

Article

Emergency Load-Shedding Decision for Frequency Stability of New Energy Power System Based on Constrained Markov Decision Process

Qiushi Fang ¹, Zhentao Han ¹, Wenhui He ^{2,*}, Yufei Jin ¹, Zewei Li ², Mingxuan Lu ¹, Weihan Chen ², Jiawen Gao ¹ and Rui Zhang ² 

¹ Economic and Technological Research Institute, State Grid Liaoning Electric Power Co., Ltd., Shenyang 110016, China

² School of Electrical Engineering, Northeast Electric Power University, Jilin 132012, China

* Correspondence: 1202400031@neepu.edu.cn

Abstract

Renewable energy systems dominated by powered electronic devices generally exhibit weak disturbance tolerance and limited grid-support capability. Following the blocking of a flexible DC transmission system, emergency load shedding in renewable-rich grid regions may induce overvoltage or undervoltage at the point of common coupling, forcing renewable energy units into a voltage ride-through state. This, in turn, reduces their active power output and threatens the frequency stability of the power system. To address this issue, this paper proposes an emergency load-shedding decision model based on a constrained Markov decision process (CMDP). First, an emergency frequency control model for AC–DC hybrid power systems is established within the Markov decision process framework, thereby formulating power system frequency stability control as a Markov decision problem. Second, Lagrange multipliers are introduced into the CMDP framework to transform the constrained optimization problem with security constraints into an unconstrained objective optimization problem. Finally, the Proximal Policy Optimization (PPO) algorithm is adopted to accelerate the training process and improve the decision accuracy of the intelligent agent. The simulation results, based on the modified IEEE 39-bus system, demonstrate that, compared with the traditional contingency strategy and the conventional Markov decision algorithm, the proposed load-shedding strategy can satisfy system frequency stability requirements, effectively avoid voltage violations at renewable energy grid-connection points, and minimize the total load shedding amount.

Keywords: frequency stability; voltage security control; constrained Markov decision process; emergency; emergency load shedding



Academic Editor: José Matas

Received: 8 May 2026

Revised: 4 June 2026

Accepted: 5 June 2026

Published: 26 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

With the continuous growth of electricity demand, the large-scale integration of renewable energy and long-distance cross-regional power transmission via high-voltage direct current (HVDC) systems has been widely adopted in modern power systems, providing strong support for the ongoing transformation of industrial structures and energy systems. However, internal converter faults may lead to the blocking of voltage-source-converter-based HVDC systems, causing the transmitted active power to drop to zero within an extremely short time [1,2]. Under such circumstances, the resulting power imbalance, together with the weak disturbance tolerance and limited grid-support capability

of power-electronic-interfaced devices, may cause load-shedding actions in regions with high renewable energy penetration to induce voltage violations at renewable energy grid-connection points. This, in turn, poses a serious threat to the secure integration of renewable energy and the frequency stability of the power system.

Existing research on emergency load shedding optimization can generally be classified into two main categories: classical mathematical-programming-based methods and heuristic optimization approaches [3]. Both categories offer distinct advantages, but they also exhibit inherent limitations. Classical mathematical programming methods typically formulate optimization models with economic efficiency, system security, and stability as the primary objectives, and then derive load-shedding schemes that satisfy the corresponding security and stability constraints. Representative models include the average system frequency (ASF) model [4] and the system frequency response (SFR) model [5].

However, classical optimization methods often struggle to capture the increasingly complex, nonlinear, and time-varying operating characteristics of modern power systems with high renewable energy penetration. Consequently, extracting the dynamic evolution characteristics of key post-fault electrical variables and developing emergency load-shedding strategies based on heuristic optimization algorithms have become effective approaches for addressing these challenges and enhancing the secure and stable operation of power systems.

Heuristic algorithms are generally used to solve nondeterministic polynomial time-optimization problems. By leveraging intuition, search rules, or empirical knowledge, they can provide feasible solutions to combinatorial optimization problems within an acceptable computational cost, and are therefore particularly suitable for complex optimization tasks. Typical heuristic approaches include particle swarm optimization, simulated annealing, random forests, and neural networks.

In Ref. [6], a cultural particle swarm optimization algorithm was employed to optimize an emergency load-shedding model under multiple operating conditions, thereby reducing the risk of convergence to local optima. In Ref. [7], a dual-driven load-shedding strategy based on event-driven and response-driven mechanisms was proposed. In that method, the load shedding optimization model of the main grid was used to determine the required shedding amount for each distribution network under emergency scenarios, and the resulting solutions were stored in an offline database for the event-driven module. Meanwhile, a random-forest model trained on the offline database was applied in the response-driven module. These methods can effectively satisfy post-fault power system stability requirements. However, they pay limited attention to the optimization of load shedding amounts, making it difficult to derive truly optimal load-shedding schemes. Moreover, the large-scale integration of renewable energy has made modern power systems increasingly high-dimensional and nonlinear, leading to a continuous increase in computational burden and solution complexity. Deep reinforcement learning provides a promising approach for addressing these challenges.

Deep reinforcement learning (DRL), which builds upon conventional reinforcement learning and has been enabled by the rapid development of deep learning models and computing hardware, can improve decision-making efficiency and accuracy while overcoming the limitations of traditional algorithms in handling problems with large-scale or even continuous state and action spaces. Reinforcement learning problems are commonly formulated as Markov decision processes (MDPs). Once an MDP is established, it can be solved using dynamic programming methods, such as value iteration and policy iteration, or policy optimization methods, such as policy gradient algorithms. Compared with conventional reinforcement learning approaches, the key feature of DRL is the introduction of deep neural networks to approximate the mappings associated with states, actions, rewards, and

value functions in an MDP. As a result, DRL is well suited to addressing complex power system problems characterized by high dimensionality and strong nonlinearity, while learning effective control policies through repeated interactions with the environment.

In Ref. [8], a Deep Q-Network (DQN)-based method was trained using large-scale historical data, in which load shedding cost was incorporated to derive an optimal load-reduction strategy. In Ref. [9], multi-Q-learning was applied to alleviate the high-dimensionality problem in power systems and to develop emergency load-shedding strategies for frequency stability under different operating scenarios. In Ref. [10], neural networks were employed to extract high-dimensional and nonlinear post-fault features of power systems, and a load-shedding penalty term was incorporated into the DQN reward function to obtain an optimal load-shedding strategy that could satisfy system stability requirements. However, most existing studies primarily focus on minimizing the amount of post-fault load shedding, while paying insufficient attention to the fact that inappropriate load-shedding actions may drive renewable energy units into voltage ride-through conditions, thereby further endangering system frequency stability.

More recent studies have further demonstrated the potential of reinforcement-learning-based emergency load shedding in renewable-energy-integrated power systems. Chen et al. proposed a DRL-based emergency load-shedding strategy for power systems with high renewable energy penetration and adopted PPO to improve the adaptability and efficiency of an MDP-based decision-making model under multiple-fault and variable-operation scenarios [11]. Hu et al. introduced knowledge-enhanced DRL into event-based load shedding, in which a knowledge-enhanced MDP and a linear decision space were designed to reduce the training difficulty caused by high-dimensional action spaces [12]. Song et al. applied PPO-based DRL to remedial action schemes for VSC-HVDC systems, demonstrating improved training stability and voltage-regulation performance under renewable energy uncertainty [13]. More recently, Li et al. combined model- and data-driven reinforcement learning methods for under-frequency load shedding control, where improved reinforcement learning strategies were used to accelerate convergence and enhance training stability [14]. These studies indicate that reinforcement learning has become an effective and promising tool for emergency control and load-shedding decision making in modern power systems.

Recent RL-based emergency control studies have shown that reinforcement learning can improve the adaptability of emergency control strategies under complex operating conditions. Huang et al. developed adaptive emergency control schemes for power systems using deep reinforcement learning and demonstrated their effectiveness in dynamic braking and undervoltage load-shedding applications. Chen et al. proposed a DRL-based emergency load-shedding strategy for power systems with high renewable energy penetration; they adopted PPO to improve the adaptability and efficiency of MDP-based decision making under multiple fault and operating scenarios [15]. Xie and Sun formulated emergency frequency control as an MDP and solved it using distributional deep reinforcement learning, thereby demonstrating the potential of DRL for emergency frequency regulation.

However, standard RL-based emergency control methods typically optimize the expected reward and do not explicitly guarantee the satisfaction of operational security constraints during either policy learning or policy execution. As a result, safe reinforcement learning has attracted increasing attention in power system control. Vu et al. introduced safe reinforcement learning into emergency load shedding to enhance post-fault voltage recovery [16], and further incorporated barrier functions into reinforcement-learning-based emergency control to suppress unsafe actions [17]. Recent review studies have also indicated that safe RL methods can be broadly classified into safety-layer-based methods and integrated-optimization-based methods [18]. Among these approaches, CMDP-based

constrained optimization is a representative formulation for explicitly incorporating safety constraints into the policy-learning process [19].

