Risk-Aware Power System Real-Time Economic Dispatch via Distributional Reinforcement Learning

Jiachen Zhang
Department of Electrical Engineering
Tsinghua University
Beijing, China
zhangjc23@mails.tsinghua.edu.cn

Haotian Zhao
Department of Electrical Engineering
Tsinghua University
Beijing, China
zhaohaotian@tsinghua.edu.cn

Qinglai Guo
Department of Electrical Engineering
Tsinghua University
Beijing, China
guoqinglai@tsinghua.edu.cn

***Abstract*—The increasing uncertainty from renewable energy sources and fluctuating loads poses significant challenges to traditional real-time economic dispatch (RTED). This paper proposes a risk-aware RTED framework based on distributional reinforcement learning (DRL). The methodology enhances the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm by integrating a quantile-based distributional critic, which enables precise modeling of the return distribution. By optimizing a distorted expectation objective, the framework can explicitly manage tail risks and formulate risk-sensitive dispatch policies. Furthermore, by incorporating action projection and reward penalties, constraint violations can be significantly mitigated. Experimental results on a modified IEEE 33-bus system demonstrate the validity of the proposed method over traditional reinforcement learning baselines.**

***Index Terms*—Distributional reinforcement learning, real-time economic dispatch, uncertainty.**

I. Introduction

THE increasing penetration of variable renewable generation and volatile demand profiles introduces significant uncertainty into the real-time economic dispatch (RTED) of power systems. Over short horizons, this uncertainty is tightly coupled with network constraints, rendering risk management a vital design objective in RTED. Traditionally, stochastic programming [1] and robust optimization [2] have been extensively employed to address uncertainty. In practical scenarios, however, operators observe only historical measurements and short-term forecasts; the true probability law remains unknown ex ante and may drift over time. Consequently, stochastic formulations that depend on assumed probability distributions are susceptible to model misspecification. Robust formulations generally rely on fixed-shape uncertainty sets that tend to be inflexible and conservative. Both approaches necessitate either an explicit distributional model or a carefully defined ambiguity set. Nevertheless, such information is seldom available, thereby restricting their practical application in real-world operations [3].

Beyond model-based methods that require pre-defined assumptions, a data-driven alternative is to learn dispatch policies directly from operational data. Reinforcement learning (RL) offers such a pathway by optimizing policies from interaction without fully specifying system dynamics or probability laws. However, most standard RL algorithms [4]–[6] maximize expected return and are therefore effectively risk-neutral. In RTED, source and load uncertainty primarily manifests as power-balance deviations, shrinking feasibility regions and causing volatile operating costs. Expectation-based policies can overlook tail realizations and perform poorly in certain scenarios. Addressing this risk sensitivity requires moving beyond mean performance to reason about the distribution of future rewards explicitly.

To this end, distributional RL models the entire distribution of returns rather than solely its expected value. This facilitates explicit management of tail risk. Early categorical methods [7] approximate the value distribution on a predetermined support through projected Bellman updates. Quantile-based approaches [8]–[10] learn quantile functions via the quantile Huber loss, yielding flexible, support-free approximations. Empirical evidence indicate that these methods improve sample efficiency and robustness, while enabling tail-aware training and evaluation. In RTED, learning a distribution over prospective rewards inherently supports the minimization of distorted expectations (also known as risk measures).

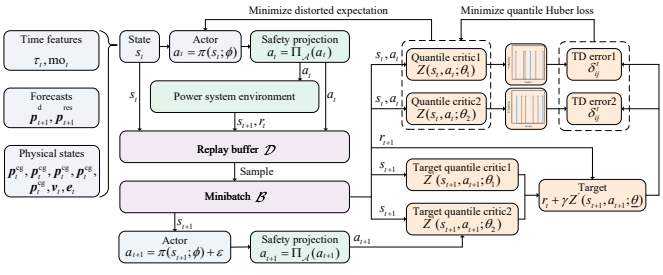

Fig. 1. Risk-aware real-time economic dispatch framework via distributional reinforcement learning.

