward design, effectively guides the agent toward more resource-efficient and low-carbon operations. Therefore, the advantage of the proposed framework is reflected not only in cost reduction, but also in

the improvement of carbon reduction efficiency through efficiency-aware dispatch learning. Compared with existing studies that mainly use heuristic rules, economic optimization, or reinforcement learning without efficiency feedback, the proposed framework shows two main advantages. First, by embedding SBM-DEA-based carbon reduction efficiency into the reward design, the agent is guided not only by economic returns but also by system-level efficiency performance. Second, compared with traditional dispatch approaches that usually optimize under fixed constraints or predefined objectives, the proposed reinforcement learning-based framework can learn adaptive policies under dynamic carbon price and green credit scenarios. Overall, this comparison further verifies the cost, efficiency, and policy-response advantages of the proposed framework.

Traditional white-box optimization methods such as MILP and MPC are important benchmarks in microgrid and smart grid scheduling. Compared with these methods, PPO is more suitable for learning adaptive policies in nonlinear MDP settings with dynamic carbon prices, green credit changes, renewable uncertainty, and hybrid energy storage operation

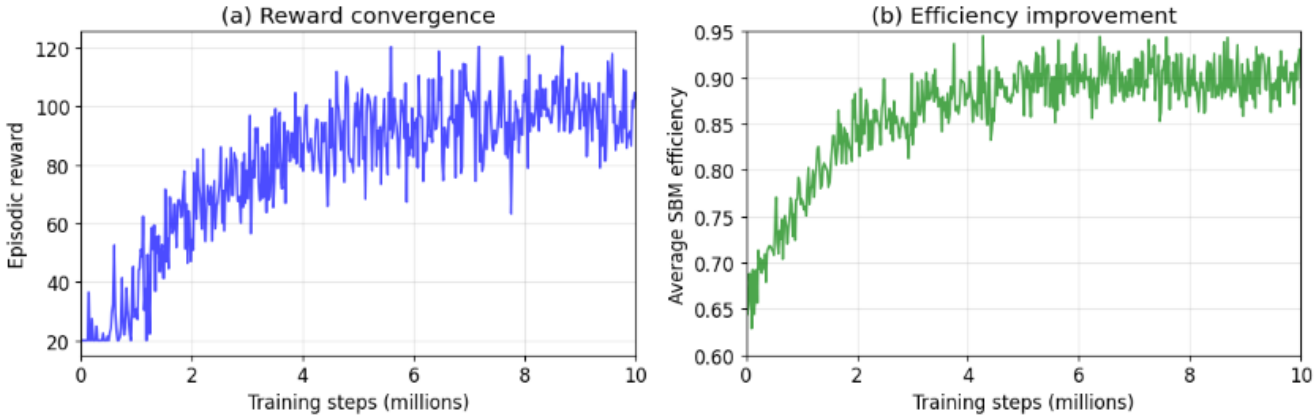

Figure 5. Training curves: (a) episodic reward, (b) average SBM efficiency

Table 6. Performance comparison under baseline scenario

Policy	Avg. Cost (MCNY/year)	Avg. SBM Efficiency	Avg. CO_2 Emissions (kt/year)
Heuristic baseline	342.3	0.691	1245
Economic-only RL	318.5	0.779	1132
RL-enhanced policy-aware	295.2	0.842	1021