Although the above studies have reported promising results, several limitations remain. First, most existing reinforcement-learning-based emergency load-shedding methods mainly address frequency recovery, transient voltage stability, or load shedding economy from a single-control perspective. The coupled relationship among MMC-HVDC blocking, post-load-shedding voltage violations at renewable energy grid-connection points, voltage-ride-through-induced active power reduction, and the resulting secondary frequency deterioration has not been sufficiently investigated. Second, conventional MDP-based reinforcement learning typically handles security requirements through reward-penalty mechanisms. Although such treatments may statistically reduce unsafe actions, they cannot explicitly represent the decision-making process under operational constraints. Third, the impact of renewable energy voltage security constraints on emergency frequency-control decisions has rarely been incorporated into load-shedding strategy design within a unified mathematical framework.

Relevant grid-connection codes for renewable energy have established corresponding technical requirements [20,21]. Under emergency operating conditions, renewable energy units are capable of rapidly regulating their active power output and, when necessary, reducing active power generation or disconnecting from the grid. After load shedding is executed, if the terminal bus voltage at a renewable energy grid-connection point falls outside the allowable range specified by the grid-connection requirements, the corresponding renewable energy units will enter a voltage ride-through state. Under such conditions, the active power output of renewable energy units decreases, which may further deteriorate the system frequency response and even lead to frequency instability. Therefore, the voltage violation problem at renewable energy grid-connection points after load-shedding actions must be explicitly considered in emergency load-shedding strategy design.

Based on the above literature review, the proposed method is positioned between standard RL-based emergency load-shedding approaches and more complex safe RL frameworks. Compared with conventional MDP/PPO-based emergency load-shedding methods, the proposed approach does not treat renewable energy voltage violations merely as undesirable outcomes in reward evaluation. Instead, renewable energy voltage security is explicitly incorporated into the CMDP formulation as an operational safety constraint. Compared with safe RL methods that rely on additional safety layers, recovery networks, barrier certificates, or action-correction modules, the proposed method provides a simpler and more interpretable safety-constrained formulation by introducing voltage-security-related constraint costs and Lagrangian penalty terms into the PPO-based policy optimization process. Therefore, the main novelty of this work lies in formulating the MMC-HVDC-blocking-induced emergency load shedding problem as a CMDP-based safety-constrained decision-making problem, rather than in proposing a new PPO optimizer.

To address the above research gaps, this paper proposes a CMDP-based emergency load-shedding decision-making strategy for renewable-energy-integrated AC/DC power systems under MMC-HVDC blocking faults. The main contributions of this work are summarized as follows:

1. The voltage violation mechanism at renewable energy grid-connection points following MMC-HVDC blocking and emergency load shedding is analyzed, thereby revealing the coupling relationship among load-shedding actions, renewable energy voltage ride-through, active power reduction, and system frequency deterioration.
2. A constrained Markov decision process is formulated for emergency load shedding, in which frequency stability, load shedding economy, maximum load shedding limits,

and renewable energy voltage security constraints are jointly considered within a unified decision-making framework.

3. Lagrange multipliers are introduced to transform the constrained load shedding optimization problem into an unconstrained policy-optimization problem, thereby reducing the likelihood of selecting actions that violate voltage security constraints during agent training.
4. A PPO-based solution framework is developed through Python-DIGSILENT co-simulation, and the proposed CMDP-based strategy is compared with both a traditional contingency strategy and a conventional MDP-based strategy to verify its effectiveness in coordinating frequency stability, voltage security, and load shedding economy.

It should be emphasized that the main novelty of this work does not lie in developing a new reinforcement learning optimizer, but in formulating the emergency load shedding problem following MMC-HVDC blocking as a safety-constrained decision-making problem. Unlike conventional MDP/PPO-based emergency control methods, the proposed framework explicitly incorporates renewable energy voltage security into the CMDP formulation and jointly coordinates frequency stability, voltage security, and load shedding economy within a unified decision-making framework.

2. Mechanism of Renewable Energy Voltage Violations Under DC Blocking and Load Shedding

This section investigates the voltage violation problem induced by inappropriate load-shedding actions following MMC-HVDC blocking. First, the causes of the weakened regional reactive power-regulation capability after MMC-HVDC blocking are analyzed. Then, the post-load-shedding nodal voltage recovery mechanism is further examined based on the nodal voltage formulation. This analysis clarifies that improper load-shedding actions following MMC-HVDC blocking may cause the voltages at renewable energy grid-connection points to deviate beyond the normal operating range, either exceeding the upper limit or falling below the lower limit. Simulation results further verify that, after the implementation of unreasonable load-shedding strategies, the voltage levels at renewable energy grid-connection points may violate the grid-connection technical requirements specified in national standards, thereby reducing renewable energy output and further deteriorating system frequency stability.

To clarify the theoretical analysis, the following assumptions are adopted in this section. First, the analysis focuses on the short-term electromechanical response following MMC-HVDC blocking, during which the network topology is assumed to remain unchanged except for the execution of load-shedding actions. Second, the load at each shedding node is represented by equivalent active and reactive power demands, and the active and reactive components are assumed to be reduced proportionally according to the prescribed load-shedding depth. Third, once the MMC-HVDC blocking fault occurs, the active power transmitted by the DC system is assumed to be fully interrupted, while the reactive current support provided by the converter is assumed to become unavailable. Fourth, renewable energy units are characterized by their terminal bus voltages and active power outputs, and their voltage ride-through behavior is evaluated according to the grid-connection voltage limits. Under these assumptions, the influence of load shedding locations and load-shedding depths on renewable energy grid-connection voltages can be analyzed through the post-fault power flow redistribution and nodal voltage recovery process.

It should be noted that the voltage security evaluation in this paper is focused on the electromechanical timescale emergency load shedding problem following MMC-HVDC blocking. The terminal voltages of renewable energy units are adopted as the primary

operational security indicators because they directly determine whether renewable energy units enter the voltage ride-through state in accordance with grid-connection requirements. Accordingly, the proposed model is intended to identify and avoid load-shedding actions that may trigger terminal voltage violations and the resulting reduction in renewable energy active power output. Detailed converter inner-loop dynamics, phase-locked-loop dynamics, current-limiting control, and interarea oscillation modes are not explicitly represented in the present framework. These issues require more detailed electromagnetic transient models or hybrid electromechanical–electromagnetic models and will therefore be investigated in future work.

2.1. AC-Side Voltage Support Mechanism of MMC-HVDC Transmission

When a system fault occurs, the MMC-HVDC transmission system is capable of providing voltage support, and the voltage at the point of common coupling (PCC) is a key variable characterizing the interaction between the MMC-HVDC system and the AC grid [22,23]. Following the fault, the MMC-HVDC system injects reactive current into the AC network to support the system voltage [24]. According to the current–balance relationship at the point of common coupling, as expressed in Equation (1):

$$\frac{U_s - U_c}{Z_{eq}} + I_{MMC} = \frac{U_c}{R_f} \quad (1)$$

where u_c denotes the voltage at the point of common coupling (PCC), u_s represents the equivalent voltage source of the AC grid, z_{eq} is the equivalent impedance, R_f denotes the fault resistance, and I_{MMC} represents the reactive current injected by the MMC-HVDC system. Accordingly, the PCC voltage can be obtained, as shown in Equation (2):

$$U_c = \frac{R_f (U_s + Z_{eq} I_{MMC})}{Z_{eq} + R_f} \quad (2)$$

When voltage fluctuations or voltage sags occur in the AC grid, the MMC-HVDC system rapidly injects reactive current into the AC network to maintain the stability of the AC bus voltage. However, after a DC blocking fault, the MMC-HVDC support current I_{MMC} becomes zero [25], and the MMC-HVDC system no longer participates in reactive power or voltage support for the grid.

When an MMC-HVDC converter is connected to the AC grid, a converter reactor is typically connected in series between the converter and the grid, with its reactance denoted as X . Assuming that the magnitude of the AC voltage output by the converter is u_c , the magnitude of the AC grid bus voltage is u_s , and the phase angle difference between them is δ , the active power P and reactive power Q injected from the converter into the grid can be approximately expressed according to the fundamental theory of power systems as follows [26]:

$$P = \frac{U_c U_s}{X} \sin \delta \quad (3)$$

$$Q = \frac{U_s (U_c \cos \delta - U_s)}{X} \sin \delta \quad (4)$$

Active power control is primarily governed by the phase angle difference δ ; meanwhile, under the small-angle approximation where $\delta \approx 1$, the reactive power is mainly determined by the voltage magnitude difference ($U_c - U_s$). Therefore, when the voltage U_s at the MMC-HVDC grid-connection point is lower than U_c , the converter delivers reactive power to the grid, exhibiting capacitive characteristics and thereby increasing the grid voltage. Conversely, when U_s is higher than U_c , the converter absorbs reactive power from the grid, exhibiting inductive characteristics and consequently reducing the grid voltage.