This paper presents a risk-aware RTED framework founded on an improved Twin Delayed Deep Deterministic Policy Gradient (TD3) [11], as shown in Figure 1. The methodology involves substituting the traditional scalar critic with a quantile-based distributional critic that estimates multiple quantiles of the reward. This representation facilitates the direct evaluation

This work was supported by the National Key R&D Program of China (No. 2022YFB2404000).

of distorted expectations and supports the refinement of risk-averse policies and decision-making processes. To promote feasibility under operational constraints, a safety projection process is incorporated, projecting potential actions onto the feasible set. Additionaly, the training process is augmented with penalties for any residual violations.

The main contributions of this paper are as follows: (i) We propose a distributional TD3 (DR-TD3) algorithm with distorted expectation that emphasizes adverse tail outcomes, thereby enhancing risk-aware decision-making in RTED. (ii) Constraint violations are mitigated through action projection and reward penalties. (iii) The efficacy of the proposed method is validated in comparison to RL baselines in risk-sensitive RTED tasks on benchmark systems.

II. Problem Formulation

We formulate real-time dispatch (RTD) over a finite horizon involving conventional generators (CGs), renewable energy sources (RESs), and battery storage systems (BSSs). The goal is to minimize the total operating cost—comprising CG generation costs and RES curtailment penalties—supplemented with soft penalties on voltage and current violations. This formulation is subject to AC power flow constraints, generator capacity and ramping limitations, RES availability and power-factor considerations, as well as BSS power and state of charge (SoC) dynamics.

A. Decision Variables and Operating Constraints

Let $\mathcal{T} := \{0, 1, \ldots, T-1\}$ denote the next T dispatch intervals of length ΔT. Our work encompasses the following three categories of controllable devices:

a) CGs ($i \in \mathcal{G}$): Decision variables $p_{i,t}^{\text{cg}}$ and $q_{i,t}^{\text{cg}}$ denote active and reactive power outputs of CGs. Each unit must operate within its respective capacity and ramping constraints:

$$\underline{P}_i^{\text{cg}} \le p_{i,t}^{\text{cg}} \le \overline{P}_i^{\text{cg}}, \quad i \in \mathcal{G}, t \in \mathcal{T} \tag{1}$$

$$\underline{Q}_i^{\text{cg}} \le q_{i,t}^{\text{cg}} \le \overline{Q}_i^{\text{cg}}, \quad i \in \mathcal{G}, t \in \mathcal{T} \tag{2}$$

$$-R_i^{\text{dn}} \le p_{i,t+1}^{\text{cg}} - p_{i,t}^{\text{cg}} \le R_i^{\text{up}}, \quad i \in \mathcal{G}, t \in \mathcal{T}\backslash\{T-1\} \tag{3}$$

where $\underline{P}_i^{\text{cg}}$ and $\overline{P}_i^{\text{cg}}$ denote the minimum and maximum active power outputs; $\underline{Q}_i^{\text{cg}}$ and $\overline{Q}_i^{\text{cg}}$ are the minimum and maximum reactive power outputs; and R_i^{dn} and R_i^{up} are the ramp-down and ramp-up limits of generator i.

b) RESs ($i \in \mathcal{R}$): Decision variables $p_{i,t}^{\text{res}}$ and $q_{i,t}^{\text{res}}$ represent the active and reactive power outputs of RESs. The active power output is restricted by the instantaneous available renewable energy power and the equipment capacity, whereas the reactive power output is limited by the minimum power factor.

$$0 \le p_{i,t}^{\text{res}} \le \min(\overline{P}_i^{\text{res}}, P_{i,t}^{\text{av}}), \quad i \in \mathcal{R}, t \in \mathcal{T} \tag{4}$$

$$|q_{i,t}^{\text{res}}| \le \tan(\overline{\varphi}_i) p_{i,t}^{\text{res}}, \quad i \in \mathcal{R}, t \in \mathcal{T} \tag{5}$$

where $\overline{P}_i^{\text{res}}$ is the maximum active power output; $P_{i,t}^{\text{av}}$ is the available active power; and $\cos(\overline{\varphi}_i)$ indicates the minimum power factor for RES i.

c) BSSs ($i \in \mathcal{S}$): Decision variables comprise the output power $p_{i,t}^{\text{bss}}$, which is constrained by the charge and discharge power limits. The SoC $e_{i,t}$ must reside within the minimum and maximum energy capacity bounds. Its evolution over time is dictated by the charge/discharge power and efficiency.