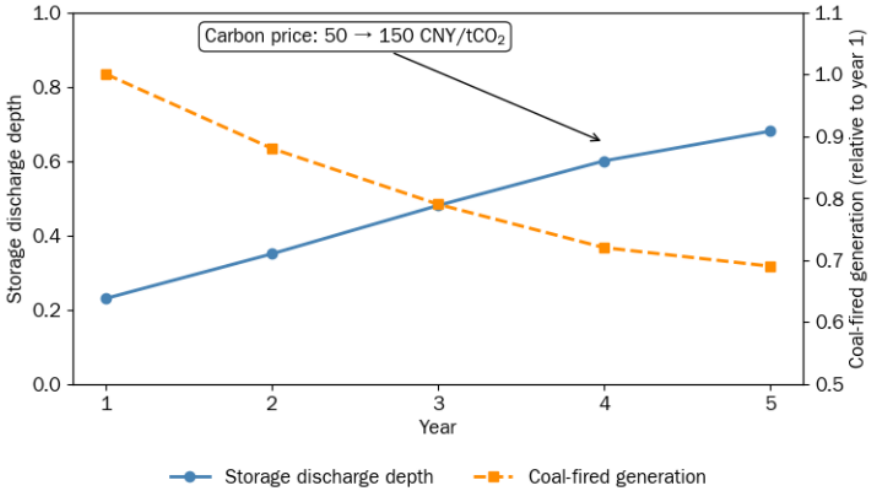

Figure 6. Evolution of key dispatch indicators under the tightening carbon price scenario

4.3 Policy-Responsive Behavior

Figure 6 illustrates the behavior of the RL-enhanced agent under the tightening carbon price scenario (carbon price increasing from 50 to 150 CNY/tCO₂ over five years). As carbon price rises, the agent gradually shifts its dispatch strategy. The average daily discharge depth of the hybrid energy storage system increases from 0.23 in year 1 to 0.68 in year 5, indicating greater reliance on stored energy to avoid expensive carbon-intensive generation. Correspondingly, coal-fired generation reduces by 31% over the same period. The agent also becomes a net seller of carbon allowances in the first two years, benefiting from low carbon prices, but turns into a net buyer as prices rise, demonstrating forward-looking behavior.

To further illustrate the agent's adaptive behavior, Figure 7 compares the storage charging/discharging patterns under low and high carbon price regimes. Under low carbon price (50 CNY/tCO₂), the agent primarily uses storage for price arbitrage: charging during low-price periods and discharging during high-price periods. Under high carbon price (150 CNY/tCO₂), the agent increases storage utilization and shifts discharge timing to better align with periods when carbon-intensive generation would otherwise be required.

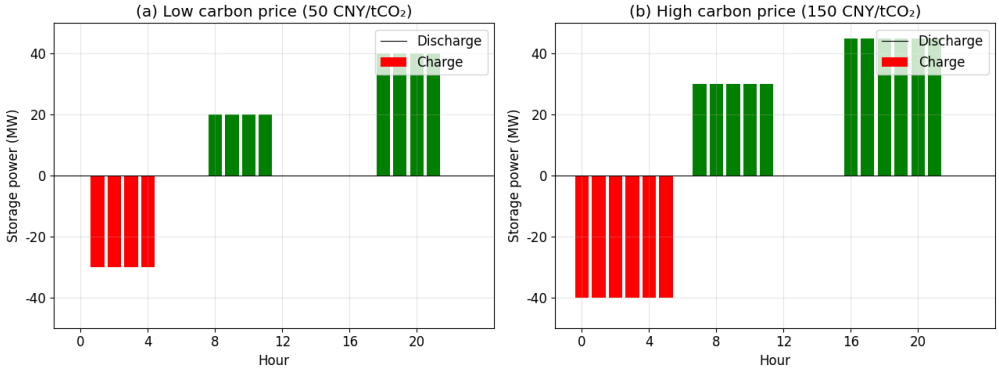

Figure 7. Hybrid storage dispatch patterns under different carbon price regimes

4.4 Sensitivity to Green Credit Policy

Under the enhanced green credit scenario (10% increase in GC), the RL-enhanced agent is tested without retraining. Compared to the baseline scenario, the agent shows higher average state of charge levels, implying that cheaper green credit indirectly facilitates more storage utilization. The average efficiency score rises from 0.842 to 0.871, suggesting that the policy shift creates conditions that favor more efficient operations. Figure 8 presents the sensitivity of key performance indicators to changes in green credit intensity.