In regions with high renewable energy penetration, the voltage support capability is relatively insufficient. When the system is subjected to a large active power deficit fault, particularly when the disturbance occurs near large-scale renewable energy grid-connection points, the overall system voltage level decreases. Due to the relatively weak transient voltage support capability in renewable-energy-dominated areas, the voltage at these locations tends to decline more rapidly, resulting in severe undervoltage at the busbars of renewable energy grid-connection points.

2.2. DC-Side Voltage Support Mechanism of MMC-HVDC Transmission

The balance between the input and output power of generators is closely related to the active power balance of the power system. During system operation, the total active power generated by power plants, $\sum P_{Gi}$, is equal at all times to the total active power consumed within the system, $\sum P_{Di}$, namely:

$$\sum P_{Gi} - \sum P_{Di} = \sum P_{Gi} - (\sum P_{Li} + \sum P_{\text{loss}}) \quad (5)$$

where the active power consumed by the system mainly includes the sum of the active power consumed by the loads in the system, $\sum P_{Li}$, and the total tie-line network losses in the system, $\sum P_{\text{loss}}$.

The direct cause of frequency fluctuations in a power system is the imbalance between the input and output power of the generator, namely the imbalance between the mechanical power input from the prime mover and the electromagnetic power output by the generator [27,28]. Taking a single-machine single-load system as an example, the system frequency is determined by the rotational speed of the synchronous generator:

$$f = \frac{np}{60} \quad (6)$$

where f denotes the system frequency, in Hz; n denotes the rotational speed of the synchronous generator, in r/min; and P denotes the number of pole pairs of the synchronous generator.

Studying the dynamic variation law of system frequency is essentially equivalent to studying the motion law of the synchronous generator rotor. Accordingly, the rotor motion equation of the generator can be obtained as follows:

$$P_m - P_e = \Delta P = \frac{J}{\omega_n} \frac{d\omega}{dt} \quad (7)$$

When an imbalance occurs between the mechanical power input by the prime mover and the electromagnetic power output by the generator, the generator rotational speed will inevitably change. In a multi-generator multi-load system, the post-disturbance mechanical power P_m is not constant; rather, it depends on the generator capacity and governor characteristics, while the variation of P_e depends on the generator's own electromagnetic characteristics and the system load characteristics. Therefore, to ensure system frequency stability, load shedding can be adopted to reduce the electromagnetic power P_e of the generator.

When an MMC-HVDC blocking fault occurs, the receiving-end power grid loses not only the active power (P) transmitted by the MMC-HVDC system but also the reactive power (Q) regulated by it, which may consequently lead to system frequency instability and voltage fluctuations. After the load-shedding strategy is implemented, the system voltage level can be improved, as shown in Equation (8) [29]:

$$V_i = \sum_{k=1}^m |K_{ik}| \cos(\theta_{ik}) V_{Gk} - \sum_{j=1}^n \left| \frac{Z_{ij}}{V_j \angle \delta_j} \right| \cos(\theta_{ij}) P_j \quad (8)$$

where V_i denotes the voltage magnitude of node i ; K_{ik} and Z_{ij} are obtained from the elements of the system nodal admittance matrix, where $K = -Y_{LL}^{-1}Y_{LG}$ and $Z = Y_{LL}^{-1}$. Here, Y_{LL} is the self-admittance matrix of the load nodes in the system; Y_{LG} is the mutual admittance matrix between the load nodes and generator nodes; V_{Gk} denotes the voltage magnitude of generator node k ; m and n denote the numbers of generator nodes and load nodes in the system, respectively; θ_{ik} is the phase angle difference between nodes i and k ; θ_{ij} is the phase angle difference between nodes i and j ; V_j denotes the voltage magnitude of node j ; ζ_j is the power factor of node j ; P_j denotes the active load at node j .

When load shedding control is implemented at load node j and a certain amount of load is disconnected, the active load power P_j decreases, thereby increasing the voltage magnitude V_j at node i . In renewable-energy-integrated regions, the voltage regulation capability is relatively weak, and MMC-HVDC blocking prevents the MMC-HVDC system from participating in system voltage regulation. Therefore, when the amount of load shedding is excessively large, the voltage V_j may recover excessively, resulting in overvoltage. Conversely, insufficient load shedding may lead to inadequate voltage recovery, causing undervoltage and ultimately deteriorating the system frequency, as shown in Figure 1. In summary, inappropriate load shedding locations and load shedding amounts are likely to cause voltage violations at renewable energy grid-connection points, reduce renewable energy output, and endanger system frequency stability.

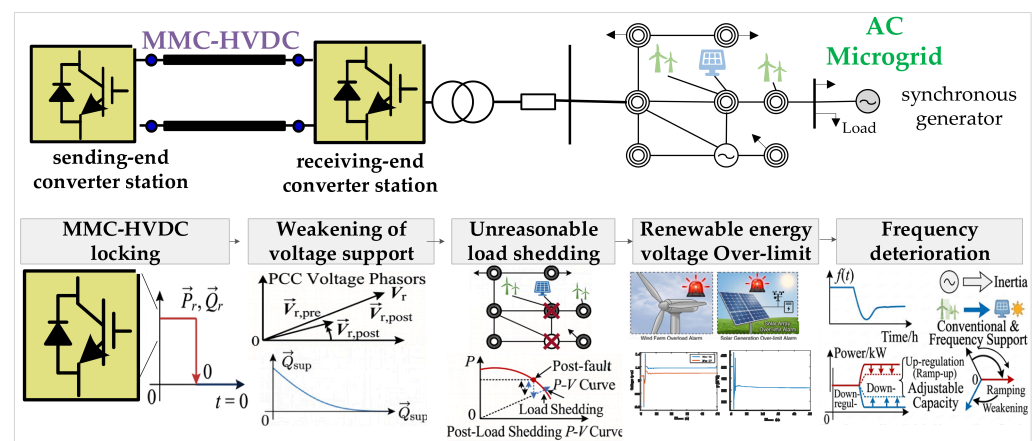


Figure 1. New energy voltage exceeding limit caused by emergency load shedding after DC blocking fault.

According to Equation (8), the voltage magnitude of a load bus is jointly influenced by generator-side voltage support, network parameters, phase-angle difference, load power factor, and active power demand. Therefore, emergency load shedding mitigates the active power imbalance, improves frequency response, and changes the post-fault power flow distribution and affects the voltage recovery of nearby renewable energy grid-connection buses. Excessive load shedding near a renewable energy bus may lead to overvoltage, whereas insufficient shedding or electrically distant shedding locations may result in undervoltage. In both cases, voltage ride-through may be triggered, reducing renewable energy active power output and further worsening the post-fault frequency response.

The above mechanism establishes the theoretical link between load-shedding actions and renewable energy voltage violations. In the proposed framework, this relationship is incorporated into the CMDP formulation in Section 3 by introducing voltage security as an operational constraint, and is further validated in Section 4 through comparisons among the contingency-based, MDP-based, and CMDP-based load-shedding strategies. Taking the modified IEEE 39-bus system shown as an example, simulations are conducted on the DiGSILENT/PowerFactory platform. An MMC-HVDC blocking fault is applied at $t = 2$ s,

resulting in an active power deficit of 832 MW. Considering the operation of reactive power compensation devices, Load 20 near the fault location is selected for load shedding, with a shedding depth of 65%, corresponding to an active load reduction of 678 MW. As shown in Figure 2, this strategy causes the terminal bus voltage of wind turbine W1 to exceed 1.18 p.u., forcing the unit into a high-voltage ride-through state. As a result, the active power output of W1 decreases by nearly 180 MW. If the wind turbine is disconnected from the grid at this moment, a secondary frequency decline may occur, seriously threatening the secure and stable operation of the system.