$$-\overline{P}_i^{\text{ch}} \le p_{i,t}^{\text{bss}} \le \overline{P}_i^{\text{dis}}, \quad i \in \mathcal{S}, t \in \mathcal{T} \tag{6}$$

$$\underline{E}_i \le e_{i,t}^{\text{bss}} \le \overline{E}_i, \quad i \in \mathcal{S}, t \in \mathcal{T} \tag{7}$$

$$e_{i,t+1}^{\text{bss}} = \begin{cases} e_{i,t}^{\text{bss}} - \dfrac{p_{i,t}^{\text{bss}}}{\eta_i^{\text{dis}}} \Delta T, & p_{i,t}^{\text{bss}} \ge 0, \\ e_{i,t}^{\text{bss}} - \eta_i^{\text{ch}} p_{i,t}^{\text{bss}} \Delta T, & p_{i,t}^{\text{bss}} < 0, \end{cases} \quad i \in \mathcal{S},\ t \in \mathcal{T}\backslash T \tag{8}$$

where $\overline{P}_i^{\text{ch}}$ and $\overline{P}_i^{\text{dis}}$ denote the maximum charge and discharge powers, respectively; $\underline{E}_i$ and $\overline{E}_i$ are the minimum and maximum energy capacities; and η_i^{ch} and η_i^{dis} are the charge and discharge efficiency of BSS i, respectively.

Beyond device constraints, the system should satisfy the AC power flow equations and network limits. Let $\mathcal{B}$ and $\mathcal{E}$ denote the sets of buses and branches, respectively. The voltage magnitude $v_{i,t}$ at each bus i should satisfy operational limits:

$$\underline{V}_i \le v_{i,t} \le \overline{V}_i, \quad i \in \mathcal{B}, t \in \mathcal{T} \tag{9}$$

The current flow $I_{ij,t}$ on each branch (i,j) should remain within thermal limits:

$$|I_{ij,t}| \le \overline{I}_{ij}, \quad (i,j) \in \mathcal{E}, t \in \mathcal{T} \tag{10}$$

B. MDP Formulation

We cast real-time dispatch over the finite horizon $\mathcal{T}$ as a standard Markov decision process (MDP), characterized by the tuple $\langle \mathcal{S}, \mathcal{A}, R, P, \gamma \rangle$. The sets $\mathcal{S}$ and $\mathcal{A}$ denote continuous state and action spaces, respectively. The state-transition dynamics are governed by an unknown probability density $P : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, \infty)$, which specifies the density of the next state given a current state-action pair. The reward function is given by $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and $\gamma \in (0, 1]$ is the discount factor.

The state s_t at time t includes: (i) temporal features, including the time index of the day τ_t and the month index mo_t; (ii) one-step-ahead forecasts of load $\hat{\boldsymbol{p}}_{t+1}^{\text{d}}$ and renewable generation $\hat{\boldsymbol{p}}_{t+1}^{\text{res}}$; (iii) current outputs of CGs $\boldsymbol{p}_t^{\text{cg}}$ & $\boldsymbol{q}_t^{\text{cg}}$, RESs $\boldsymbol{p}_t^{\text{res}}$ & $\boldsymbol{q}_t^{\text{res}}$, and BSSs $\boldsymbol{p}_t^{\text{bss}}$; (iv) current bus voltage magnitudes $\boldsymbol{v}_t$; (v) current SoC of BSSs $\boldsymbol{e}_t$. The vector representation can be expressed as:

$$s_t = [\tau_t, \text{mo}_t, \hat{\boldsymbol{p}}_{t+1}^{\text{d}}, \hat{\boldsymbol{p}}_{t+1}^{\text{res}}, \boldsymbol{p}_t^{\text{cg}}, \boldsymbol{q}_t^{\text{cg}}, \boldsymbol{p}_t^{\text{res}}, \boldsymbol{p}_t^{\text{res}}, \boldsymbol{p}_t^{\text{bss}}, \boldsymbol{v}_t, \boldsymbol{e}_t]$$

