ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Deep Reinforcement Learning for Uncertainty-Aware Dispatch Optimization in Power Systems

Shuo Yu[1], *Jingbo Wang*[1,*], *Qiang Li*[1], *Rui Yang*[2] *and Hongyu Tang*[2]

[1] Inner Mongolia Power (Group) Co., Ltd., Saihan District, Hohhot 010020, Inner Mongolia, China

[2] Beijing Tsintergy Technology Co., Ltd., Haidian District, Beijing 100084, China

SUMMARY: *This paper proposes a deep reinforcement learning method for uncertainty aware scheduling to address the problem of power system scheduling decisions being susceptible to prediction bias, related disturbances, and extreme scenarios under high proportion wind and photovoltaic power integration conditions. Firstly, construct a scheduling environment that includes joint error characterization of wind power, photovoltaic power, and load, and explicitly embed multi-source related deviations into the state space. Secondly, design a Soft Actor Critic (SAC) scheduler that integrates risk sensitive rewards and safety action mapping layers to achieve coordinated optimization between operating costs, wind and solar power curtailment, carbon emissions, insufficient backup, and constraint violations. Based on publicly available time series data and combined with typical days, extreme disturbances, and sample scenarios outside the training set for validation. The results showed that the total operating cost of the proposed method was 52.47 × 10^4 CNY/day, a decrease of 3.39% compared to the original SAC method and a decrease of 7.75% compared to Model Predictive Control (MPC). And the wind and solar abandonment rate is 3.79%, the constraint violation rate is only 0.21%, and the average single step solving time is 0.045 seconds. At the same time, this method shows better stability and generalization ability under high uncertainty, cross month testing, and extreme weather conditions. Research has shown that this method can provide intelligent decision support with engineering feasibility for online scheduling of new power systems.*

KEYWORDS: *deep reinforcement learning; uncertainty-aware dispatch; risk-sensitive optimization; safety action mapping; power system operation optimization*

1 Introduction

Against the backdrop of the "dual carbon" goal and the continuous promotion of the construction of new power systems, the installed capacity of renewable energy such as wind power and photovoltaics has grown rapidly, and the operation mode of the power system has also undergone profound changes. Traditional scheduling is more based on the premise that controllable power sources dominate and system fluctuations are relatively limited. Its core task is usually to solve deterministic economic scheduling problems around load demand and unit constraints. However, when a high proportion of wind and photovoltaic power is connected, power fluctuations, prediction deviations, time period coupling, and multi-source uncertainties work together, and scheduling decisions are no longer static optimization at a

*etxpublic01@163.com
https://doi.org/10.65102/is2026804
Published: 30 April, 2026

single moment, but closer to a sequential decision-making process that requires continuous perception of the environment, dynamic correction of actions, and real-time response feedback. In recent years, Deep Reinforcement Learning (DRL) has demonstrated strong adaptability in high-dimensional nonlinear control problems through a closed-loop mechanism of "state observation policy output environment feedback policy update", providing a new path for complex, continuous, and strongly time-varying scheduling tasks in power systems that differs from traditional optimization methods [1]. At the same time, the application of reinforcement learning in power systems is gradually expanding from device control to scenarios such as planning, operation scheduling, and multi-stage decision-making. This indicates that it can not only serve as a local control tool, but also has the potential to support more complex system decisions.

However, applying DRL to system level scheduling with uncertainty awareness still faces significant theoretical and engineering challenges. Existing research has validated the feasibility of learning scheduling strategies in multiple scenarios. For example, robust federated DRL for collaborative control of multiple virtual power plants shows that the learning method has certain adaptability in complex distributed environments [3]. The research on short-term optimization scheduling of water wind solar complementary systems also shows that reinforcement learning can effectively handle multi energy complementarity and dynamic operation processes [4]. In addition, the economic dispatch of integrated energy systems under low-carbon orientation further demonstrates that reinforcement learning frameworks can naturally accommodate issues such as multi-objective trade-offs, carbon costs, and coupled energy flows [5]. In the dispatch of power grids containing renewable energy, relevant methods also demonstrate certain online decision-making potential and real-time optimization value [6]. However, overall, these studies are more focused on smaller boundaries such as microgrids, parks, virtual power plants, or local multi energy systems. The research objects are relatively closed, and the network size and operational constraint complexity are limited. For the "network level, rolling, continuous time domain" scheduling problems that are closer to the actual power system, especially the uncertainty aware dispatch under the joint action of wind power, photovoltaic power, and load related disturbances, there is still a lack of in-depth and systematic discussions.

Further analysis of existing work reveals at least three areas worth advancing. Firstly, although uncertainty has become an important topic in scheduling research in recent years, many methods still mainly treat it as independent error terms, single step prediction deviations, or simple scene disturbances. The description of the related structures, resonance changes, and temporal evolution laws between wind power, photovoltaic power, and loads is insufficient, resulting in incomplete understanding of complex disturbance environments by intelligent agents. Secondly, existing research often emphasizes the reduction of operating costs or improvement of convergence performance, but rarely integrates risk measurement, constraint feasibility, and strategy security into the analysis framework. When new energy fluctuations increase, net load jumps rapidly, or extreme periods occur, relying solely on average cost optimization may be difficult to avoid operating beyond limits, insufficient backup, or even strategy failure. Again, many studies still focus on overall cost comparison or training curves in the presentation of results, while paying insufficient attention to extreme scenario response, parameter sensitivity, key module contributions, and interpretability of scheduling trajectories. This makes it difficult for methods, even if they have numerical advantages, to fully demonstrate their robustness and generalization ability in practical operating environments. Recently, research on forward-looking economic dispatch has begun to introduce look ahead mechanisms to enhance the perception and responsiveness of strategies to future system states [7]. The research on temporal feature enhanced scheduling for virtual power plants also

indicates that historical information extraction and temporal pattern modeling are of great significance for the formation of learning strategies [8]. These achievements provide direct inspiration for this article, but there is still room for further expansion in risk sensitive modeling, security constraint embedding, and system level uncertainty awareness.

From an engineering practice perspective, the reason why system level scheduling is more difficult to handle than local energy management is because it has more decision-making objects, stronger time period coupling, and a more rigorous constraint system. On the one hand, there is a significant cross time linkage between conventional unit ramp up, energy storage state of charge, reserve capacity configuration, power balance, and network security boundaries. Local optima at a single time often translate into feasible domain contraction or increased operating costs in subsequent time periods. On the other hand, the output error of new energy does not occur in isolation. It often evolves together with factors such as weather changes, load response, and regional exchange power, causing the system to face a chain effect of continuous amplification of "prediction deviation control action operation feedback". Therefore, the focus of this article is not simply to replace traditional solvers with reinforcement learning, but to attempt a more targeted restatement at the problem modeling level: to extend the traditional static scheduling approach of "prediction first, optimization later" to a dynamic decision-making framework of "perception of uncertainty, assessment of operational risks, and real-time correction of strategies". This also constitutes the basic starting point for the subsequent method design and experimental verification in this article.

Based on the above understanding, this article focuses on the uncertainty aware scheduling problem in the power system and constructs a DRL method that balances economy, feasibility, and robustness. Specifically, this article first establishes an uncertainty aware dispatch environment that considers wind power, photovoltaic, and load related disturbances. Multiple sources of uncertain information are embedded into the decision-making process in the form of joint state or scene features, enabling the agent to not only recognize the current system state, but also to some extent sense the correlation structure and changing trends between disturbances. Secondly, to address the issues of policy boundary crossing and operational risk accumulation in the context of high proportion renewable energy access, a DRL scheduler with risk penalty terms and a security mapping layer is designed. A constraint correction mechanism is added between policy output and physical feasible domain to balance policy learning efficiency and practical operational safety. Again, based on the characteristics of multiple costs coexisting in real scheduling, a multi-objective reward function is constructed that simultaneously considers conventional unit output, energy storage charging and discharging, backup configuration, wind and solar power curtailment, carbon emissions, and constraint violation penalties, in order to avoid the model only pursuing the minimum single cost and neglecting system resilience and safety margin. Finally, comparative experiments were conducted on typical days, extreme days, and different levels of uncertainty to systematically evaluate the comprehensive performance of the proposed method in terms of economy, disturbance rejection, and generalization ability.

The research motivation of this article is closely related to the recent progress in related uncertainty and distribution robust optimization. Previous studies have shown that when wind power uncertainty has significant dependency structures, using independent error assumptions often underestimates system operational risks. However, incorporating relevant structures into opportunity constraints or robust optimization frameworks can more accurately depict the relationship between safety boundaries and scheduling costs [9]. This article absorbs its core idea of "relevant uncertainty cannot be ignored" and further transforms it into key principles in reinforcement learning environment design, reward shaping, and security mechanism construction. Based on this, this article attempts to answer a more practical question, namely

whether DRL can form an uncertainty aware dispatch optimization paradigm that is more suitable for new power systems under the conditions of simultaneous existence of relevant disturbances, rolling decisions, and safety constraints.

2 Methods

2.1 Uncertainty-aware dispatch model and MDP formulation

After the integration of high proportion wind and photovoltaic power, power system scheduling is no longer just based on static output allocation based on point prediction results, but is closer to a rolling control process that continuously adjusts decisions under continuous disturbance driving [10]. In order to enable the scheduling strategy to truly respond to the operational disturbances caused by the high proportion of wind and photovoltaic power integration, this paper describes the system scheduling process as a Markov Decision Process (MDP) with relevant uncertainties [11]. This article also organizes the prediction biases of wind power, photovoltaic power, and load into a joint disturbance vector, allowing the intelligent agent to not only "see" the prediction results during decision-making, but also simultaneously perceive the direction, magnitude, and linkage changes of multi-source biases. After such treatment, the scheduling problem is no longer regarded as a static output allocation, but is restated as a "sequential control process that rolls over time and is continuously driven by uncertainty". The overall framework of uncertainty aware DRL scheduling is shown in Figure 1.