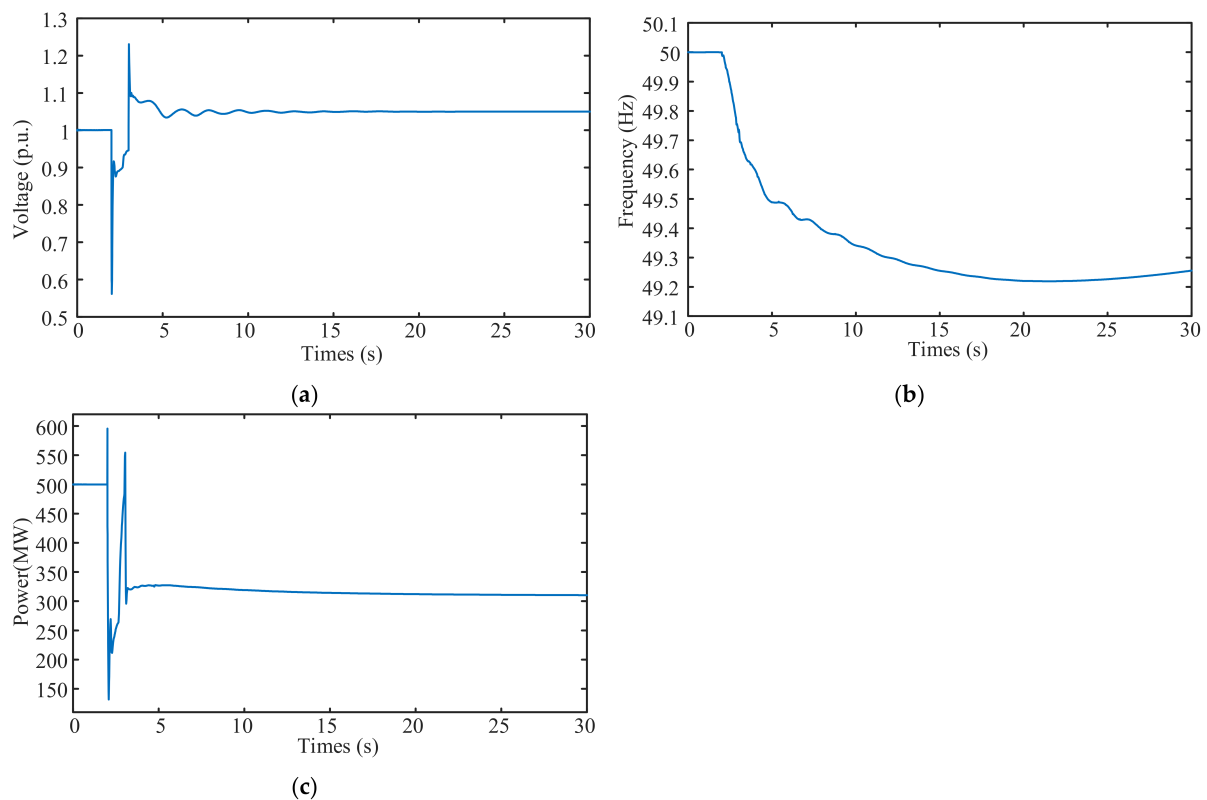


Figure 2. Effect diagram after implementing the load-shedding strategy: (a) new-energy grid-connected voltage after load shedding; (b) system frequency; (c) active power output of fan W1 after load shedding.

3. Safe-RL-Based Emergency Load-Shedding Decision-Making Model

In this section, a constrained Markov decision process is formulated on the basis of the conventional Markov decision process by introducing operational security constraints. Then, Lagrange multipliers are adopted to transform the constrained policy optimization problem into an unconstrained optimization problem, so that frequency stability, voltage security, and load shedding economy can be coordinated in the decision-making process. By considering frequency stability, load shedding cost, maximum load shedding limits, and voltage security constraints, a mathematical model for emergency load shedding is established. Based on this model, the electrical variables required for state observation are determined, the emergency load-shedding actions are defined, and the reward function is constructed, thereby enabling dynamic interaction between reinforcement learning and the power system.

3.1. Constrained Markov Decision Process

To address the rapid frequency decline and the possible triggering of operational constraints following MMC-HVDC blocking, this paper establishes a constrained Markov

decision process model that jointly considers frequency support performance and system security constraints. In reinforcement learning, the interaction process is described by a Markov decision process (MDP) which satisfies the Markov property; that is, the state at the next time instant depends only on the current state. When actions are taken into account, the next state is determined by both the current state and the current action.

An MDP is represented by a tuple containing multiple elements:

$$M = \{S, A, P, R, \gamma\} \quad (9)$$

where S denotes the state space, A denotes the action space, and P denotes the transition probability, namely the probability of transitioning from the current state $s_t \in S$ to the next state $s_{t+1} \in S$ after taking action $a \in A$. R is the reward function, and $R(S, A)$ represents the reward obtained by executing action a . γ is the reward discount factor used to calculate the cumulative return, where $\gamma \in [0, 1]$. It defines the importance of future rewards to the current decision. A larger value of γ indicates that future rewards have a greater influence on the decision-making process, while a value of γ closer to 0 indicates that the agent pays more attention to the short-term return of the action policy. At time t , the agent is in state s_t , executes action a_t , and transitions to state s_{t+1} with probability P_t . Then, according to s_t , a_t , and s_{t+1} , the reward value $r(s_{t+1}, s_t, a_t)$ is obtained through the reward function and transmitted to the agent. The above steps are repeated throughout the reinforcement learning process until the episode terminates.

When an MMC-HVDC blocking fault occurs, the power system transitions from a steady-state operating condition to a transient process. The state variation composed of voltage and frequency is defined as s_t . After the load-shedding strategy, a , is executed, the system frequency and voltage recover to s_{t+1} . The agent evaluates the system according to the predefined security indices and obtains the optimal load-shedding strategy after multiple explorations based on the reward r_t .

The actions executed by the agent are determined by its policy, which defines the mapping from states to actions. The state transition of the environment is governed by its dynamic model, while the reward function specifies how rewards are generated. Together with the state space and action space, the environment dynamics and reward function constitute a Markov decision process.

According to renewable energy grid-connection requirements, the voltage at renewable energy grid-connection points must remain within the prescribed technical limits. However, after MMC-HVDC blocking, voltage support in renewable-energy-integrated regions becomes insufficient, and load shedding may easily induce voltage violations. Therefore, a constrained Markov decision process is introduced on the basis of the conventional Markov decision process. To explicitly describe the security-constrained emergency load shedding problem, the CMDP is defined as:

$$M_c = \langle S, A, P, r, c, d, \gamma, \rho_0 \rangle \quad (10)$$

where S denotes the state space, A denotes the action space, $P(s_{t+1}, s_t, a_t)$ denotes the state transition probability, $r(s_t, a_t)$ denotes the reward function, $c(s_t, a_t)$ denotes the constraint cost function, d denotes the allowable constraint boundary, γ is the discount factor, and ρ_0 is the initial state distribution. Compared with the conventional MDP, the CMDP requires the agent to maximize the expected cumulative reward while satisfying the expected cumulative constraint cost.

For a given policy, π , the expected discounted reward and constraint cost are defined as:

$$\begin{aligned} J_r(\pi) &= \mathbb{E}_\pi \left[\sum_{t=0}^T \gamma^t r(s_t, a_t) \right] \\ J_c(\pi) &= \mathbb{E}_\pi \left[\sum_{t=0}^T \gamma^t c(s_t, a_t) \right] \end{aligned} \quad (11)$$

Therefore, the emergency load-shedding decision-making problem can be formulated as:

$$\begin{aligned} \max_{\pi} & J_r(\pi) \\ \text{s.t.} & J_c(\pi) \leq d \end{aligned} \quad (12)$$

In this paper, the state s_t consists of post-fault frequency and voltage measurements. The action a_t represents the selected load shedding locations and corresponding shedding depths. The reward $r(s_t, a_t)$ is designed to encourage frequency recovery and reduce unnecessary load shedding. The constraint cost $c(s_t, a_t)$ is associated with violations of operational security constraints, including the maximum load shedding limit, post-fault frequency security constraint, and renewable energy grid-connection voltage security constraint.

The constraint cost can be expressed as a weighted sum of the violation degrees of different security constraints:

$$c(s_t, a_t) = \omega_p C_p(s_t, a_t) + \omega_f C_f(s_t, a_t) + \omega_U C_U(s_t, a_t) \quad (13)$$

where C_p , C_f , and C_U denote the violation degrees of the load shedding capacity constraint, frequency security constraint, and renewable energy voltage security constraint, respectively. ω_p , ω_f , and ω_U are the corresponding weighting coefficients. When all operational constraints are satisfied, the constraint cost is equal to zero or remains below the allowable boundary. Once over-shedding, frequency instability, or renewable energy voltage violation occurs, the corresponding constraint cost increases and affects the subsequent policy update.

Compared with the standard MDP/PPO framework, the proposed CMDP-based framework does not merely modify the reward function by adding penalty terms. Instead, it explicitly introduces operational security constraints through constraint costs and allowable constraint boundaries. In this way, voltage violations at renewable energy grid-connection points are treated as constrained decision-making conditions rather than simply undesirable outcomes in reward evaluation.

CMDP is an extension of MDP, as shown in Figure 3. Compared with conventional MDP, CMDP incorporates constraint cost functions and admissible constraint boundaries into the optimization process.