The action a_t at time t includes: (i) active and reactive power setpoints of CGs $\boldsymbol{p}_{t+1}^{\text{cg}}$, $\boldsymbol{q}_{t+1}^{\text{cg}}$; (ii) active and reactive power setpoints of RESs $\boldsymbol{p}_{t+1}^{\text{res}}$, $\boldsymbol{q}_{t+1}^{\text{res}}$; (iii) active power setpoints of BSSs $\boldsymbol{p}_{t+1}^{\text{bss}}$. The vector representation can be expressed as:

$$a_t = [\boldsymbol{p}_{t+1}^{\text{cg}}, \boldsymbol{q}_{t+1}^{\text{cg}}, \boldsymbol{p}_{t+1}^{\text{res}}, \boldsymbol{q}_{t+1}^{\text{res}}, \boldsymbol{p}_{t+1}^{\text{bss}}]$$

The reward r_t at time t is defined as the negative of the total operating cost, which includes generation costs of CGs

and penalty costs for RES curtailment, augmented with soft penalties for voltage violations:

$$r_t = -\lambda_C \left(\sum_{i\in\mathcal{G}} C_i\big(p^{\mathrm{cg}}_{i,t}\big) + \sum_{i\in\mathcal{R}} C^{\mathrm{w}}_i\big(P^{\mathrm{av}}_{i,t} - p^{\mathrm{res}}_{i,t}\big) \right) - \lambda_V \mathcal{P}_V(\boldsymbol{v}_t)$$

$$\mathcal{P}_V(\boldsymbol{v}_t) = \sum_{i\in\mathcal{B}} \Big([\, v_{i,t} - \overline{V}_i \,]^2_+ + [\, \underline{V}_i - v_{i,t} \,]^2_+ \Big)$$

where $[x]_+ := \max(0, x)$, $C_i(\cdot)$ is the generation cost of CG i, and $C^{\mathrm{w}}_i(\cdot)$ is the curtailment penalty for RES i. Here $\mathcal{P}_V(\boldsymbol{v}_t)$ is a nonnegative penalty functional that measures the aggregated violations of voltage magnitude limits at time t, and $\lambda_C, \lambda_V \geq 0$ is the weighting coefficient.

III. Distributional TD3 for Risk-Aware RTED

A. Risk-aware Dispatch Objective

A policy $\pi : \mathcal{S} \to \mathcal{A}$ is a mapping from each state to an action. The set of all policies is denoted by Π. For a policy π, the action-value distribution from state-action (s_t, a_t) is defined as follows:

$$Z^\pi(s_t, a_t) = \sum_{k=0}^{T-1} \gamma^k r_{t+k}(s_{t+k}, a_{t+k}),$$
$$a_{t+k} \sim \pi(\cdot|s_{t+k}),\ \ s_{t+k+1} \sim P(\cdot|s_{t+k}, a_{t+k}) \quad (11)$$

which induces a full distribution over future cumulative costs due to source/load uncertainty and exogenous forecast errors. Classical risk-neutral objectives minimize the expectation $\mathbb{E}[Z^\pi(s,a)]$, which can cause poor performance in certain scenarios and weaken the robustness of dispatch policies.

To explicitly model tail risk, we adopt distorted expectations of the value distribution. Let F_Z be the cumulative distribution function (CDF) of $Z^\pi(s,a)$ and $F_Z^{-1}(\tau)$ its quantile function. A distortion function $\psi : [0,1] \to [0,1]$ is increasing with $\psi(0) = 0, \psi(1) = 1$. The distorted expectation is

$$\mathbb{D}_\psi[Z^\pi(s,a)] = \int_0^1 F_Z^{-1}(\tau)\, \mathrm{d}\psi(\tau) = \int_0^1 w(\tau)\, F_Z^{-1}(\tau) \mathrm{d}\tau \quad (12)$$

where $w(\tau) := \psi'(\tau)$ denotes the implied quantile weight. The risk-neutral case corresponds to $\psi(\tau) = \tau$.

An important and practical choice is the Wang risk measure [12], defined via the distortion $\psi_\beta(u) = \Phi(\Phi^{-1}(u) + \beta)$ with parameter $\beta \in \mathbb{R}$ and standard normal CDF Φ, where the implied quantile weight is $w_\beta(\tau) = \dfrac{\phi(\Phi^{-1}(\tau) + \beta)}{\phi(\Phi^{-1}(\tau))}$ with ϕ the standard normal probability density function (PDF). For rewards, as shown in Figure 2, $\beta > 0$ over-weights low quantiles (risk-averse), while $\beta < 0$ under-weights them (risk-seeking); $\beta = 0$ reduces to the expectation.