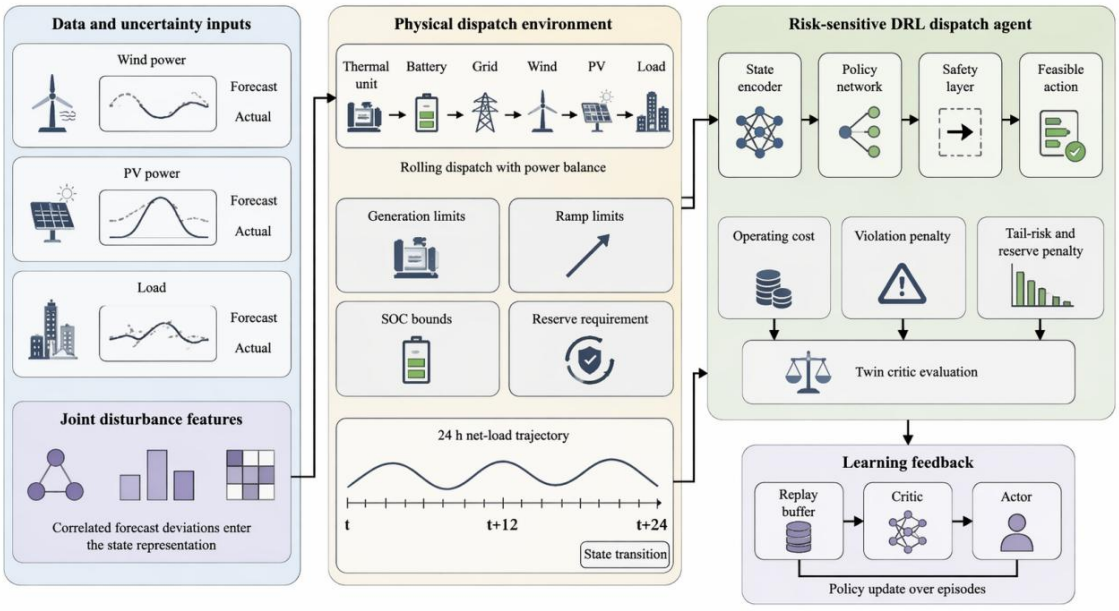


Figure 1: Overall framework of uncertainty-aware DRL dispatch.

In Figure 1, the left module provides historical sequences, actual values, and predicted values of wind power, photovoltaic power, and load. The middle module corresponds to a scheduling environment with power balance, energy storage status, and backup constraints. The right module is a DRL intelligent agent with risk shaping and safety action mapping. The method framework shown in Figure 1 illustrates the three argument points of the proposed method. Firstly, uncertainty does not remain at the point prediction of the input, but is organized into joint perturbation features into the state space. Secondly, the output of the

strategy network is not directly executed instructions, but security actions that require feasible domain correction [13]. Thirdly, training feedback not only comes from operating costs, but is also affected by tail risks, violation costs, and reserve shortages [14]. These three factors together determine the difference between our method and ordinary scheduling based reinforcement learning. The system considered in this article consists of conventional units, wind farms, photovoltaic power stations, energy storage units, and external grid interfaces. At each time period t, the system needs to satisfy the power balance relationship, as shown in formula (1).

$$\sum_{i=1}^{N_g} P_{i,t}^{g}+P_t^{dis}+P_t^{grid}+P_t^{w}+P_t^{pv}=L_t+P_t^{\mathrm{ch}}+P_t^{curt} \tag{1}$$

In Equation (1), $P_{i,t}^{g}$ represents the output of the i-th conventional unit. P_t^{dis} and P_t^{ch} denote the energy storage discharge and charge power, respectively. P_t^{grid} represents the exchange power with the external grid. P_t^{w}, P_t^{pv} and L_t represent the actual values of wind power, PV power, and load, respectively. Ptcurt represents the wind/PV curtailment power. Equation (1) is the constraint core of the entire scheduling environment, and all subsequent action corrections and reward feedback revolve around this balance relationship. The energy storage unit plays a role in smoothing net load fluctuations and providing flexible adjustment capabilities, therefore the energy storage state is expressed recursively across time periods, as shown in formula (2).

$$E_{t+1}=E_t+\eta^{\mathrm{ch}}P_t^{\mathrm{ch}}\Delta t-\frac{P_t^{dis}\Delta t}{\eta^{dis}} \tag{2}$$

In Equation (2), E_{t+1} and E_t represent the state of charge (SOC) of the energy storage at time period t+1 and t, respectively. η^{ch} and η^{dis} denote the charge and discharge efficiencies, respectively. Equation (2) directly links the current action with the scheduling space of the next time period, making policy learning have obvious temporal coupling characteristics. In addition to the two core constraints mentioned above, this article also considers conditions such as the upper and lower limits of conventional unit output, ramp rate, backup lower limit, and the State of Charge (SOC) boundary of energy storage batteries synchronously in the environment [15, 16]. To enhance the ability to perceive uncertainty, this article organizes the prediction errors of wind power, photovoltaic power, and load into a joint disturbance vector ξ t, as shown in formula (3).

$$\xi_t=\begin{bmatrix} e_t^{w} \\ e_t^{pv} \\ e_t^{L} \end{bmatrix}=\begin{bmatrix} P_t^{w}-\widehat{P}_t^{w} \\ P_t^{pv}-\widehat{P}_t^{pv} \\ L_t-\hat{L}_t \end{bmatrix} \tag{3}$$

In Equation (3), e_t^{w}, e_t^{pv} and e_t^{L} represent the prediction errors for wind power, PV power, and load, respectively. $\widehat{P}_t^{w}$, $\widehat{P}_t^{pv}$ and $\hat{L}_t$ represent the forecasted values for wind power, PV power, and load, respectively. The disturbance vector directly enters the state space, enabling the strategy network to learn composite disturbance patterns such as "high load and low photovoltaic", "wind and solar deviation from prediction", and "rapid increase in net load", which are more closely related to the common scenarios of rapid increase in net load or wave superposition in actual scheduling. Considering that errors often exhibit significant correlation in continuous time periods, this paper also embeds the error statistics of the last H time periods into the state to enhance the strategy's perception of local trends and related

structural changes. In terms of state representation, this article comprehensively introduces information such as point prediction values of wind power, photovoltaic power, and load, joint error vectors, energy storage state of charge, available reserve level, and previous period unit output. The resulting state vector contains both forward-looking predictive information and dynamic feedback and device inertia information. Correspondingly, the action space mainly includes conventional unit output adjustment, energy storage charging and discharging power, grid exchange power, backup allocation, and necessary power abandonment control. Through this state action design, the intelligent agent can not only perceive the current power balance relationship, but also identify the scheduling pressure under multi-source disturbance coupling [17].

In summary, this paper formalizes the uncertainty-aware dispatch environment as a quintuple (S,A,P,R,γ). Here, S represents the state space. A represents the continuous action space. P is determined by the realizations of wind, PV, and load along with equipment constraints. Given the current state s_t, the agent outputs an action a_t. The environment completes the state transition $s_t \rightarrow s_{t+1}$ based on the actual realizations of wind, PV, and load and the equipment constraints, and returns an immediate reward r_t. After a dispatch cycle concludes, the cumulative return is used to update the policy parameters.

2.2 Risk-sensitive deep reinforcement learning with safety layer

After completing the modeling of uncertain environments, this paper further constructs a risk sensitive DRL method for power dispatch. Considering that the scheduling actions are mainly continuous variables and the system operation has significant cross time coupling characteristics, this paper adopts Soft Actor Critic (SAC) as the basic framework, and adds a safety action mapping layer and tail risk shaping mechanism on this basis, so that the strategy can maintain sensitivity to constraint feasibility and extreme scenario risks while pursuing low cost [18, 19]. The MDP structure and security layer architecture are shown in Figure 2.

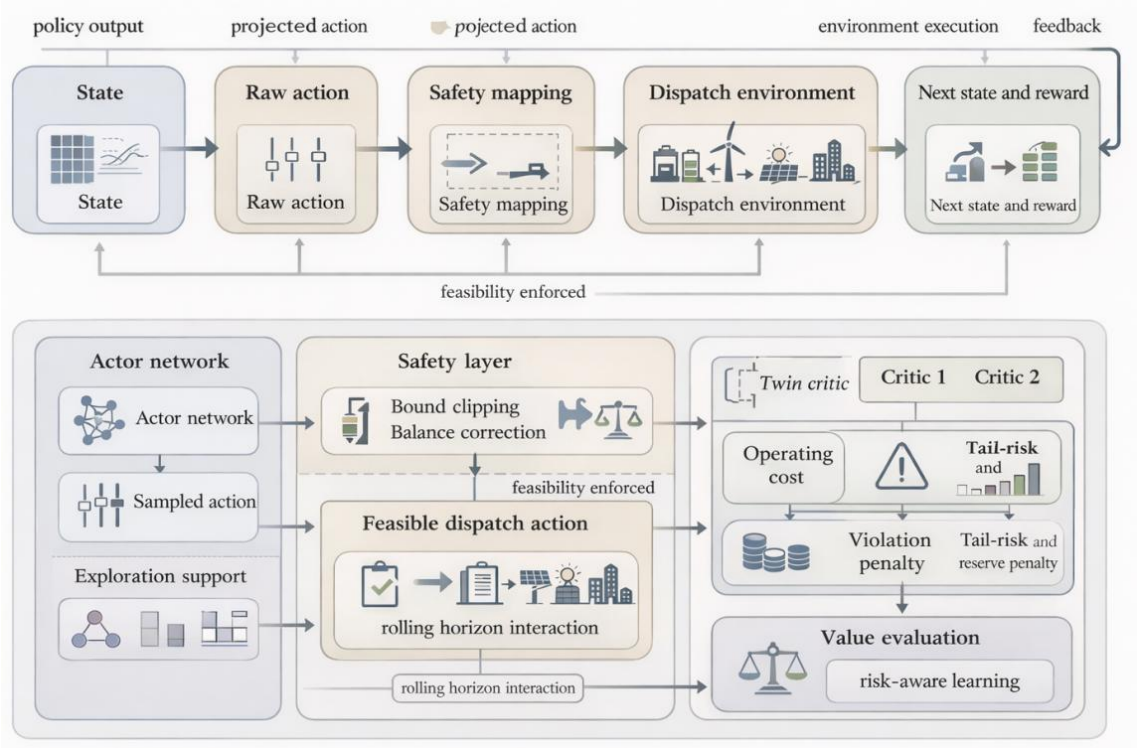


Figure 2: MDP formulation and safety-layer architecture.