To solve the above CMDP-based constrained optimization problem, the Lagrangian relaxation method is introduced. By introducing the Lagrange multiplier λ , the constrained policy optimization problem can be transformed into an unconstrained Lagrangian optimization problem:

$$\mathcal{L}(\pi, \lambda) = J_r(\pi) - \lambda(J_c(\pi) - d), \quad \lambda \geq 0 \quad (14)$$

where λ is used to adjust the penalty intensity of constraint violations. When the policy leads to voltage violations or frequency security violations, $J_c(\pi) > d$, and the Lagrange multiplier increases. As a result, unsafe load-shedding actions receive a larger penalty during subsequent policy updates. Conversely, when the policy satisfies the security constraints, $J_c(\pi) \leq d$, the Lagrange multiplier decreases or remains at a small value, allowing the agent to further optimize the load shedding amount while maintaining system security.

The Lagrange multiplier is updated iteratively according to the constraint violation degree:

$$\lambda_{k+1} = [\lambda_k + \eta_\lambda (J_c(\pi_k) - d)]^+ \quad (15)$$

where k denotes the training iteration, η_λ is the learning rate of the Lagrange multiplier, and $[\cdot]^+ = \max(\cdot, 0)$ ensures the non-negativity of λ . In the proposed emergency load shedding framework, PPO updates the policy parameters under the current λ , while λ is updated according to the constraint feedback returned by the power system simulation. In the case study, $\lambda = 1.2$ is adopted as the penalty coefficient after preliminary tuning to balance voltage security and load shedding economy. Through this primal dual update mechanism, the agent gradually learns a load-shedding strategy that coordinates frequency stability, voltage security, and load shedding economy.

Traditional emergency load-shedding strategies mainly aim to prevent frequency collapse. In contrast, the constrained Markov decision process considered in this paper seeks to restore system frequency while simultaneously maintaining the voltages at renewable energy grid-connection points within the safe operating range. Therefore, ensuring that the bus voltages of renewable energy units and high-voltage transmission systems remain within the prescribed security limits after emergency load shedding is a key consideration in strategy formulation.

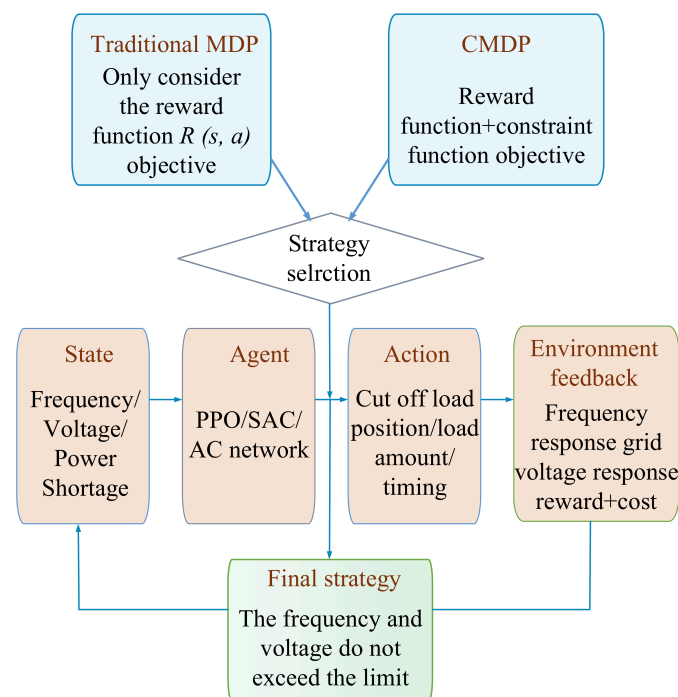


Figure 3. The interaction process between the agent and the environment in CMDP.

3.2. Construction of the Emergency Load-Shedding Model

To address emergency load shedding after power system faults, this paper constructs a load-shedding model with security constraints to ensure secure and stable system operation. In addition to maintaining post-fault security and stability, the load shedding amount should also be minimized. Therefore, a mathematical model for load-shedding strategy optimization is established based on the principle of minimizing the total load shedding amount while guaranteeing post-fault system stability:

$$F = \min \Delta P = \min \sum_{i=1}^n P_{\text{shed},i} \quad (16)$$

where ΔP denotes the total load shedding amount, $P_{\text{shed},i}$ denotes the total load shed at node i , and $P_{\text{shed},i} > 0$.

The above mathematical model for load shedding optimization should satisfy the following constraints:

(1) The load shedding amount at each load shedding location shall not exceed its own total load, and the total load shedding amount shall not exceed the total system load:

$$\begin{cases} P_{\text{shed},i} \leq P_{\text{load},i} \\ \sum_{i=1}^n P_{\text{shed},i} \leq \sum_{i=1}^n P_{\text{load},i} \end{cases} \quad (17)$$

where $P_{\text{load},i}$ denotes the total load at node i , which depends on the actual operating condition of the system.

(2) Post-fault frequency security constraint: after the emergency load-shedding action is executed, the minimum system frequency should not be lower than the first-round setting threshold of under-frequency load shedding, and the maximum frequency should not threaten the stable operation of the system:

$$f_{\text{UFLS}} < f_{\text{shed,after}} < f_{\text{max}} \quad (18)$$

where f_{UFLS} denotes the first-round setting value of under-frequency load shedding, $f_{\text{shed,after}}$ denotes the dynamic system frequency after load shedding, and f_{max} denotes the maximum operating frequency acceptable for the system.

(3) Bus voltage security constraint: to prevent the bus voltage level of renewable energy units from becoming excessively high or excessively low after emergency load shedding, which may cause renewable energy units to disconnect from the grid and induce cascading faults, thereby threatening the secure and stable operation of the system, the bus voltage security constraint is defined as follows:

$$U_{i,\min} < U_i < U_{i,\max} \quad (19)$$

where U_i denotes the bus voltage at node i , $U_{i,\min}$ denotes the minimum allowable operating voltage of the equipment at node i , and $U_{i,\max}$ denotes the maximum allowable operating voltage of the equipment at node i .

The voltage constraint in Equation (19) is a grid-code-oriented operational security constraint rather than a complete converter-control stability model. It is introduced to prevent renewable energy units from entering high-voltage or low-voltage ride-through states after emergency load shedding. In the proposed CMDP framework, this constraint provides an observable and computationally efficient security boundary for agent training. Although it does not replace detailed converter-control stability modeling, it can effectively capture the operational risk that terminal voltage violations may cause renewable energy active power reduction and secondary frequency deterioration.

Post-fault frequency variation and bus-voltage recovery are closely related to the load shedding locations and amounts. Therefore, the key objective of emergency load shedding optimization is to determine appropriate shedding locations and satisfy system stability requirements with the minimum total load shedding amount.

3.3. Construction of the State–Action and Reward–Constraint Framework

To address the rapid frequency decline caused by MMC-HVDC blocking faults, this paper develops a safe-RL-based emergency load-shedding decision-making model to achieve fast and effective load-shedding actions while satisfying system security constraints. Most

traditional controllers make control decisions based on variations in measured electrical quantities. Similarly, the data-driven RL agent responds by acquiring real-time electrical variables from the system. Therefore, it is feasible to treat the agent as the control center for formulating post-fault emergency load-shedding strategies. In this paper, reinforcement learning is adopted to model and solve the emergency load shedding problem in power systems. To make the interaction process between the agent and the power system environment clearer, the state variables, action space, reward function, and constraint-handling mechanism are further described below. To improve the rigor and reproducibility of the CMDP formulation, the selected observation points, action discretization rules, penalty terms, and key parameter settings are explicitly summarized in this subsection and in the parameter table in Section 5.1.

(1) **State variable (s_t):** The system state at any time interval t is defined as:

$$s_t = \{f_{1,t-T}, f_{2,t-T}, \dots, f_{n,t-T}, U_{1,t-T}, \dots, U_{m,t-T}\} \quad (20)$$

where $f_{1,t-T}, f_{2,t-T}, \dots, f_{n,t-T}$ are the frequency values of the selected n observation points within the time interval $t - T$, and $U_{1,t-T}, U_{2,t-T}, \dots, U_{m,t-T}$ are the voltage values of the selected m observation points within the time interval $t - T$.

The observed state variables have the following functions:

- (1) the agent determines the degree of system instability according to the current electrical measurements and the variation trends of electrical quantities over the previous several time instants, and then provides corresponding stability control measures;
- (2) the control effect is evaluated according to the electrical quantity variations within one step after the previous stability control measure is executed;
- (3) whether the constraints are violated is judged according to the state feedback after load shedding is implemented.

Specifically, the state vector contains two types of post-fault dynamic information. The first type is frequency-related information, including the frequency values of selected observation buses or generator buses within the observation window. These variables reflect the severity and evolution trend of the post-fault active power imbalance. The second type is voltage-related information, including the terminal bus voltages of renewable energy units and other key buses. These variables are used to judge whether the renewable energy units are close to the voltage ride-through boundary. In order to improve training stability, the observed frequency and voltage variables are normalized before being input into the neural network. Therefore, the agent can evaluate both the frequency recovery requirement and the voltage security margin when selecting emergency load-shedding actions.