Our risk-aware dispatch objective is to learn a policy π that minimizes a distorted expectation of the action-value distribution:

$$J_\psi(\pi) = \mathbb{E}_{s_t\sim\mathcal{D}}\Big[\mathbb{D}_\psi\big[Z^\pi(s_t, \pi(s_t))\big]\Big] \quad (13)$$

where $\mathcal{D}$ represents the transitions replay buffer.

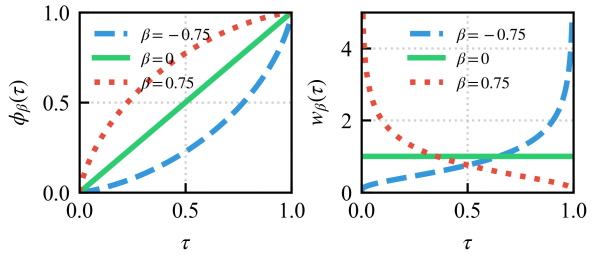

Fig. 2. Distortion functions ψ_β and implied quantile weights w_β for the Wang risk measure with different β.

B. Quantile-based Distributional Critic

To optimize the risk-aware dispatch objective $J_\psi(\pi)$, we approximate the action-value distribution with a quantile critic. Let $\{\tau_n\}_{n=1}^N$ be a uniformly spaced set of quantile levels. The critic $Z(s,a;\theta)$ parameterizes $Z^\pi(x,a)$ and provides the estimated quantiles $\{Z_{\tau_n}(s,a;\theta)\}_{n=1}^N$.

We employ twin target critics with delayed policy updates, as exemplified in TD3 [11], extended to the distributional framework. Two quantile critics, $Z(s,a;\theta_1)$ and $Z(s,a;\theta_2)$, are maintained. The temporal difference error is calculated using target networks:

$$\delta^t_{ij}(\theta) = r_t + \gamma Z'_{\tau_i}(s_{t+1}, a_{t+1}; \underline{\theta}) - Z_{\tau_j}(s_t, a_t; \theta) \quad (14)$$

where $Z'(s,a;\theta)$ is the target critic network; $\underline{\theta}$ are the parameters of the critic with a lower distorted expectation; $a_{t+1} = \pi(s_{t+1};\phi)$ is the action from the current policy. The next action is computed as:

$$\tilde{a}_{t+1} = \pi'(s_{t+1};\phi) + \epsilon_\pi, \quad \epsilon_\pi \sim \mathrm{clip}\big(\mathcal{N}(0,\sigma^2), -c, c\big) \quad (15)$$

$$a_{t+1} = \Pi_{\mathcal{A}}(\tilde{a}_{t+1}) \quad (16)$$

where $\pi'(s;\phi)$ is the target policy network, ϵ_π is clipped Gaussian noise to smooth the target policy, and $\Pi_{\mathcal{A}}(\cdot)$ is a projection operator that clips the action to satisfy the device constraints (1)-(8) at execution.

Given a replay batch $\mathcal{B}$ of transitions, the critic is trained with the quantile Huber loss:

$$\mathcal{L}(\theta) = \sum_{\mathcal{B}} \sum_{i=1}^{N} \sum_{j=1}^{N} (\tau_{i+1} - \tau_i) \rho^\kappa_{\tau_n}(\delta_{ij}(\theta)) \quad (17)$$

$$\rho^\kappa_\tau(u) = |\tau - \mathbb{I}\{u<0\}| \cdot \begin{cases} \frac{1}{2}u^2, & |u| \leq \kappa \\ \kappa(|u| - \frac{1}{2}\kappa), & |u| > \kappa \end{cases} \quad (18)$$

where $\kappa > 0$ is the Huber threshold and $\mathbb{I}\{\cdot\}$ is the indicator function. This loss represents the quantile regression error between the predicted and the target quantiles.

The target critic network is slowly updated by:

$$\theta'_i \leftarrow \alpha\theta_i + (1-\alpha)\theta'_i, \quad i = 1,2 \quad (19)$$

where $\alpha \in (0,1)$ is the soft update rate.