As shown in Figure 2, the upper part illustrates the basic relationships among the current state s_t, the raw action $\hat{a}_t$, the safety action a_t, and the environment transition. The lower part connects the actor, twin critic, security layer, environment, and risk modules together. The MDP structure and security layer architecture indicate that the method proposed in this paper separates the processing of "policy learning" and "action correction". The former is responsible for generating exploratory candidate actions, while the latter compresses actions into feasible domains based on unit, energy storage, and backup constraints, thereby avoiding frequent issues such as out of bounds output, state of charge overflow, or insufficient backup in the later stages of training [20]. Assume the policy network generates a raw candidate action $\hat{a}_t$ at time period t based on the state s_t. Due to the fact that the action is directly output by the neural network, its value may violate unit boundaries, power balance, or energy storage constraints. Therefore, this paper introduces a safety action mapping layer before execution to project candidate actions onto the feasible domain Ω, as shown in formula (4).

$$a_t^{safe}=\Pi_{\Omega}(\hat{a}_t) \tag{4}$$

In Equation (4), $\Pi_{\Omega}(\cdot)$ represents the constraint projection operator. In practical implementation, this mapping includes three levels. Firstly, boundary trimming is applied to the output of the unit, the charging and discharging power of energy storage, and the purchasing power. Secondly, the remaining deviation is redistributed based on the power balance relationship [21]. Finally, check the backup lower limit, SOC boundary, and climbing requirements. If there are still parts that cannot be completely repaired, convert them into violation penalties and provide feedback to the environment. This design can effectively reduce sample contamination caused by a large number of infeasible actions during the training phase, and also enable strategy updates to adapt to the "constrained correction" operating environment earlier. In order to avoid the intelligent agent only pursuing the minimum average cost, this paper explicitly introduces tail risk term, violation term, and spare shortage term in the reward function. The definition of immediate return is shown in formula (5).

$$r_t=-\left(C_t+\lambda_1 V_t+\lambda_2 R_t^{short}+\lambda_3 \mathrm{CVaR}_{\alpha,t}\right) \tag{5}$$

In Equation (5), C_t represents the comprehensive operating cost, including conventional unit generation cost, energy storage degradation cost, electricity purchase cost, wind/PV curtailment penalty, and carbon emission cost. V_t represents the cost of constraint violations, such as power limit violations, SOC boundary violations, or ramp conflicts. R_t^{short} represents the penalty for reserve inadequacy. $\mathrm{CVaR}_{\alpha,t}$ represents the tail risk under extreme loss scenarios. λ_1, λ_2 and λ_3 denote the weights for constraint violation penalty, reserve shortfall penalty, and tail risk penalty, respectively. By introducing equation (5), the strategy no longer only focuses on the "lowest average cost", but actively avoids decision-making methods that, although have lower short-term costs, are more likely to lead to load loss or large-scale out of bounds under strong fluctuation conditions. The conditional value at risk used in this article is shown in formula (6).

$$\mathrm{CVaR}_{\alpha}(Z)=\mathrm{E}[\,Z \mid Z\geq \mathrm{VaR}_{\alpha}(Z)\,] \tag{6}$$

In Equation (6), Z represents the random variable of dispatch loss. α denotes the confidence level. E denotes the mathematical expectation. $\mathrm{VaR}_{\alpha}(Z)$ epresents the quantile threshold of the random loss variable Z at confidence level α. Conditional Value at Risk is

used to characterize the average exposure level of the right tail of the loss distribution. Compared with the objective function based solely on expected values, conditional value at risk can more directly penalize large losses under extreme deviation conditions, making agents tend to retain more sufficient backup and more cautious energy storage strategies during high-risk periods. In actual training, the conditional value at risk is approximately estimated from a high loss subset of batch samples, without the need to explicitly solve complex risk optimization subproblems at each step, thus balancing feasibility and risk characterization ability [22].

At the network training level, this article retains the advantages of SAC's double-Q network and entropy regularization, and improves training stability through experience replay and target network soft updates. The main difference from the original SAC is that both value assessment and policy updates revolve around actions after security mapping, which means that critic learns the value of "executable actions" [23]. This can avoid the strategy repeatedly relying on post-processing to "remedy" infeasible actions during training, and also facilitate the formation of control preferences that are closer to actual scheduling logic.

Overall, the algorithm in this article constructs a closed-loop learning process consisting of state input, candidate action generation, safety action correction, risk feedback, and policy updates. Its core advantage lies not in simply replacing reinforcement learning algorithms, but in unifying continuous action control, risk avoidance, and physical feasibility into the same learning framework. For the rolling scheduling problem under high proportion of new energy access, this design is more conducive to improving the stability of the strategy in extreme fluctuations, backup shortages, and amplified prediction deviations, and also provides a methodological basis for subsequent typical days, extreme days, and generalization testing.

2.3 Training setup, benchmarks, and evaluation metrics

To ensure the training convergence of the proposed method and the fairness of comparison between different algorithms, this paper adopts a training mechanism driven by historical samples combined with scene disturbance enhancement. The scheduling cycle is set to 1 hour, and each training round corresponds to 24 consecutive rolling scheduling days. Training, validation, and testing are separated from publicly available time series data and kept in chronological order to avoid information leakage. In addition to using regular historical samples during the training phase, different intensities of joint error perturbations are also injected, including typical scenarios such as low wind solar synchronization, sudden load increase, rapid net load climb, and sustained cloudy and low wind, to enhance the adaptability of the strategy in unseen scenarios. The core consideration here is that if only stationary samples are used for training, the agent is prone to learn conservative or optimistic strategies, but shows significant degradation under extreme fluctuations. The relevant testing system parameters, uncertainty settings, and DRL training hyperparameters are shown in Table 1.

Table 1: Test system parameters, uncertainty settings, and DRL hyperparameters.

Category	Item	Value / Setting
Test system	Dispatch horizon	24 h rolling horizon
Test system	Time resolution	1 h
Test system	Conventional units	6 thermal units
Test system	Renewable units	1 wind farm + 1 PV plant
Test system	Storage unit	1 battery system
Test system	Grid interaction	Bidirectional exchange allowed
Data split	Training / validation / testing	70% / 15% / 15%
Uncertainty setting	Wind error level	Low / medium / high
Uncertainty setting	PV error level	Low / medium / high
Uncertainty setting	Load error level	Low / medium / high
Uncertainty setting	Extreme scenarios	Wind drop, PV drop, load spike, joint disturbance
DRL hyperparameters	Backbone	SAC
DRL hyperparameters	Actor learning rate	3×10^{-4}
DRL hyperparameters	Critic learning rate	1×10^{-3}
DRL hyperparameters	Discount factor γ	0.99
DRL hyperparameters	Batch size	256
DRL hyperparameters	Replay buffer size	1×10^{5}
DRL hyperparameters	Soft update coefficient τ	0.005
DRL hyperparameters	Risk weights	$\lambda1$, $\lambda2$, $\lambda3$ tuned on validation set

Table 1 not only defines the complexity boundaries of the experimental environment in this article, such as rolling time domain, renewable energy composition, and energy storage participation methods. Also clarify the uncertainty intensity and training hyperparameters to make the subsequent results more reproducible. Especially the three error settings of "low/medium/high" and the extreme scenario column provide a unified reference for sensitivity analysis and recovery ability comparison in the following text.

To comprehensively evaluate the proposed methods, this article sets six types of baselines. The first type is Deterministic Economic Dispatch (ED), which refers to deterministic economic dispatch solved only based on point prediction values, and is used to reflect the basic performance of traditional static optimization in uncertain environments [24]. The second type is stochastic ED, which explicitly considers new energy and load fluctuations through scenario sampling. The third type is Robust ED, which deals with uncertainty through interval or worst-case constraints and represents a more conservative strategy. The fourth type is Model Predictive Control (MPC), which repeatedly solves finite time domain optimization problems in the rolling time domain, representing the online "prediction optimization correction" framework. The fifth category includes Proximal Policy Optimization (PPO), Twin Delayed Deep Deterministic Policy Gradient (TD3), and SAC, which respectively characterize mainstream DRL scheduling methods based on policy gradient, deterministic actor critic, and random actor critic. The sixth category is the risk sensitive SAC with safety layer proposed in this article. This baseline setting helps to expand the comparison scope from "the advantages and disadvantages between different reinforcement learning algorithms" to "the overall differences between traditional optimization, rolling optimization, and learning scheduling methods".