In the case study, the selected observation variables are explicitly defined according to the electrical quantities exported from the DIgSILENT/PowerFactory simulation. The frequency-related observation is the frequency of Bus 39, denoted as $f_{39,t}$. The voltage-related observation is the voltage magnitude of Bus 34, denoted as $U_{34,t}$, which is the grid-connection bus of the monitored wind turbine. In addition, the active power output of the monitored wind turbine is also observed and denoted as $P_{W1,t}$. Therefore, the state vector at decision time t can be expressed as:

$$s_t = [f_{39,t}, U_{34,t}, P_{W1,t}] \quad (21)$$

If a time observation window is adopted, the state vector can be extended as:

$$s_t = [f_{39,t-L+1:t}, U_{34,t-L+1:t}, P_{W1,t-L+1:t}] \quad (22)$$

where L denotes the length of the observation window. Here, $f_{39,t}$ reflects the post-fault frequency response of the system, $U_{34,t}$ is used to evaluate the voltage security margin of the renewable energy grid-connection point, and $P_{W1,t}$ characterizes the active power response of the monitored wind turbine after voltage variations. These observed variables are normalized before being input into the neural network to improve training stability.

(2) **Action (a_t):** One to three selected load shedding locations in the system are chosen to formulate stability control measures:

$$\begin{cases} a_t = \{a_{1,t}, a_{2,t}, a_{3,t}\} \\ a_{1,t} = [P_{loc,a}, D_{deep,i}] \\ a_{2,t} = [P_{loc,b}, D_{deep,j}] \\ a_{3,t} = [P_{loc,c}, D_{deep,k}] \end{cases} \quad (23)$$

where $a_{1,t}$, $a_{2,t}$, $a_{3,t}$ represent the three load shedding locations, respectively; $P_{loc,a}$, $P_{loc,b}$, $P_{loc,c}$ denote the selected load shedding locations, respectively; and $D_{deep,i}$, $D_{deep,j}$, $D_{deep,k}$ denote the selected load shedding depths, respectively. When $a \neq b \neq c$, three load shedding locations are selected; when $i \neq j \neq k$, different load shedding depths are selected.

In the case study, the candidate load shedding locations are selected from buses with controllable loads and strong electrical relevance to the disturbed area, including Load 39, Load 04, Load 23, Load 27, Load 20, Load 24, and Load 16. The shedding depths are discretized as 0%, 20%, 35%, 50%, 65%, 80%, and 100%. At each decision step, the agent selects one to three load shedding locations and their corresponding shedding depths to form a combined control action. Repeated selections of the same load shedding location are avoided during action decoding. If the selected action violates the load shedding capacity limit or causes security violations, it is penalized through the reward and constraint cost functions. This discrete action-space design reduces the difficulty of agent training and is also consistent with practical emergency load shedding schemes, where load shedding is usually implemented in preset locations and discrete stages.

In the power system simulation environment, the stability control measures generated according to the system state are defined as actions. An action may include multiple control operations executed at different times within a single interaction step. Its parameters include the target equipment, control depth, and execution time. In this framework, reward design is particularly important because it directly determines the evaluation of control effectiveness and affects both the convergence of reinforcement learning and the quality of the resulting stability-control strategy. The reward is constructed based on the electrical measurements obtained after the fault or after the execution of control actions, while also accounting for security constraints.

(3) **Reward (r_t):** The reward function of the proposed method is designed to maximize the objective of the mathematical model for optimized load-shedding strategy execution under the premise of satisfying the emergency load-shedding strategy constraints formulated in Section 2.2, as follows:

$$\begin{cases} F = \min \Delta P = \min \sum_{i=1}^n P_{shed,i} \\ r_t = -(F + C + \lambda G(u)) \end{cases} \quad (24)$$

where F denotes the mathematical model for the optimized load-shedding strategy formulated in Equation (16), C denotes the extremely large penalty assigned when Constraints 1 and 2 are violated, λ denotes the Lagrange multiplier, and $G(u)$ denotes the penalty assigned when a constraint is violated. The value of $G(u)$ should be much smaller than C , because

violations of Constraints 1 and 2 are intolerable for the secure and stable operation of the system. The proposed method aims to satisfy the voltage security of system components through the influence of Constraint 3 while adopting load shedding measures that are as small as possible.

The reward function is tuned according to three objectives: maintaining frequency stability, reducing unnecessary load shedding, and avoiding renewable energy voltage violations. The frequency-related reward is assigned when the post-load-shedding frequency remains within the secure range, while a large penalty is imposed when frequency instability occurs. The load-shedding-related term is designed as a cost term proportional to the total amount of shed load, so that the agent is encouraged to use the minimum necessary load shedding amount. The voltage-related penalty is activated when the terminal voltage of a renewable energy unit exceeds the allowable operating range. In the case study, the minimum frequency requirement is set to 49.2 Hz, and the renewable energy terminal voltage is constrained within 0.90–1.10 p.u. The specific reward and penalty coefficients are selected through preliminary tuning to ensure that frequency security and voltage security have higher priority than load shedding economy, while the load shedding amount is further optimized after the security constraints are satisfied.

To improve the reproducibility of the reward design, the reward and penalty coefficients are selected according to a hierarchical security priority principle rather than only by trial-and-error. The first priority is to avoid unacceptable operating conditions, including frequency instability and infeasible load-shedding actions that exceed the available controllable load. Therefore, these violations are assigned large penalties so that their negative effects dominate the reward evaluation. The second priority is to maintain the terminal voltages of renewable energy units within the allowable voltage ride-through range. The voltage-related penalty $G(u)$ is determined according to the voltage deviation degree and the grid-connection voltage requirements listed in Table 1. The third priority is to reduce the total load shedding amount after the basic frequency and voltage security requirements are satisfied. Therefore, the load shedding amount is treated as an optimization cost rather than a hard penalty.

Table 1. Response requirements for abnormal voltage in photovoltaic power plants. (Note: Italics are used for mathematical variables in voltage expressions, such as U and U_N).

Voltage at the Grid-Connection Point	Maximum Tripping Time
$U < 0.5 U_N$	0.1 s
$50\% U_N \leq U < 85\% U_N$	2.0 s
$85\% U_N \leq U \leq 110\% U_N$	Continuous operation
$110\% U_N < U < 135\% U_N$	2.0 s
$135\% U_N \leq U$	0.1 s

In this design, the penalties for frequency instability and infeasible load-shedding actions should be larger than the maximum possible benefit obtained from reducing the load shedding amount. This ensures that the agent does not sacrifice system security for load shedding economy. The Lagrange multiplier λ is used to adjust the relative importance of voltage security constraints. A larger λ makes the agent more conservative and reduces the probability of renewable energy voltage violations, whereas a smaller λ allows the agent to focus more on load shedding economy but may weaken voltage security guidance. Therefore, $\lambda = 1.2$ is adopted after preliminary tuning to balance voltage security and load shedding economy in the studied system.

(4) **Security operation constraints ($C, G(u)$):** The constraint-handling mechanism consists of two levels. First, hard operational constraints are explicitly defined in the emergency load-shedding model, including the maximum available load at each shedding location, the acceptable post-fault frequency range, and the allowable voltage range of renewable energy grid-connection buses. These constraints correspond to Equations (14)–(16). Second, during agent training, violations of these constraints are converted into constraint costs and penalty terms. Severe violations, such as excessive load shedding beyond the available load or frequency instability, are assigned large penalties because they are unacceptable for secure system operation. Voltage violations at renewable energy grid-connection points are penalized according to the voltage ride-through requirements, since such violations may reduce renewable energy active power output and further deteriorate frequency stability. Through this mechanism, unsafe actions are suppressed during policy updates, and the final policy tends to satisfy frequency stability, voltage security, and load shedding economy simultaneously. An extremely large penalty is assigned for violations corresponding to C , which will not be further elaborated here. As shown in Table 1, according to the requirements of the national standard Technical Requirements for Grid Connection of Photovoltaic Systems GB/T 19939-2005 [23], the required grid-connected operation time of photovoltaic power stations varies with the voltage at the grid-connection point. Therefore, different penalty operations are adopted for $G(u)$:

$$G(u) = \begin{cases} 1000, & 110\% U_N < u < 135\% U_N \\ 2000, & 135\% U_N < u \end{cases} \quad (25)$$

4. Solution of the PPO-Based Emergency Load Shedding Decision-Making Model

To formulate an emergency load-shedding strategy that satisfies both frequency stability and voltage security requirements, this section adopts the Proximal Policy Optimization (PPO) algorithm based on the Markov decision model established in Section 3. The optimal load-shedding scheme is obtained through policy iteration to satisfy frequency stability requirements while preventing renewable energy units from entering voltage ride-through states after load shedding. Finally, the reinforcement-learning-based emergency load-shedding training is implemented through a Python–DIGSILENT co-simulation platform.