Algorithm 1 DR-TD3 for Risk-aware RTED

Initialize critic networks $Z(s,a;\theta_1)$, $Z(s,a;\theta_2)$, and policy network $\pi(s;\phi)$ with random parameters θ_1, θ_2, ϕ
Initialize target networks with weights $\theta_1' \leftarrow \theta_1$, $\theta_2' \leftarrow \theta_2$, $\phi' \leftarrow \phi$
Initialize replay buffer $\mathcal{D}$
Choose quantile levels $\{\tau_k\}_{k=1}^N$ and Wang risk weights $w_k = w_\beta(\tau_k)$; set $\gamma, \alpha, d_\pi, \sigma, c, \kappa$, batch size B
for each episode **do**
  Randomly sample initial state s_0
  for $t = 0, 1, \ldots, T-1$ **do**
    Execute exploration action a_t, observe r_t, s_{t+1}; store (s_t, a_t, r_t, s_{t+1}) in $\mathcal{D}$
    Sample minibatch $\mathcal{B} = \{(s,a,r,s')\}_{b=1}^B$ from $\mathcal{D}$
    Target smoothing: $\tilde{a}' = \pi'(s';\phi') + \epsilon_\pi$, with $\epsilon_\pi \sim \text{clip}(\mathcal{N}(0,\sigma^2), -c, c)$
    Safety projection: $a' = \Pi_{\mathcal{A}}(\tilde{a}')$
    for $i = 0, 1, \ldots, N-1$ **do**
      for $j = 0, 1, \ldots, N-1$ **do**
        $\delta_{ij}^t = r_t + \gamma Z'_{\tau_i}(s_{t+1}, a_{t+1}; \underline{\theta}) - Z_{\tau_j}(s_t, a_t; \theta)$
      end for
    end for
    Update θ_i with $\nabla_{\theta_k}\mathcal{L}(\theta_k)$, $k = 1, 2$
    if t mod $d_\pi = 0$ **then**
      Update ϕ with $\nabla_\phi J(\phi)$
      $\theta_i' \leftarrow \alpha\theta_i + (1-\alpha)\theta_i', \quad i = 1, 2;$
      $\phi' \leftarrow \alpha\phi + (1-\alpha)\phi'$
    end if
  end for
end for

C. Risk-aware Policy Update and Action Selection

Given quantile critics $Z_{\tau_k}(s,a;\theta_1)$ and $Z_{\tau_k}(s,a;\theta_2)$ for levels $\{\tau_n\}_{n=1}^N$, the actor $\pi(s;\phi)$ is updated to minimize a distorted expectation of the action-value distribution as estimated by one of the twin critics:

$$J(\phi) = \mathbb{E}_{s_t \sim \mathcal{D}}\left[\sum_{k=1}^{N} w_\beta(\tau_k) Z_{\tau_k}(s_t, a; \theta_1)\right] \tag{20}$$

Using the chain rule, a practical policy gradient for (20) is

$$\nabla_\phi J(\phi) = \mathbb{E}_{s_t \sim \mathcal{D}}\left[\sum_{k=1}^{N} w_\beta(\tau_k) \nabla_a Z_{\tau_k}(s_t, a; \theta_1) \nabla_\phi \pi(s_t; \phi)\right] \tag{21}$$

To stabilize the learning process, the policy is updated following every d_π critic update steps. The target policy network is also slowly updated by:

$$\phi' \leftarrow \alpha\phi + (1-\alpha)\phi' \tag{22}$$

The overall algorithm is summarized in Algorithm 1.

IV. Case Studies

A. Experimental Setup

The proposed method is validated on a modified IEEE 33-bus distribution system, as shown in Figure 3. The system comprises one gas turbine, four wind turbines, and two battery storage systems. Hyperparameters are listed in Table I.

To simulate the uncertainties of wind power and load, we utilize historical data from [13], [14]. At the start of each training episode, we randomly sample a starting day from the corresponding month for the load and wind generation profiles. The simulation then proceeds by sequentially reading from these time-series datasets. Each episode has a maximum length of 200 steps, where each step represents a 15-minute interval.