In terms of evaluation indicators, this article conducts a joint evaluation from six aspects: economy, renewable energy utilization level, environmental cost, feasibility, real-time

performance, and resilience to abnormal operating conditions. Specifically, it includes: total operating costs, which are used to calculate expenses related to unit output, energy storage losses, power purchase, backup, and abandoned power [25]. Abandoned wind and solar power rate, used to measure the level of renewable energy consumption. Carbon emission cost, used to reflect the environmental burden caused by conventional unit scheduling. Constrained violation rate, used to calculate the proportion of phenomena such as power out of bounds, SOC out of bounds, climbing conflicts, and insufficient backup in all scheduling periods. The average solution time is used to compare the online computational efficiency of different methods. And the ability to recover from extreme scenarios, which refers to the number of time steps required for the system to return to non violation and backup to a safe threshold after being impacted by high disturbances. The "resilience" indicator is specifically retained here because simply looking at costs can easily mask the vulnerability of scheduling strategies under abnormal conditions, while recovery time can more directly reflect the resilience and emergency adjustment quality of the strategy.

In summary, the training settings, baseline system, and evaluation indicators in this section collectively serve three questions: whether the proposed method can reduce operating costs under general conditions, maintain higher constraint satisfaction rates under high uncertainty conditions, and recover to a safe operating range faster in extreme scenarios. Through this experimental design of "multiple baselines+multiple indicators", the subsequent result analysis will no longer be limited to simple cost comparisons, but can more fully reveal the differences in robustness, real-time performance, and risk control capabilities of different scheduling strategies.

3 Results and Discussion

3.1 Overall performance comparison and training behavior

To determine whether the proposed method can converge stably during the training phase and further identify whether its performance improvement comes from faster convergence, improved constraint satisfaction, or later wave suppression, this paper first compares the dynamic performance of PPO, TD3, SAC and the proposed method on the training and validation sets, as shown in Figure 3.

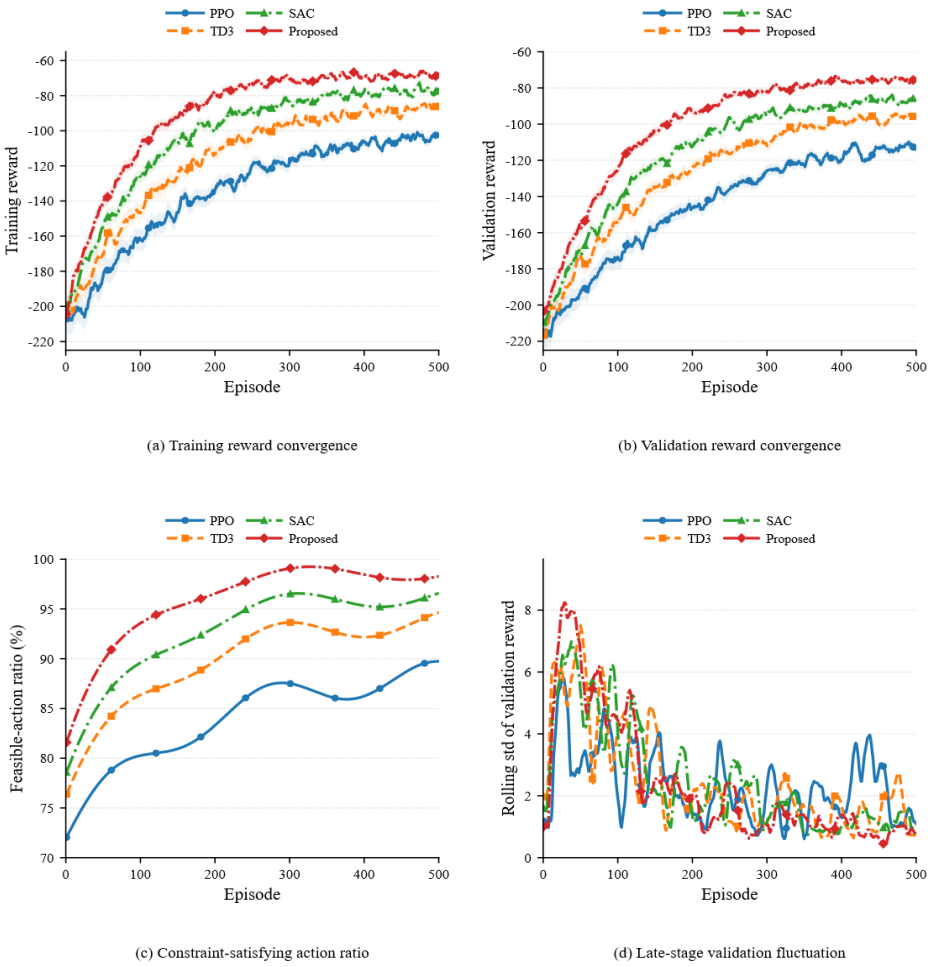


Figure 3: Training behavior of PPO, TD3, SAC, and the proposed method.

Figure 3 (a) shows that all four learning methods can gradually improve cumulative returns during the training process, but there are significant differences in convergence slope and tail stationarity. PPO rises slowly within the first 150 episodes and has not yet fully entered the stable zone around 400 episodes. The improvement speed of TD3 and SAC is significantly faster, achieving higher training returns in the mid-term; The upward trend of the proposed method is most concentrated, stabilizing after about 300 episodes, and the training return in the final stage remains around -70. The validation return variation shown in Figure 3 (b) is consistent with the training curve, but the differences between the methods are clearer. The final validation return for PPO is approximately -104, TD3 is -91, and SAC is -84. However, the proposed method has been improved to -75, and the validation end still retains an advantage of approximately 10 reward points compared to SAC. This indicates that the benefits of the proposed method are not only present within the training samples, but also maintain higher strategy quality on data that has not been updated.

Figure 3 (c) further illustrates the proportion of constraint satisfied actions. This indicator reflects the proportion of the output of the strategy network that can directly fall into the feasible control interval after environmental interaction, and can therefore be regarded as an intuitive representation of the learner's adaptability to physical constraints. From the results, PPO remained at around 89% in the later stage, while TD3 and SAC were close to 94% and 96%, respectively. The proposed method ultimately reached about 98.5%. This difference is closely related to the security action mapping layer introduced earlier. Due to the fact that the original candidate actions have been projected into the feasible domain before execution, the

feedback obtained during the policy update process is closer to the real scheduling constraints, resulting in a higher proportion of executable actions in the later stage. The rolling standard deviation of the validation report in Figure 3 (d) also supports this judgment. PPO still has significant fluctuations in the post training stage, while TD3 and SAC show significant weakening, but local rebound still occurs after high disturbance samples enter the replay pool. The rolling volatility of the proposed method is the lowest, and the last 100 episodes remain within a relatively narrow range, indicating that its later strategy is more stable and less sensitive to sample perturbations. From the four subplots in Figure 3, it can be seen that the performance advantage of the proposed method is not caused by a single increase in returns, but is reflected in four aspects: faster training convergence, more stable validation performance, more sufficient constraint adaptation, and weaker tail oscillation.

On this basis, this article further examines the comprehensive performance of various scheduling methods on the test set. Considering that comparing only the total operating cost can easily mask the differences between violations, power abandonment, and real-time performance, Figure 4 organizes the overall performance from four perspectives: the original indicator matrix, normalized score matrix, relative baseline gain matrix, and coefficient of variation matrix, which are used to answer the four questions of "how absolute the values of each method are", "how comprehensive the ranking is", "how much the proposed method improves relative to the baseline", and "whether the results are stable enough".

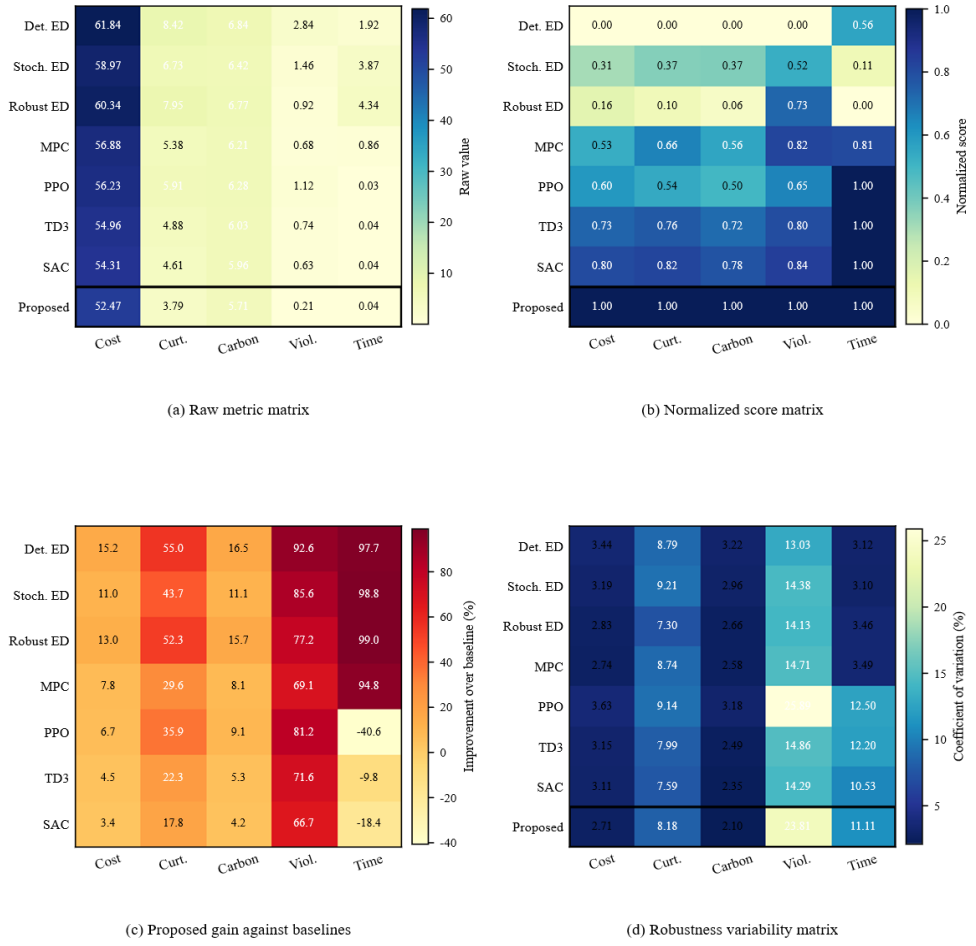


Figure 4: Overall performance comparison across deterministic, optimization-based, and DRL-based dispatch methods.