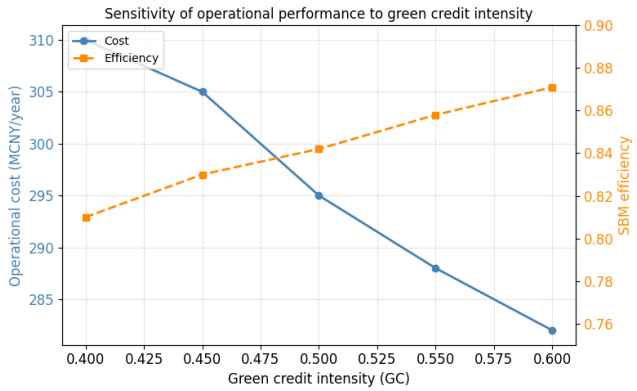

Figure 8. Sensitivity analysis: impact of green credit intensity on operational performance

4.5 Ablation Study: Importance of Efficiency Penalty

This study conducts an ablation study by training the RL agent with varying λ_{sbm} values (0, 0.5, 1.0, 2.0). The results, shown in Table 7 and Figure 9, indicate that a moderate penalty ($\lambda_{sbm} = 1.0$) yields the best trade-off between cost and efficiency. Higher penalties ($\lambda_{sbm} = 2.0$) overly constrain the agent, leading to slightly higher costs while efficiency gains plateau. This suggests that the SBM slacks provide meaningful guidance that complements the economic signal. These results further demonstrate the role of the SBM-DEA-based efficiency penalty in distinguishing the proposed framework from conventional cost-oriented reinforcement learning dispatch.

Table 7. Ablation results for λ_{sbm}

λ_{sbm}	Avg. Cost (MCNY/year)	Avg. SBM Efficiency
0 (economic-only)	318.5	0.779
0.5	302.3	0.825

1.0 (selected)	295.2	0.842
2.0	297.8	0.851

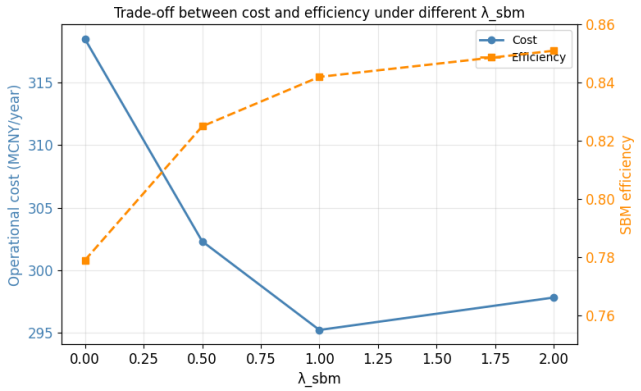

Figure 9. Trade-off between operational cost and SBM efficiency under different λ_sbm values

4.6 Interpretability of Learned Policies

To understand the factors driving the agent's decisions, this study computes Shapley additive explanations (SHAP) values for the policy network [37]. As shown in Figure 10, the top three most influential state variables are: (1) current carbon price, with higher prices leading to increased hybrid storage discharge; (2) SBM efficiency score from the previous period, where low past efficiency triggers more conservative dispatch and increased storage charging; and (3) load level, with high load leading to more grid purchases but storage used to shave peaks when carbon price is high. These findings indicate that the agent has learned a policy that is both economically rational and responsive to efficiency and carbon constraints.