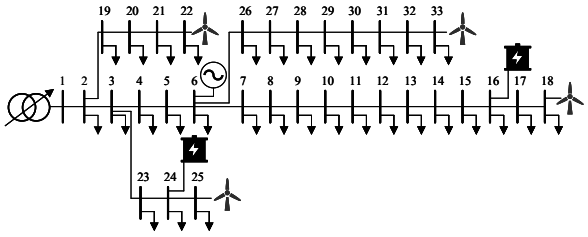

Fig. 3. Topology of the modified IEEE 33-bus system.

TABLE I
Hyperparameter Settings

Parameter	Value	Parameter	Value
Learning rate	3e-4	Update rate (α)	0.005
Smoothing noise (σ)	0.2	Policy delay (d_π)	2
Smoothing noise clip (c)	0.5	Cost coefficient (λ_C)	0.5
Batch size (B)	256	Penalty coefficient (λ_V)	0.1
Number of quantiles (N)	32	State dimension	156
Discount factor (γ)	0.99	Action dimension	14

B. Results and Discussion

To validate the efficacy of the proposed method, a comparative analysis is conducted against two RL baselines: TD3 and SAC. Meanwhile, the DR-TD3 agent is configured with four distinct risk-aversion levels, corresponding to β values of 0, 0.5, 0.75, and 0.95. The performance of all methods is evaluated on a fixed test set comprising 2000 episodes.

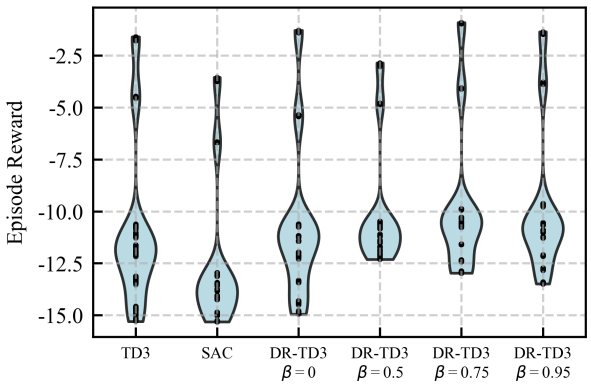

Fig. 4. Comparison of episode reward distributions with different methods.

Figure 4 presents the resulting distributions of episode rewards for each configuration. Even risk-neutral DR-TD3

with $\beta = 0$ demonstrates superior performance compared to TD3 and SAC. This improvement is attributable to its use of distributional information, which facilitates a more accurate training of the critic network.

Upon introducing the distorted expectation objective, a significant enhancement in tail-end performance is observed. As β increases, greater weight is placed on the lower quantiles of the return distribution. This incentivizes the agent to prioritize improving less favorable outcomes, effectively shifting the learned return distribution toward higher values. Consequently, the agent with $\beta = 0.75$ achieves better average performance than the one with $\beta = 0.5$.

However, an excessively high β can be detrimental. Excessive emphasis on low-reward outcomes steers the policy toward overly conservative updates, sacrificing high-reward opportunities and effectively overfitting to tail events, which ultimately degrades performance relative to the more balanced setting $\beta = 0.75$. This is also evident from the training curves in Figure 5. The agent with $\beta = 0.95$ achieves the highest reward during training, yet generalizes worse on the test set, suggesting overfitting. It can also be observed that TD3, SAC, and the risk-neutral DR-TD3 exhibit weaker performance throughout the training process, which explains their relatively poor performance on the test set.

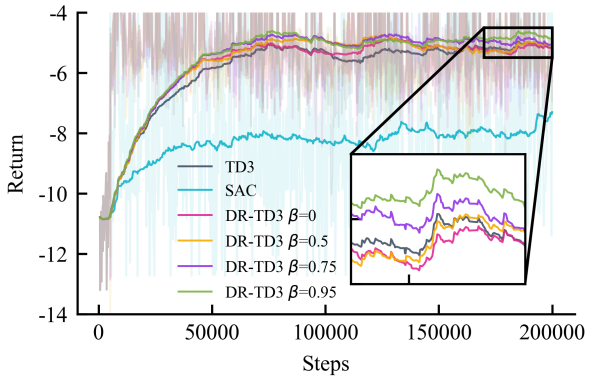

Fig. 5. Training curve of different methods.