Figure 4 (a) shows five original indicators of the test set, including total operating cost, wind and solar curtailment rate, carbon emission cost, constraint violation rate, and average solution time. Deterministic ED performs the worst in terms of cost, power curtailment, and violation, with a total cost of 61.84 × 104 CNY/day, a wind and solar curtailment rate of 8.42%, and a violation rate of 2.84%. Stochastic ED and Robust ED can reduce risk exposure to varying degrees, but at the cost of significantly increasing solution time, with the average single step time of Robust ED reaching 4.34 seconds. MPC is generally superior to traditional ED, but its online solution cost is still significantly higher than that of learning methods. Among the three original DRL methods, SAC has the best overall performance, reducing the total cost to 54.31×10^4 CNY/day, with a violation rate of 0.63%. The proposed method further compresses the total cost to 52.47×10^4 CNY/day, reduces the curtailment rate to 3.79%, carbon emission cost to 5.71×10^3 CNY/day, and the violation rate to only 0.21%, indicating a better balance between economy and operational feasibility.

Figure 4 (b) compares the five indicators horizontally after normalizing them uniformly. Due to the alignment of numerical scales, the comprehensive contours of different methods are more intuitive. The proposed method is in the optimal region in terms of Cost, Curtailment, Carbon, and Violation, with only the Time metric slightly lower than the fastest in pure inference DRL methods. Combining Figure 4 (a), it can be seen that this slight time disadvantage comes from the additional calculations brought by the security layer and risk assessment module, but its absolute value is still only 0.045 s/step, far lower than the online solving time of MPC and two types of ED methods, so it will not weaken its real-time scheduling applicability.

In order to more clearly measure the improvement of the proposed method relative to each baseline, the gain matrix is shown in Figure 4 (c). Compared with SAC, the proposed method improves total cost, abandonment rate, carbon emission cost, and violation rate by approximately 3.39%, 17.79%, 4.19%, and 66.67%, respectively; Compared with MPC, the corresponding gains are 7.75%, 29.55%, 8.05%, and 69.12%, respectively. Even compared to the more conservative Robust ED, the proposed method still has over 77% room for improvement in violation rate. This result indicates that the improvement brought by our method is not limited to a single indicator, but covers multiple dimensions such as economy, consumption capacity, and safety. Figure 4 (d) further compares the dispersion of the results of each method from a stability perspective, using the coefficient of variation to characterize the fluctuation level of the test set. Deterministic ED and PPO show high dispersion in multiple indicators, indicating that they are more sensitive to sample perturbations; The fluctuation of SAC has significantly converged; The coefficient of variation of the proposed method is generally the lowest or close to the lowest in the five indicators, especially in the violation rate and abandoned wind and solar power rate. This means that the proposed method not only achieves better results at the mean level, but also has more stable performance output, and is less prone to significant degradation when facing different day patterns and error intensities.

In summary, the advantages of the proposed method are reflected in two aspects. Firstly, during the training phase, the combination of safety action mapping and risk sensitive rewards enhances the feasibility and convergence stability of strategy learning, resulting in higher validation returns and smaller fluctuations in the later stages. Secondly, during the testing phase, this training benefit is transformed into lower total costs, lower wind and solar curtailment rates, lower carbon emission costs, and significantly reduced constraint violation rates, while still maintaining a sufficiently fast online solving speed. It can be concluded that the proposed method has strong comprehensive performance advantages in uncertainty aware dispatch scenarios, providing a reliable foundation for subsequent typical day analysis and

extreme scenario discussions.

3.2 Typical-day dispatch trajectories and extreme-scenario analysis

To further investigate the actual scheduling behavior of the proposed method in high volatility scenarios, this paper selects summer high photovoltaic days and winter high load days from the test set as typical operating conditions, and constructs two extreme events, wind power sudden drops and load spikes, in independent disturbance samples. Unlike the overall indicators in the previous section, this section focuses more on the linkage process of various regulatory resources on the timeline, namely how energy storage, heat engines, backup, and external power purchases share the responsibility of balancing when disturbances intensify, and why this collaborative regulation can bring lower violation rates and faster recovery speeds. Firstly, discuss typical daily scenarios. Figure 5 shows the intraday scheduling trajectories and their corresponding response information for two representative types of days.

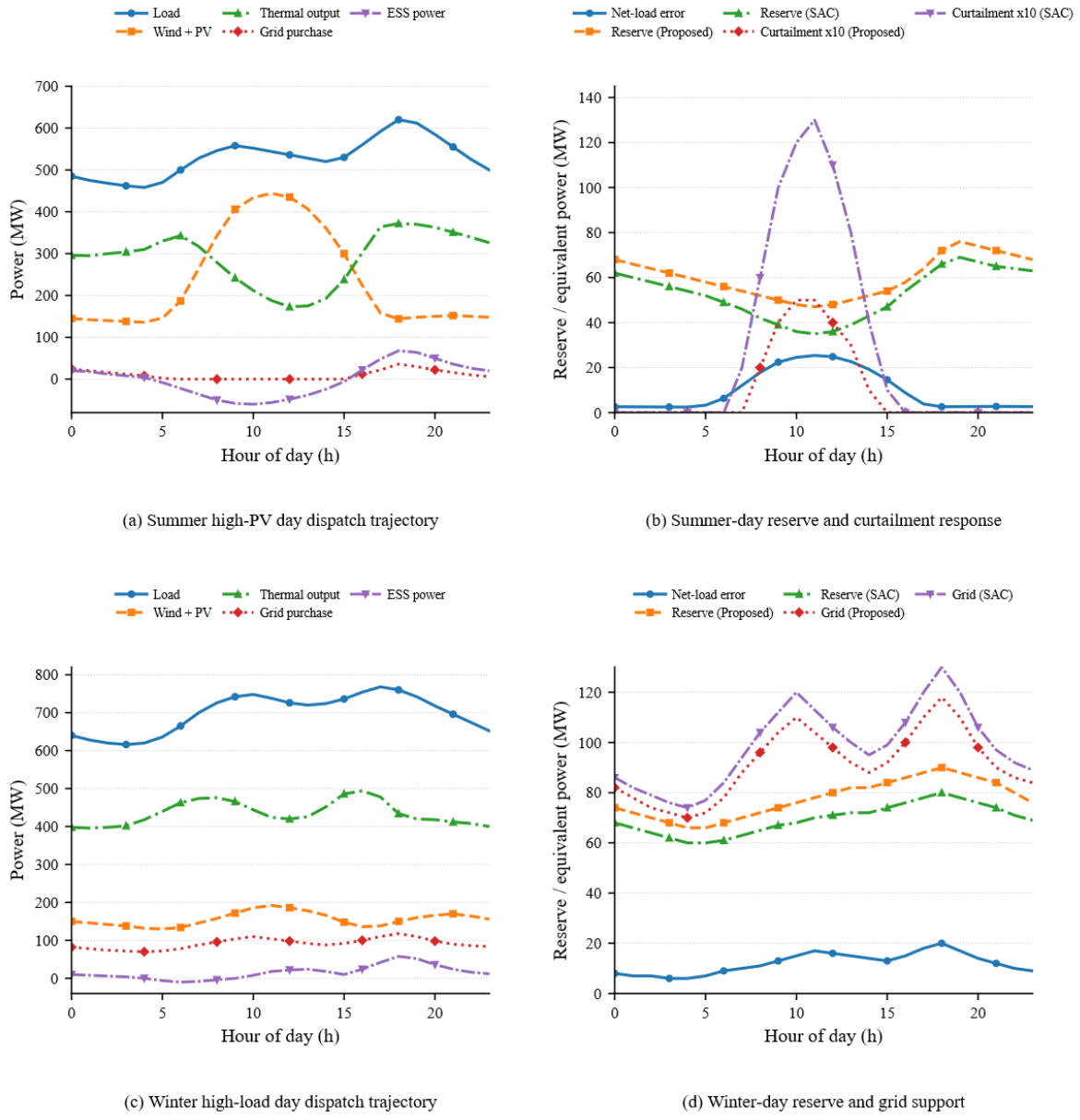


Figure 5: Dispatch trajectories on typical summer and winter days.

From Figure 5 (a), it can be seen that the main characteristic of high photovoltaic days in summer is the rapid rise of photovoltaics during the noon period. The total output of wind and solar energy from 10:00 to 14:00 is close to 440 MW. In response to this characteristic, the proposed method significantly enhances the energy storage and absorption capacity around noon, with the lowest charging power reaching around -60 MW. At the same time, the thermal engine output is reduced to the range of 220MW to 240 MW, and the grid exchange is gradually shifted from purchasing electricity to small-scale external transmission. In this way, the newly added renewable energy output during the noon period is not directly converted into large-scale power abandonment, but partially enters energy storage and partially replaces heat engines and external electricity. The accompanying Figure 5 (b) further indicates that although the proposed method actively reduces the reserve at noon in summer, the reserve remains within the safe range of 47MW to 50 MW, which is higher than the lowest level of SAC during the same period; The corresponding peak amount of abandoned electricity is only about 5 MW, while SAC reaches 13 MW at the same time. This indicates that this strategy has stronger adaptability to the "high photovoltaic low net load" state at noon, and can improve the level of new energy consumption while maintaining basic regulation margin.