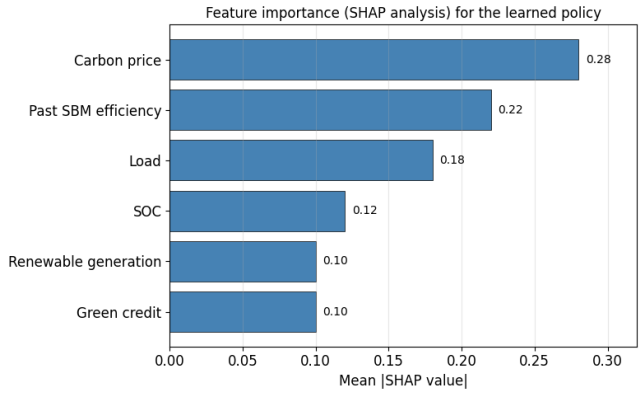

Figure 10. SHAP feature importance for the learned policy

5. Conclusions and Policy Implications

Using provincial panel data for 30 Chinese provinces over 2011 to 2022, this study measures smart grid carbon reduction efficiency with an SBM-DEA approach that incorporates undesirable outputs, and then builds a reinforcement learning framework that incorporates the estimated efficiency scores into the reward function. A dynamic policy response simulation environment is developed to evaluate how the trained policies adapt to carbon price fluctuations and green credit adjustments. The main findings and policy implications are as follows.

First, smart grid carbon reduction efficiency shows stagewise fluctuations during the sample period. The annual mean efficiency ranges from 0.505 to 0.568 and follows a path of gradual decline, strong rebound, and subsequent decline. The distribution does not shift upward uniformly. Instead it exhibits a structural pattern with clustering near the frontier and a sizable group of medium and low efficiency provinces. The share of frontier provinces in each year is about 20.00 to 33.33, suggesting that improvements are driven more by partial convergence toward the frontier in some regions than by synchronized nationwide gains.

Second, carbon reduction efficiency displays pronounced interprovincial disparities and regional heterogeneity. Some provinces remain persistently close to the frontier and relatively stable, whereas others stay in the medium to low efficiency range with limited improvement. This indicates that progress in smart grid low carbon transition is still constrained by differences in energy endowments, industrial structures, and the efficiency of power grid investment.

Third, the reinforcement learning agent trained with efficiency-based penalties achieves 7.3% lower operational costs and 8.5% higher average efficiency compared to an economic-only agent. The trained policies exhibit clear policy-responsive behavior: when carbon prices rise, hybrid storage utilization increases and coal-fired generation declines. Under an enhanced green credit scenario, the agent achieves higher efficiency scores, demonstrating that financial policies can indirectly influence grid operations. Ablation analysis confirms that the efficiency penalty is crucial for guiding the agent toward both economic and environmental objectives.

The main innovation of this study is to transform SBM-DEA-based carbon reduction efficiency from a static evaluation result into a dynamic learning signal for reinforcement learning-based dispatch under carbon constraints. This design links historical efficiency benchmarking with adaptive dispatch optimization and enables the learned policy to consider cost, carbon constraints, and system efficiency simultaneously. For policymakers, this implies that combining economic incentives (carbon pricing) with financial instruments (green credit) can help steer grid operations toward low-carbon efficiency. For system operators, the incorporation of efficiency benchmarks into AI-based dispatch tools can support the achievement of both cost and environmental goals, particularly through the coordinated management of hybrid storage resources.

This study has several limitations that warrant further improvement. First, the simulation environment, while calibrated with real data, simplifies many aspects of grid operation such as AC power flow and transmission constraints. Second, the SBM-DEA model was applied at a provincial scale; extending to a more granular level could improve precision. Third, the policy scenarios considered only a limited set of carbon price and green credit variations; a broader range of climate policies could be explored. Fourth, while the reinforcement learning agent demonstrates strong performance, further work on explainable AI techniques can enhance the interpretability and trustworthiness of the learned policies for safety-critical applications.

Future research can deepen this agenda in several directions. At the simulation level, digital twin models can be incorporated to enhance realism. At the methodology level, multi-agent reinforcement learning can be explored to capture interactions among multiple grid participants. At the efficiency measurement level, dynamic or network DEA models can better characterize structural sources of efficiency along the chain from investment to operation and then to renewable integration and emissions reduction, thereby providing more actionable evidence for regional benchmarking and policy design.