To further investigate the impact of incorporating the soft penalties for voltage violations, we conduct an ablation study under the $\beta = 0.75$ configuration. As shown in Figure 6, enabling the voltage penalty substantially reduces the severity of voltage violations compared to its counterpart. It empirically validates the effectiveness of the proposed method in enhancing operational safety and maintaining system stability.

V. Conclusion

This paper introduced a novel risk-aware framework for real-time economic dispatch, leveraging an advanced distributional reinforcement learning agent, DR-TD3. Experimental results demonstrated that our distributional approach significantly outperforms standard RL baselines. Furthermore, the introduction of a risk-averse objective led to superior performance in managing tail-end events. An ablation study

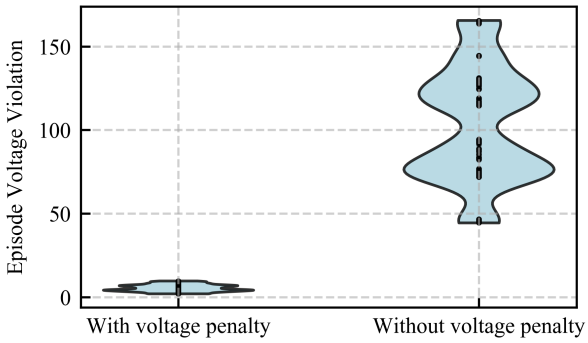

Fig. 6. Comparison of episode voltage violation distributions.

also empirically confirmed the effectiveness of the soft reward penalties, which substantially reduced the severity of voltage constraint violations. These findings underscore the value of integrating distributional perspectives and explicit safety mechanisms for developing robust, risk-sensitive control policies in uncertain power system environments.

References

[1] R. Lu, T. Ding, B. Qin, J. Ma, X. Fang, and Z. Dong, "Multi-stage stochastic programming to joint economic dispatch for energy and reserve with uncertain renewable energy," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 3, pp. 1140–1151, 2019.

[2] R. Jiang, J. Wang, and Y. Guan, "Robust unit commitment with wind power and pumped storage hydro," *IEEE Transactions on Power Systems*, vol. 27, no. 2, pp. 800–810, 2011.

[3] C. Ning and F. You, "Optimization under uncertainty in the era of big data and deep learning: When machine learning meets mathematical programming," *Computers & Chemical Engineering*, vol. 125, pp. 434–448, 2019.

[4] H. Li and H. He, "Learning to operate distribution networks with safe deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 13, no. 3, pp. 1860–1872, 2022.

[5] Z. Yan and Y. Xu, "Real-time optimal power flow: A lagrangian based deep reinforcement learning approach," *IEEE Transactions on Power Systems*, vol. 35, no. 4, pp. 3270–3273, 2020.

[6] T. Wu, A. Scaglione, and D. Arnold, "Constrained reinforcement learning for predictive control in real-time stochastic dynamic optimal power flow," *IEEE Transactions on Power Systems*, vol. 39, no. 3, pp. 5077–5090, 2023.

[7] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *International conference on machine learning*. PMLR, 2017, pp. 449–458.

[8] W. Dabney, M. Rowland, M. Bellemare, and R. Munos, "Distributional reinforcement learning with quantile regression," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[9] W. Dabney, G. Ostrovski, D. Silver, and R. Munos, "Implicit quantile networks for distributional reinforcement learning," in *International conference on machine learning*. PMLR, 2018, pp. 1096–1105.

[10] X. Ma, J. Chen, L. Xia, J. Yang, Q. Zhao, and Z. Zhou, "Dsac: Distributional soft actor-critic for risk-sensitive reinforcement learning," *Journal of Artificial Intelligence Research*, vol. 83, 2025.

[11] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International conference on machine learning*. PMLR, 2018, pp. 1587–1596.

[12] S. S. Wang, "A class of distortion operators for pricing financial and insurance risks," *Journal of risk and insurance*, pp. 15–36, 2000.

[13] "Wind power production estimation and forecast on belgian grid (historical)," https://www.elia.be/en/grid-data/generation-data/wind-power-generation, 2025.

[14] "Wind power production estimation and forecast on belgian grid (historical)," https://www.elia.be/en/grid-data/load-and-load-forecasts, 2025.