More importantly, the proposed method did not establish the low-cost operation at noon on the basis of excessive overdraft for subsequent periods. Figure 5 (a) shows that as photovoltaics rapidly decline after 16:00, energy storage shifts from charging to discharging, and the discharge power increases to 68 MW around 18:00, with the thermal output also recovering synchronously. In Figure 5 (b), the standby level was raised again to above 70 MW in the evening, indicating that the strategy had reserved sufficient adjustment margin for climbing during the evening peak at noon. Therefore, the performance improvement in the summer samples is not only due to the absorption of several MW of photovoltaic energy at noon, but also comes from the closer coordination arrangement between the pre - and post day periods. Corresponding test results show that the total cost of the proposed method on the summer high-PV day is 49.82×10^4 CNY/day, lower than SAC's 51.43×10^4 CNY/day, and the violation rate decreases from 0.44% to 0.12%.

The operating logic of high load days in winter is significantly different from that in summer. In Figure 5 (c), the system load reaches its peak around 17:00-20:00, reaching a maximum of nearly 770 MW, while the contribution of renewable energy is significantly weaker than in the summer scenario. Under these conditions, the proposed method does not rely on a single regulation link to fill the load gap, but simultaneously increases the output of the heat engine, energy storage and discharge, and external power purchase. During the evening rush hour, the purchased power increased to 110MW to 118 MW, the energy storage discharge reached 58 MW, and the thermal engine remained in a relatively high but not overly aggressive output range. This arrangement is directly reflected in Figure 5 (d): the proposed method still maintains a reserve of about 80MW to 90MW during high load periods, while the overall reserve level of SAC is lower; At the same time, SAC requires higher purchasing power to maintain the same load balance. In other words, the advantages of the proposed method in winter scenarios mainly come from a more reasonable grasp of the relationship between "thermal regulation speed - energy storage release rhythm - power purchase support strength". Correspondingly, the total cost of winter high load days decreased from 59.18 × 104 CNY/day in SAC to 57.91 × 104 CNY/day, and the violation rate decreased from 0.76% to 0.28%.

Outside of typical days, extreme events better reflect the practical value of uncertainty aware scheduling. Figure 6 adopts a more compact isosurface form to display the dynamic response intensity and multi method comparison results under sudden wind power drop and load peak, respectively.

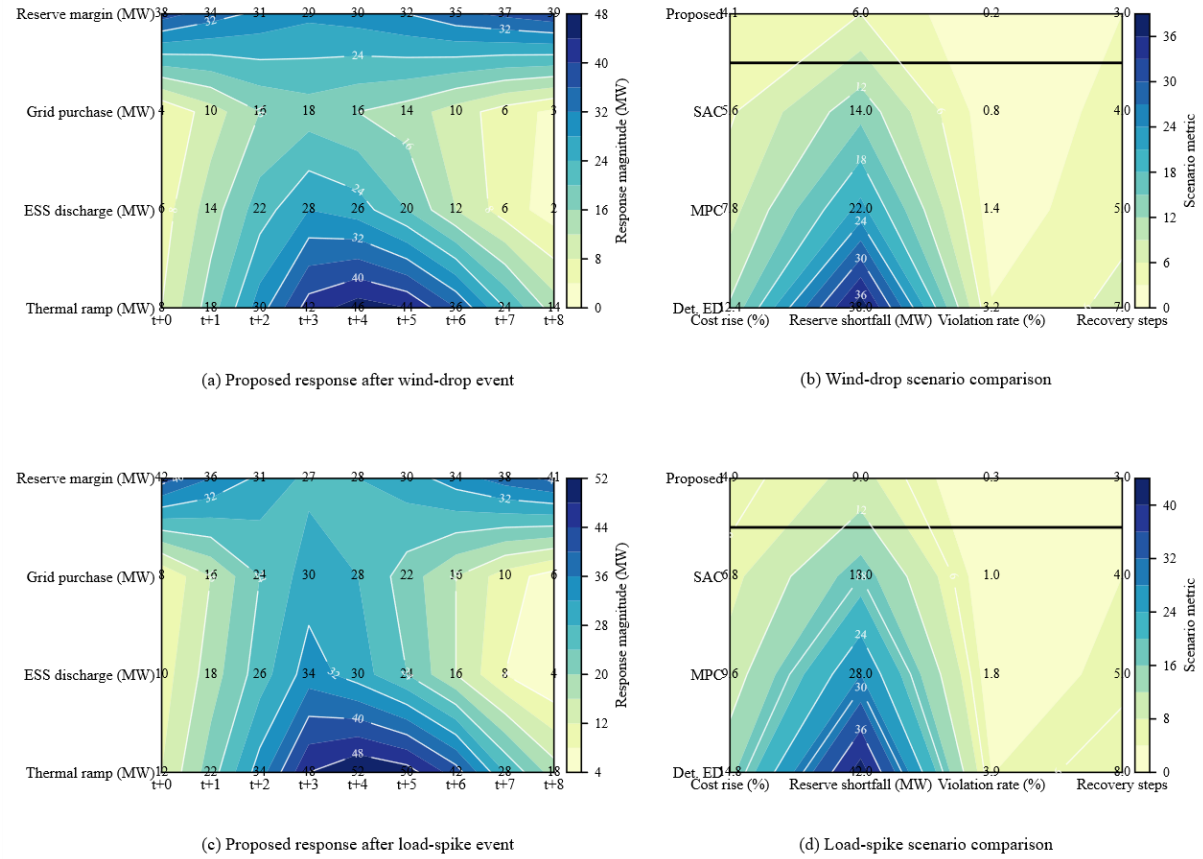


Figure 6: Extreme-scenario response and comparison under wind-drop and load-spike events.

Figure 6 (a) shows that during the nine consecutive periods after the sudden drop in wind power event, the proposed method first rapidly increases the output of the thermal engine, with a ramp range of 42-46 MW from t+3t+3t+3 to t+5t+5t+5. At the same time, the energy storage discharge reaches 26-28 MW during the same period, with a purchase increment of 16-18 MW and a minimum reserve margin of around 29 MW. This response process indicates that the proposed strategy did not concentrate the compensation responsibility on a single resource when facing a sudden decrease in wind power, but instead mitigated the impact through the allocation of heat engines, energy storage, and electricity purchases. The comparison results in Figure 6 (b) further confirm this point, where in the scenario of sudden wind power drop, the cost increase of Deterministic ED is 12.4%, and the recovery time is 7 steps. SAC is 5.6% and 4 steps respectively; The proposed method compresses these two indicators to 4.1% and 3 steps, while reducing the reserve shortfall from SAC's 14 MW to 6 MW, and lowering the violation rate from 0.8% to 0.2%. This indicates that the proposed method can quickly bring the system back to the safe operating range and is less likely to trigger new constraint conflicts during the recovery process.

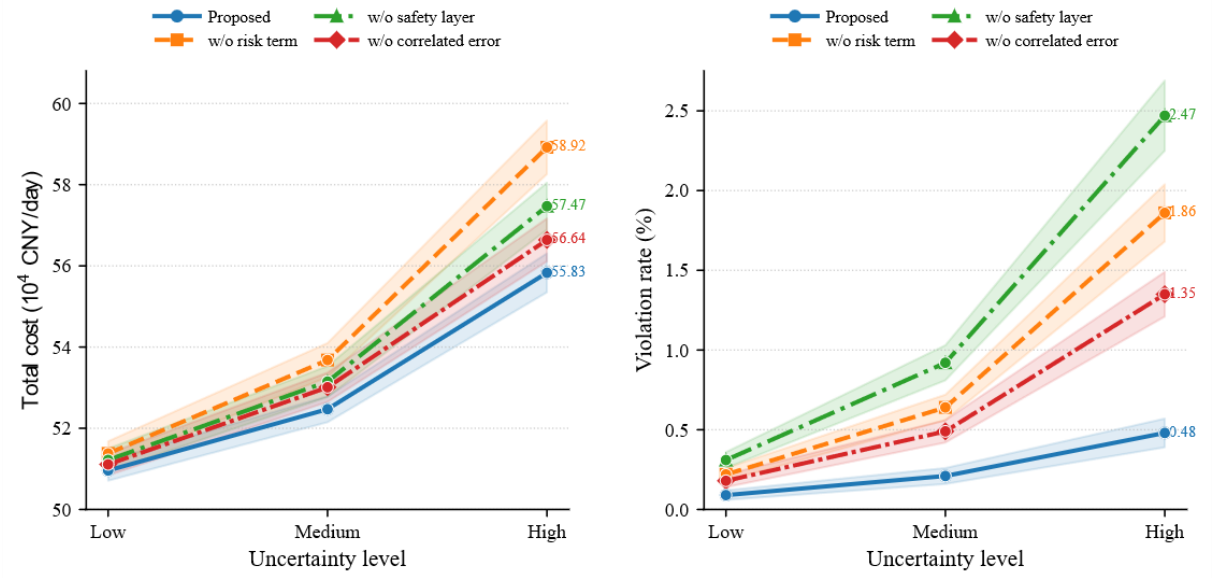
The response to peak load events exhibits similar but not identical characteristics. In Figure 6 (c), after the event, the increase in thermal power generation further increased, reaching a maximum of 52 MW; the energy storage discharge increased to 34 MW. The maximum increase in power purchase was about 30 MW. The reserve margin decreased to a minimum of 27 MW and rebounded to over 30 MW after three time steps. Compared with the sudden drop in wind power, the direct impact of load spikes on the system is more concentrated on the demand side, so heat engines and power purchases bear a higher proportion of compensation tasks. Figure 6 (d) shows that in this scenario, the cost increase of

the proposed method is 4.9%, which is lower than SAC's 6.8%; The reserve shortfall is 9 MW, while SAC is 18 MW. The violation rate has decreased from 1.0% to 0.3%, and the recovery time has also been shortened from 4 steps to 3 steps. Based on Figure 6 (c), it can be seen that the key to the faster recovery of the proposed method lies in its ability to simultaneously call multiple types of adjustment resources at the beginning of the event, rather than passively switching to the next method after a certain resource reaches the boundary.