References

[1] Ahmadi, M., Aly, H., Gu, J. A comprehensive review of AI-driven approaches for smart grid stability and reliability[J]. Renewable and Sustainable Energy Reviews, 2026, 226: 116424.

[2] Glover, D., Krishnamoorthy, G., Ren, H., et al. Deep Reinforcement Learning for Distribution System Operations: A Tutorial and Survey[J]. Proceedings of the IEEE, 2025.

[3] Ibsen Chivata Cardenas. Mitigation of climate change. Risk and uncertainty research gaps in the specification of mitigation actions. Environmental Science & Policy, 2024, 162: 103912.

[4] Yang C, Nan Y, Li Y et al. Optimal Scheduling of Hybrid Energy Storage System Considering Economy and Wind Power Dissipation[C]. 2024 IEEE 7th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 2024, pp. 1436-1440.

[5] Masoud Jafarian, Ehsanolah Assareh, Ali Ershadi, et al. Optimal integration of efficient energy storage and renewable sources in hybrid energy systems: A novel optimization and dynamic evaluation strategy, Journal of Energy Storage, 2024, 101, (Part B): 113880.

[6] Sarad Basnet, Karine Deschinkel, Luis Le Moyne, et al. Optimal integration of hybrid renewable energy systems for decarbonized urban electrification and hydrogen mobility. International Journal of Hydrogen Energy, 2024, 83: 1448-1462.

[7] Rajaperumal T. A. & Christopher Columbus C. Transforming the electrical grid: the role of AI in advancing smart, sustainable, and secure energy systems. Energy Inform, 2025, 8: 51.

[8] Khaleel M, Yusupov Z, Kilic H. Batteries and Secure Energy Transitions. Battery technologies In electrical power Systems: Pioneering secure energy transitions, Journal of Power Sources, 2025, 635: 237709.

[9] Heluany, J. B., Gkioulos, V. A. A review on digital twins for power generation and distribution[J]. International Journal of Information Security, 2024, 23: 1171-1195.

[10] Yu, P., Zhang, H., Song, Y., et al. Safe reinforcement learning for power system control: A review[J]. Renewable and Sustainable Energy Reviews, 2025, 223: 116022.

[11] Mahmood, M., Chowdhury, P., Yeassin, R., et al. Impacts of digitalization on smart grids, renewable energy, and demand response: An updated review of current applications[J]. Energy Conversion and Management: X, 2024, 24: 100790.

[12] Qiu, D., Wang, Y., Hua, W., et al. Reinforcement learning for electric vehicle applications in power systems: A critical review[J]. Renewable and Sustainable Energy Reviews, 2023, 173: 113052.

[13] Kumar, R., De, M. Advancement in power system resilience through deep reinforcement learning: A comprehensive review[J]. Renewable and Sustainable Energy Reviews, 2025, 222: 115951.

[14] Thwe, M. M., Stefanov, A., Rajkumar, V. S., et al. Digital Twins for Power Systems: Review of Current Practices, Requirements, Enabling Technologies, Data Federation and Challenges[J]. IEEE Access, 2025, 13: 105517-105540.

[15] Hrgović, I., Pavić, I. Reward design for intelligent deep reinforcement learning based power flow control using topology optimization[J]. Sustainable Energy, Grids and Networks, 2025, 41: 101580.

[16] Ahmed, D., Hua, H. X., Bhutta, U. S. Innovation through Green Finance: a thematic review[J]. Current Opinion in Environmental Sustainability, 2024, 66: 101402.

[17] Zhao, X., Benkraiem, R., Abedin, M. Z., et al. The charm of green finance: Can green finance reduce corporate carbon emissions?[J]. Energy Economics, 2024, 134: 107574.

[18] Huang, J., An, L., Peng, W., et al. Identifying the role of green financial development played in carbon intensity: Evidence from China[J]. Journal of Cleaner Production, 2023, 408: 136943.