4.1. Proximal Policy Optimization Algorithm

In this paper, Proximal Policy Optimization (PPO) is adopted to improve convergence speed and decision-making efficiency following DC blocking faults. PPO is a policy-based reinforcement learning algorithm that employs two neural networks: the policy network (actor) and the value network (critic). The actor generates actions according to the current state, while the critic evaluates the quality of the selected action based on the current state and the next state. This actor–critic framework alleviates the high-variance problem of conventional policy gradient methods and enables more efficient training. PPO is selected in this study for the following reasons. First, the emergency load shedding problem after MMC-HVDC blocking is characterized by high-dimensional nonlinear post-fault dynamics, discrete combined control actions, and strict operational security constraints. Compared with value-based methods such as DQN, PPO directly optimizes the stochastic policy and is more suitable for handling combined action decisions involving load shedding locations and shedding depths. Second, compared with conventional policy-gradient methods, PPO introduces a clipped surrogate objective to restrict excessive policy updates, which improves training stability when unsafe actions may lead to frequency instability or renewable energy voltage violations. Third, compared with TRPO, PPO avoids the complicated

second-order optimization process and is easier to implement in the Python–DigSILENT co-simulation framework. Therefore, PPO provides a suitable balance among decision-making performance, training stability, and computational implementation complexity for the proposed CMDP-based emergency load shedding problem.

Traditional policy gradient algorithms work by computing an estimator of the policy gradient and substituting it into a stochastic gradient ascent algorithm. The most commonly used gradient estimators are given as follows:

$$\hat{g} = \hat{\mathbb{E}}_t [\nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \hat{A}_t] \quad (26)$$

$$\hat{A}_t = -V(s_t) + r_t + \gamma r_{t+1} + \dots + \gamma^{T-t-1} r_{T-1} + \gamma^{T-t} V(s_T) \quad (27)$$

where π_{θ} denotes the stochastic policy, A_t denotes the estimator of the advantage function at time step t , also referred to as the reward function, $V(s_t)$ denotes the state-value function under state s_t , r_t denotes the reward at time t , and γ denotes the discount factor. Here, the expectation $\mathbb{E}_t[\dots]$ represents the empirical average over a finite batch of samples in an algorithm that alternates between sampling and optimization. By constructing an objective function whose gradient is the policy gradient estimator, the estimator g can be obtained through differentiation, and the loss function $L^{PG}(\theta)$ is therefore expressed as:

$$L^{PG}(\theta) = \hat{\mathbb{E}}_t [\log \pi_{\theta}(a_t | s_t) \hat{A}_t] \quad (28)$$

Although it appears feasible to perform multiple optimization steps on the loss function $L^{PG}(\theta)$ using the same trajectory, namely the set of all states and actions within one executed segment, empirical results show that the algorithm is prone to generating destructively updated policies. In the process of formulating emergency load-shedding strategies, when under-shedding or over-shedding strategies are selected and system instability is caused, the previously generated policies will be greatly affected, which is unfavorable for policy selection and updating.

To avoid such problems, the PPO algorithm constructs the clipped loss function L^{CLIP} to replace L^{PG} , and performs multiple steps of stochastic gradient ascent on this objective. Let $r_t(\theta)$ denote the probability ratio:

$$r_t(\theta) = \frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{old}}(a_t | s_t)}, \quad r_t(\theta_{old}) = 1 \quad (29)$$

where $\pi_{\theta_{old}}$ is the policy extracted from historical data, which is referred to as the empirical policy. By maximizing its objective, we obtain:

$$L^{CPI} = \hat{\mathbb{E}}_t \left[\frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{old}}(a_t | s_t)} \hat{A}_t \right] = \hat{\mathbb{E}}_t [r_t(\theta) \hat{A}_t] \quad (30)$$

If no constraint is imposed, maximizing L^{CPI} will lead to an excessively large policy update, that is, $r_t(\theta)$ becomes very large. In this case, the selected policy differs significantly from the empirical policy, which is unfavorable for strategy formulation and is not allowed in the optimization process. Therefore, it is necessary to consider how to modify the objective function so as to penalize policy changes that cause $r_t(\theta)$ to deviate from 1.

$$L^{CLIP} = \hat{\mathbb{E}}_t [\min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t)] \quad (31)$$

where ϵ is a hyperparameter. The first term in the minimization is L^{CPI} , while the second term is modified by clipping the probability ratio, restricting $r_t(\theta)$ within $[1 - \epsilon, 1 + \epsilon]$, thereby preventing the policy π_{θ} from deviating excessively.

The PPO-based actor–critic training procedure is summarized in Algorithm 1. During each iteration, N parallel actors collect data within the time horizon T , calculate the reward function, and update the policy over K epochs with minibatch size M .

Algorithm 1. PPO-Based Actor–Critic Training Procedure

```

For iteration = 1, 2 ... do
  For actor = 1, 2 ...,  $N$  do
    Run policy  $\pi_{\theta_{old}}$  in environment for  $T$  timesteps
    Compute advantage estimates  $\hat{A}_1, \dots, \hat{A}_T$ 
  end for
  Optimize surrogate  $L$  wrt  $\theta$ , with  $K$  epochs and minibatch size  $M \leq NT$ 
   $\theta_{old} \leftarrow \theta$ 

```

The convergence stability of PPO mainly benefits from the clipped surrogate objective. During policy optimization, the probability ratio between the updated policy and the old policy is restricted within a predefined interval. This mechanism prevents excessively large policy updates and reduces the probability of destructive policy changes caused by unsafe load-shedding actions. In the proposed CMDP framework, the Lagrange multiplier and constraint penalty further guide the policy away from actions that violate frequency or voltage security constraints. Therefore, although the CMDP-based PPO agent may require more training steps than the conventional MDP agent due to the additional voltage security constraint, it can gradually converge to a safer policy that coordinates frequency stability, voltage security, and load shedding economy.

The computational complexity of the PPO-based training process mainly consists of three parts: environment simulation, neural network forward propagation, and policy/value network parameter updates. Let N denote the number of parallel actors, T denote the interaction horizon of each actor, K denote the number of training epochs for each batch of sampled trajectories, M denote the minibatch size, and $|\theta_A|$ and $|\theta_C|$ denote the numbers of parameters in the actor and critic networks, respectively. For each iteration, NT samples are collected through interaction with the PowerFactory environment. The computational cost of policy and value network updates can be approximated as:

$$O(KNT(|\theta_A| + |\theta_C|)) \quad (32)$$

If the simulation cost of one PowerFactory interaction step is denoted as C_{sim} , the overall computational complexity of one PPO training iteration can be expressed as:

$$O(NTC_{sim} + KNT(|\theta_A| + |\theta_C|)) \quad (33)$$

In the proposed Python–DIgSILENT co-simulation framework, the power system time domain simulation usually dominates the total training time, while the neural network update cost remains relatively limited due to the moderate network scale used in this paper. In the case study, both the actor and critic networks adopt hidden layers with 64 neurons, and the total training step is set to 20,000, which provides an acceptable computational burden for offline emergency control strategy training.

4.2. Decision-Making Model Solver Based on Agent Training

To better implement the emergency load-shedding strategy for DC blocking faults, this section develops a decision-making solver based on agent training. The adopted power system simulation software is DIgSILENT/PowerFactory (2022), abbreviated as PF. PF

provides an interface for deep interaction with Python 3.8, which facilitates the integration of Python-based reinforcement learning agents with the power system.

As shown in Figure 4, the RL agent is encapsulated in a Gym environment, which serves as a toolkit for developing and evaluating reinforcement learning algorithms. It is implemented in Python and consists of two main modules: the initialization module (reset) and the action module (step_action). The initialization module is responsible for initializing the program and PF parameters, selecting a fault scenario, applying it in PF, and returning the required post-fault electrical variables as the state. The action module identifies and executes the control measures generated by the agent, advances PF by one interaction step, returns the updated state, evaluates action effectiveness and constraint violations, and finally calculates the reward.

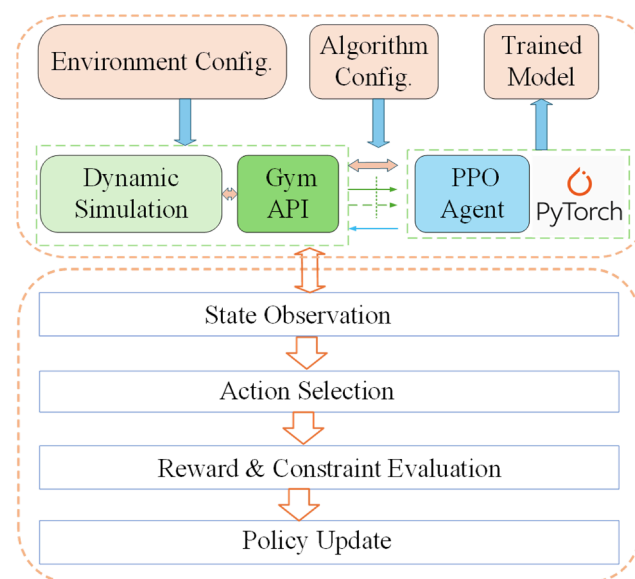


Figure 4. Software interaction framework of the power system co-simulation environment and PPO-based training platform.