In summary, the advantages of the proposed method in high volatility scenarios are mainly reflected in smoother linkage between adjustment resources, more proactive backup management, and shorter recovery processes. For high photovoltaic days in summer, its value lies in more effectively absorbing surplus new energy at noon and reserving adjustment margin for evening climbing. For winter high load days, its value lies in reducing the boundary pressure of the heat engine and improving the quality of support during late peak hours. For extreme disturbances, their value lies in the ability to identify risk exposure earlier and quickly organize the joint response of heat engines, energy storage, and power purchase. From this, it can be seen that uncertainty aware state representation not only improves the average cost, but also substantially enhances the adaptability of strategies to complex fluctuations and abnormal operating conditions.

3.3 Ablation, sensitivity, and generalization discussion

To clarify the sources of performance improvement and test the adaptability of the model outside the training distribution, this paper conducted three sets of ablation experiments, tests with different levels of uncertainty, month/cross year validation outside the training set, and extreme weather disturbance tests in sequence. The focus here is not only on mean performance, but also on the stability and recovery quality of the model under high volatility conditions. Firstly, Figure 7 shows the cost and violation rate changes of four model settings at low, medium, and high levels of uncertainty, including three ablation versions: complete model, risk item removal, safety layer removal, and related error input removal.



(a) Cost sensitivity under ablation settings

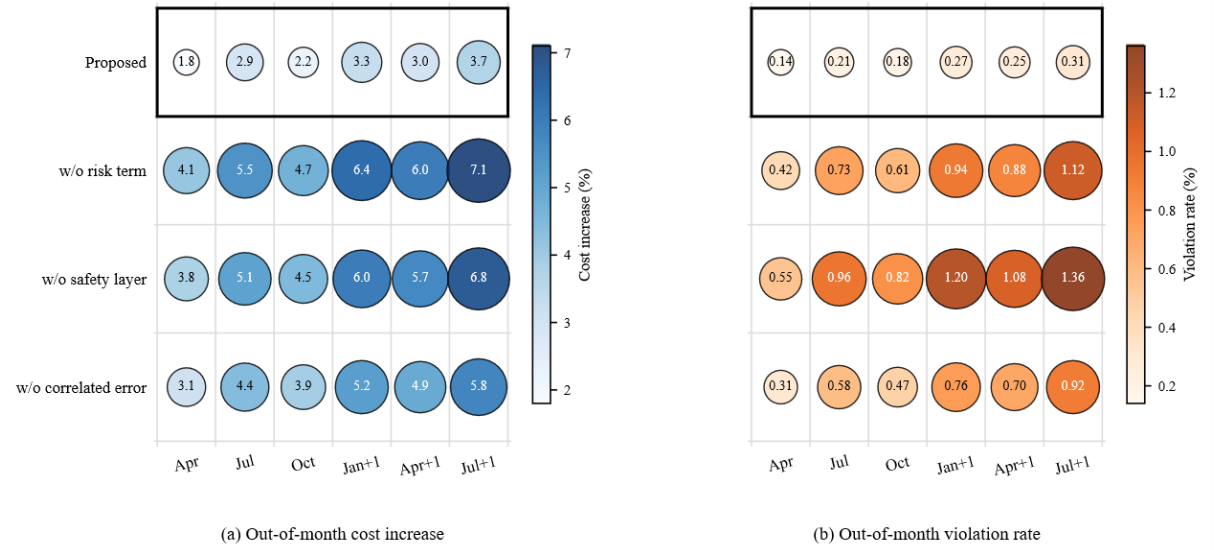
(b) Violation sensitivity under ablation settings

Figure 7: Ablation and sensitivity results under uncertainty scaling.

Figure 7 (a) shows that as the uncertainty level increases from Low to High, the operating

costs of all models significantly increase, but the growth rate of the complete model is the smallest. The total cost increased from 50.96 × 10^4 CNY/day to 55.83 × 10^4 CNY/day, and after removing the risk item, the corresponding value increased to 58.92 × 10^4 CNY/day. After removing the security layer and related error inputs, they are 57.47 × 10^4 and 56.64 × 10^4 CNY/day, respectively. This indicates that the risk term is particularly crucial for suppressing cost amplification under high disturbance conditions, and the related error input does indeed improve the model's adaptability to uncertainty escalation. Figure 7 (b) further indicates that the security layer has the most direct impact on operational feasibility. Under high conditions, the violation rate of the complete model is 0.48%, the risk-free model increases to 1.86%, the unsafe layer model reaches 2.47%, and the uncorrelated error input model also reaches 1.35%. This result indicates that risk shaping helps to reduce the tendency for aggressive scheduling to exceed boundaries, while the security layer further pushes the policy output back into the physically executable range. The absence of either will significantly reduce stability in high disturbance scenarios.

After confirming the roles of each module, this article further examines the model's generalization performance outside of the training samples. Figure 8 shows the cost increase and violation rate in six sets of tests, including April, July, October, and Jan+1, Apr+1, and July+1 after the New Year.



(a) Out-of-month cost increase

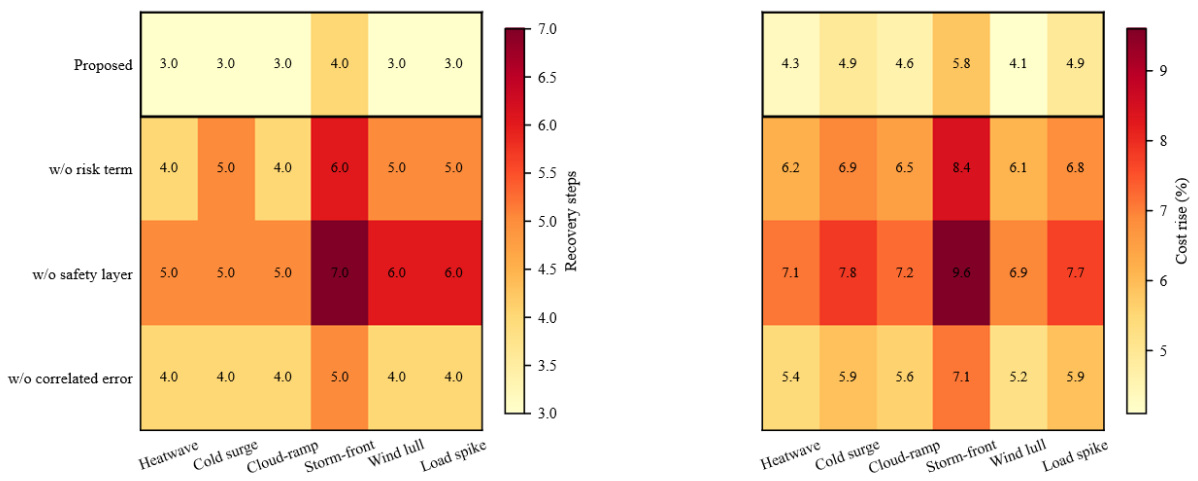
(b) Out-of-month violation rate

Figure 8: Out-of-month and out-of-year generalization performance.

From Figure 8 (a), it can be seen that the cost increase of the complete model remains between 1.8% and 3.7% in all months outside the training set, while the risk-free model expands to 4.1% -7.1%, the unsafe layer model is 3.8% -6.8%, and the uncorrelated error input model is 3.1% -5.8%. This difference is more pronounced in cross year samples, indicating that the complete model is less sensitive to temporal distribution drift. The violation rate results shown in Figure 8 (b) are consistent with this: the complete model maintains a violation rate of 0.14% -0.31% on all external samples, significantly lower than the other three ablation versions. The model without a security layer reached 1.36% in the July+1 sample, which is close to four times the upper bound of the complete model. Combining the two sets of results in Figure 8, it can be seen that the correlation error input and risk sensitive design not only improve the performance within the training set, but also enhance the transferability of the model when facing unseen months and cross year samples.

The extreme weather disturbance further tested the resilience of the model under abnormal

conditions. Figure 9 compares the recovery steps and cost increases in six scenarios: Heatwave, Cold surge, Cloud ram, Storm front, Wind roll, and Load spike.



(a) Recovery steps under extreme weather

(b) Cost rise under extreme weather

Figure 9: Robustness comparison under extreme-weather disturbances.

Figure 9 (a) shows that the recovery steps of the complete model under six types of extreme weather are concentrated between 3-4 steps, while the risk-free models mostly have 4-6 steps. The model without a safety layer is further expanded to 5-7 steps, and the model without relevant error input has 4-5 steps. Storm front is the most impactful scenario, with a complete model requiring 4 steps to recover, while a model without a security layer requires 7 steps. Figure 9 (b) shows that the cost increase of the complete model under extreme weather conditions is 4.1% -5.8%, corresponding to 6.1% -8.4% for the risk-free model, 6.9% -9.6% for the unsafe layer model, and 5.2% -7.1% for the uncorrelated error input model. This result indicates that both the safety layer and risk items have a sustained contribution to the scheduling quality under abnormal operating conditions, with the safety layer having a more direct impact on recovery speed and the risk item more significantly suppressing cost deterioration in extreme scenarios.