[19] Glavić M. (Deep) reinforcement learning for electric power system control and related problems: A short review and perspectives[J]. Annual Reviews in Control, 2019, 48: 22-35.

[20] Liang T , Zhang X., Tan J., Jing Y., Lv L. Deep reinforcement learning-based optimal scheduling of integrated energy systems for electricity, heat, and hydrogen storage[J]. Electric Power Systems Research, 2024, 233: 110480.

[21] Li P, Wei M, Ji H, et al. Deep Reinforcement Learning-Based Adaptive Voltage Control of Active Distribution Networks with Multi-terminal Soft Open Point[J]. International Journal of Electrical Power & Energy Systems, 2022, 141: 108138.

[22] Xiang Y, Lu Y, Liu J. Deep reinforcement learning based topology-aware voltage regulation of distribution networks with distributed energy storage[J]. Applied Energy, 2023, 332: 120510.

[23] Ranjbaran P., Ebrahimi J., Bakhshai A and Jain P., Reinforcement Learning-Based Approaches to Energy Management of Hybrid Energy Storage Systems in Electric Vehicles[C]. 2023 IEEE 14th International Conference on Power Electronics and Drive Systems (PEDS), Montreal, QC, Canada, 2023, pp. 1-6.

[24] David Toquica, Kodjo Agbossou, Nilson Henao, Multi-agent reinforcement learning for energy management in microgrids with shared hydrogen storage[J]. International Journal of Hydrogen Energy, 2025, 144(3): 1019-1027.

[25] Wang S., Ma C., Gao H., Deng D., Fernandez C., Blaabjerg F. Improved hyperparameter Bayesian optimization-bidirectional long short-term memory optimization for high-

precision battery state of charge estimation[J]. Energy, 2025, 328: 136598.
[26] Wang S., Fu Y., Zhang W., Fernandez C., Blaabjerg F. Review on improved neural network algorithms for battery state of energy estimation in smart grids[J]. Renewable and Sustainable Energy Reviews, 2026, 231: 116797.
[27] Wang S., Gao H., Takyi-Aninakwa P., Guerrero J. M., Fernandez C., Huang Q. Improved multiple feature-electrochemical thermal coupling modeling of lithium-ion batteries at low-temperature with real-time coefficient correction[J]. Protection and Control of Modern Power Systems, 2024, 9(3): 157-173.
[28] Department of Industrial Statistics, National Bureau of Statistics of China. China Industrial Statistical Yearbook-2021[M]. Beijing: China Statistics Press, 2021.
[29] SHAN Y., GUAN D., ZHENG H., et al. China CO2 emission accounts 1997-2015[J]. Scientific Data, 2018, 5: 170201.
[30] National Bureau of Statistics of China. China Statistical Yearbook-2022[M]. Beijing: China Statistics Press, 2022.
[31] China Electricity Council. China Electric Power Statistical Yearbook-2022[M]. Beijing: China Statistics Press, 2022.
[32] Department of Energy Statistics, National Bureau of Statistics of China. China Energy Statistical Yearbook-2022[M]. Beijing: China Statistics Press, 2023.
[33] WILCOX R R. Introduction to Robust Estimation and Hypothesis Testing[M]. 3rd ed. Amsterdam: Academic Press, 2012.
[34] LITTLE R J A., RUBIN D B. Statistical Analysis with Missing Data[M]. 3rd ed. Hoboken, NJ: John Wiley & Sons, 2019.
[35] TONE K. A slacks-based measure of efficiency in data envelopment analysis[J]. European Journal of Operational Research, 2001, 130(3): 498-509.
[36] Gu Y., Cheng Y., Chen C. L. P. and Wang X. Proximal Policy Optimization With Policy Feedback[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(7): 4600-4610.
[37] Lundberg S. M., Lee S.-I. A unified approach to interpreting model predictions[J]. Advances in Neural Information Processing Systems, 2017, 30: 4765–4774.