In summary, the risk items mainly improve cost control under high disturbance and extreme weather conditions, while the safety layer is mainly responsible for reducing violation rates and shortening recovery time. The related error inputs enhance the model's adaptability to uncertainty enhancement and cross month samples. After the combined action of the three factors, the complete model maintained the most stable overall performance in ablation testing, validation outside the training set, and extreme weather disturbances. Therefore, the performance advantage of our method has a clear structural source, rather than the result of a single parameter setting accidentally.

4 Conclusion

This article focuses on the power system scheduling problem under high proportion of new energy access conditions, and constructs a DRL method that integrates joint disturbance representation, risk sensitive rewards, and safety action mapping. The system is validated by combining typical days, extreme disturbances, and samples outside the training set. The results indicate that the proposed framework can maintain good economy, feasibility, and operational stability under conditions of increased uncertainty, providing a feasible intelligent

optimization approach for online scheduling of new power systems. The main conclusions are as follows:

(1) The proposed method can adapt well to continuous rolling scheduling scenarios, and can still achieve coordinated allocation of thermal, energy storage, backup, and power purchasing resources even when wind and solar fluctuations intensify and net load changes rapidly, demonstrating the potential application of deep reinforcement learning in uncertainty aware dispatch.

(2) Risk sensitive mechanisms can effectively suppress cost amplification in high loss scenarios. After incorporating tail risk, reserve shortfall, and violation penalties into the reward design, the strategy exhibits better stability and resilience under high uncertainty and abnormal disturbance conditions.

(3) The security action mapping layer has a direct effect on improving the feasibility of policy execution. By correcting candidate actions within the physically feasible domain, it is possible to significantly reduce out of bounds scheduling, backup shortages, and state violations, thereby enhancing operational reliability.

Funding

This work was supported by Science and Technology Plan of Inner Mongolia Autonomous Region (2022JBGS0044).

About the Author

Shuo Yu, male, postgraduate, engineer. He works at the Power Dispatching and Control Branch of Inner Mongolia Power (Group) Co., Ltd. His main research direction is power system.

Jingbo Wang, male, engineer. He works at the Power Dispatching and Control Branch of Inner Mongolia Power (Group) Co., Ltd. His main research directions include power system stability analysis, electricity market, and friendly grid integration of new energy.

Qiang Li, male, senior engineer. He works at the Power Dispatching and Control Branch of Inner Mongolia Power (Group) Co., Ltd. His main research directions include power system dispatching, power system stability analysis, etc.

Rui Yang, female, Han, master's degree, engineer. She works at Beijing Tsintergy Technology Co., Ltd. Her main research directions include power system dispatch optimization and electricity market.

Hongyu Tang, male, Han, master's degree, engineer. He works at Beijing Tsintergy Technology Co., Ltd. His main research directions include power system dispatch optimization and electricity market.

References

[1] Li, Q., Lin, T., Yu, Q., et al. (2023). Review of deep reinforcement learning and its application in modern renewable power system control. Energies, 16(10), 4143. DOI: 10.3390/en16104143

[2] Pesántez, G., Guamán, W., Córdova, J., et al. (2024). Reinforcement learning for efficient power systems planning: A review of operational and expansion strategies. Energies, 17(9), 2167. DOI: 10.3390/en17092167

[3] Feng, B., Liu, Z., Huang, G., et al. (2023). Robust federated deep reinforcement learning for optimal control in multiple virtual power plants with electric vehicles. Applied Energy, 349, 121615. DOI: 10.1016/j.apenergy.2023.121615

[4] Jiang, W., Liu, Y., Fang, G., et al. (2023). Research on short-term optimal scheduling of hydro-wind-solar multi-energy power system based on deep reinforcement learning. Journal of Cleaner Production, 385, 135704. DOI: 10.1016/j.jclepro.2022.135704

[5] Zhang, Y., Han, Y., Liu, D., et al. (2023). Low-carbon economic dispatch of electricity-heat-gas integrated energy systems based on deep reinforcement learning. Journal of Modern Power Systems and Clean Energy, 11(6), 1827-1841. DOI: 10.35833/MPCE.2022.000671

[6] Ebrie, A. S., Kim, Y. J. (2024). Reinforcement learning-based optimization for power scheduling in a renewable energy connected grid. Renewable Energy, 230, 120886. DOI: 10.1016/j.renene.2024.120886

[7] Wang, X., Zhong, H., Zhang, G., et al. (2024). Adaptive look-ahead economic dispatch based on deep reinforcement learning. Applied Energy, 353, 122121. DOI: 10.1016/j.apenergy.2023.122121

[8] Gao, Z., Kang, W., Chen, X., et al. (2024). Optimal economic dispatch of a virtual power plant based on gated recurrent unit proximal policy optimization. Frontiers in Energy Research, 12, 1357406. DOI: 10.3389/fenrg.2024.1357406

[9] Huang, W., Qian, T., Tang, W., et al. (2025). A distributionally robust chance constrained optimization approach for security-constrained optimal power flow problems considering dependent uncertainty of wind power. Applied Energy, 383, 125264. DOI: 10.1016/j.apenergy.2024.125264

[10] Shi, J., Wang, B., Yuan, R., et al. (2023). Rolling horizon wind-thermal unit commitment optimization based on deep reinforcement learning. Applied Intelligence, 53, 19591-19609. DOI: 10.1007/s10489-023-04489-5

[11] Wang, C., Zhang, J., Wang, A., et al. (2024). Prioritized sum-tree experience replay TD3 DRL-based online energy management of a residential microgrid. Applied Energy, 368, 123471. DOI: 10.1016/j.apenergy.2024.123471

[12] Liu, W. C., Mao, Z. Z. (2025). Microgrid economic dispatch using information-enhanced deep reinforcement learning with consideration of control periods. Electric Power Systems Research, 239, 111244. DOI: 10.1016/j.epsr.2024.111244

[13] Wu, P., Chen, C., Lai, D., et al. (2025). Real-time optimal power flow method via safe deep reinforcement learning based on primal-dual and prior knowledge guidance. IEEE Transactions on Power Systems, 40(1), 597-611. DOI: 10.1109/TPWRS.2024.3395248

[14] Sayed, A., Al Jaafari, K., Zhang, X., et al. (2025). Efficient optimal power flow learning: A deep reinforcement learning with physics-driven critic model. International Journal of Electrical Power & Energy Systems, 167, 110621. DOI: 10.1016/j.ijepes.2025.110621

[15] Feng, J., Wang, H., Yang, Z., et al. (2023). Economic dispatch of industrial park

considering uncertainty of renewable energy based on a deep reinforcement learning approach. Sustainable Energy, Grids and Networks, 34, 101050. DOI: 10.1016/j.segan.2023.101050

[16] Feng, W., Deng, B., Zhang, Z., et al. (2024). Low-carbon economic dispatch strategy for integrated electrical and gas system with GCCP based on multi-agent deep reinforcement learning. Frontiers in Energy Research, 12, 1428624. DOI: 10.3389/fenrg.2024.1428624

[17] Liang, T., Zhang, X., Tan, J., et al. (2024). Deep reinforcement learning-based optimal scheduling of integrated energy systems for electricity, heat, and hydrogen storage. Electric Power Systems Research, 233, 110480. DOI: 10.1016/j.epsr.2024.110480

[18] Ye, J., Wang, X., Hua, Q., et al. (2024). Deep reinforcement learning based energy management of a hybrid electricity-heat-hydrogen energy system with demand response. Energy, 305, 131874. DOI: 10.1016/j.energy.2024.131874

[19] Liu, J., Meng, X., Wu, J. (2025). Data-driven optimal scheduling for integrated electricity-heat-gas-hydrogen energy system considering demand-side management: A deep reinforcement learning approach. International Journal of Hydrogen Energy, 103, 147-165. DOI: 10.1016/j.ijhydene.2025.01.185

[20] Shuai, Q., Yin, Y., Huang, S., et al. (2025). Deep reinforcement learning-based real-time energy management for an integrated electric-thermal energy system. Sustainability, 17(2), 407. DOI: 10.3390/su17020407

[21] Liu, M., Zhu, J., Liu, M. (2025). Economic dispatch of virtual power plant using reinforcement learning method with improved state space and hybrid action representation. Engineering Applications of Artificial Intelligence, 111725. DOI: 10.1016/j.engappai.2025.111725

[22] Zhou, L., Huo, L., Liu, L., et al. (2025). Optimal power flow for high spatial and temporal resolution power systems with high renewable energy penetration using multi-agent deep reinforcement learning. Energies, 18(7), 1809. DOI: 10.3390/en18071809

[23] Zhang, H., Zhang, Y., Zhang, J., et al. (2025). Resilient dispatching optimization of power system driven by deep reinforcement learning model. Discover Artificial Intelligence, 5, 189. DOI: 10.1007/s44163-025-00451-1

[24] Xiong, B., Zhang, L., Hu, Y., et al. (2025). Deep reinforcement learning for optimal microgrid energy management with renewable energy and electric vehicle integration. Applied Soft Computing, 176, 113180. DOI: 10.1016/j.asoc.2025.113180

[25] Wang, C., Ma, Y., Xie, J., et al. (2025). Multi-objective energy dispatch with deep reinforcement learning for wind-solar-thermal-storage hybrid systems. Journal of Energy Storage, 105, 114635. DOI: 10.1016/j.est.2024.114635