4.3. Training Process of the Decision-Making Model Based on Agent Training

To address the risk of frequency instability caused by DC blocking, this section develops an adaptive emergency load-shedding scheme through agent training, as shown in Figure 5. In each episode, the interaction time instants are denoted by $T = \{t_1, t_2, t_3, \dots, t_n\}$, the observed electrical variables by $S = \{s_1, s_2, s_3, \dots, s_n\}$, and the generated control actions by $A = \{a_1, a_2, a_3, \dots, a_n\}$. The number of interaction steps in each episode is predefined.

During initialization, a fault scenario that may cause frequency instability is randomly selected from the anticipated fault set and applied to the power system. The system is then simulated until time t_1 , and the observed electrical variables at t_1 , denoted by s_1 , are returned. Based on s_1 , the agent generates the corresponding control action a_1 , and the system is advanced to time t_2 . At t_2 , the updated observation s_2 is returned, and the quality of action a_1 is evaluated according to the system response, from which the corresponding reward is calculated. The agent then generates the next control action based on s_2 . This process is repeated until the predefined number of interaction steps is reached, after which the next episode begins.

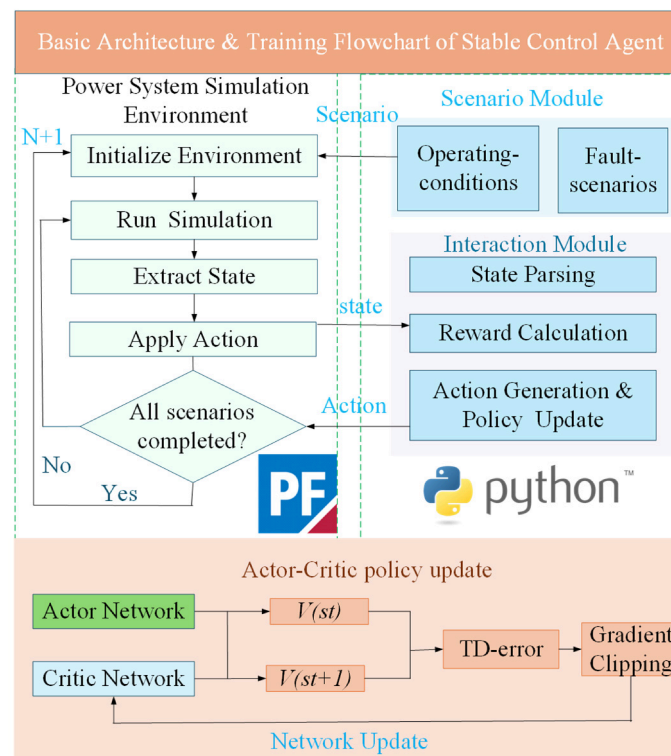


Figure 5. Co-simulation architecture and PPO training workflow of the proposed CMDP emergency load-shedding agent.

5. Case Study Analysis

To verify the effectiveness of the constrained Markov decision process for emergency load shedding in power system frequency stability control, this paper conducts case studies on the modified IEEE 39-bus system. The test system is shown in Figure 6.

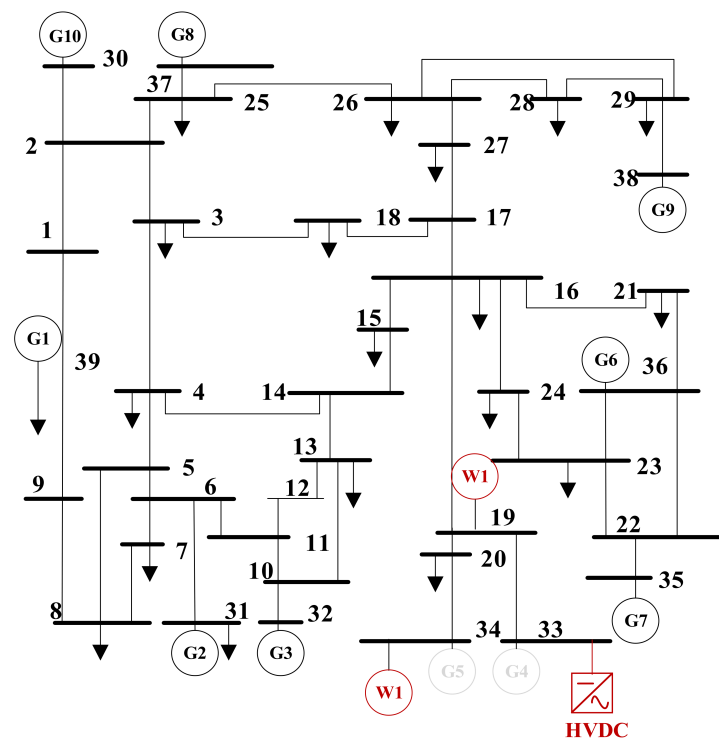


Figure 6. Improved IEEE-39 power system.

The system consists of 10 generating units equipped with governors, automatic voltage regulators, and power system stabilizers, together with 19 loads. Wind turbines are connected to Bus 1 and Bus 34, while the HVDC system is connected to Bus 33. Generators G4 and G5 are out of service. Under the initial operating condition, the HVDC system imports 832 MW from the external grid. The load shedding cost is assumed to be identical at all locations. The candidate shedding locations are Load 39, Load 04, Load 23, Load 27, Load 20, Load 24, and Load 16, and the candidate shedding depths are 0, 20, 35, 50, 65, 80, and 100. The safety constraints require the wind turbine terminal voltage to remain within 0.90–1.10 p.u. The control objective is to maintain the minimum system frequency above 49.2 Hz while minimizing the load shedding amount after an HVDC blocking fault.

For reproducibility, the observation variables used for agent training are also specified in this case study. The frequency observation is the frequency of Bus 39, the voltage observation is the voltage magnitude of Bus 34, and the active power observation is the active power output of the monitored wind turbine connected to Bus 34. These observed quantities correspond to $f_{39,t}$, $U_{34,t}$, and $P_{W1,t}$, respectively, and are used to construct the state vector of the agent.

5.1. Comparison Between the Traditional Contingency Strategy and the CMDP-Based Strategy

This section verifies the effectiveness of the CMDP-based strategy by comparing it with the traditional contingency-based strategy. The main training parameters of the agent are as follows: both the actor and critic networks use one hidden layer with 64 neurons and the ReLU activation function; the policy layer of the actor network consists of three layers with $64 \times 64 \times 64$ neurons; the value function layer of the critic network also consists of three layers with $64 \times 64 \times 64$ neurons; the total number of training steps is 20,000; the learning rate is 0.004010403089631225; the gradient clipping ratio is [0.7, 1.3]; the reward discount factor is $\gamma = 0.98$; and the Lagrange multiplier coefficient is $\lambda = 1.2$. In addition, based on Equation (24), the reward function is designed by assuming identical load shedding costs at all locations, using the total load shedding amount as the penalty term, assigning a penalty of $-10,000$ for excessive load shedding, a reward of 3000 when the system frequency remains within the stable range after load shedding, and a penalty of 5000 when frequency instability occurs. These reward and penalty coefficients are selected according to priority among emergency control objectives. Frequency instability and severe constraint violations are assigned large penalties, while the load shedding amount is treated as an optimization cost after the basic security requirements are satisfied.

The above reward and penalty settings follow the same security priority principle as the reward design in Section 3.3. Specifically, the over-shedding penalty $-10,000$ is set as the largest negative penalty because actions exceeding the available controllable load are physically infeasible. The frequency stability reward and frequency instability penalty are used to ensure that the agent first learns to maintain the post-fault frequency within the secure range. After the basic frequency security requirement is satisfied, the load shedding amount is further minimized through the load shedding cost term. The voltage security penalty is weighted by the Lagrange multiplier to guide the agent away from actions that may cause renewable energy terminal voltage violations. These parameter values are reported explicitly to improve the reproducibility of the training configuration.

To improve the reproducibility of the proposed CMDP-based emergency load-shedding strategy, the main observation variables, action discretization rules, reward and penalty coefficients, and PPO training parameters used in the case study are summarized in Table 2. These parameters are consistent with the CMDP formulation in Section 3.3 and the simulation settings of the modified IEEE 39-bus system.

Table 2. Main CMDP settings, reward function design, and PPO training parameters.

Parameter	Value	Description
Frequency observation	Bus 39 frequency f_{39}	Post-fault frequency response
Voltage observation	Bus 34 voltage U_{34}	Voltage security of the wind turbine grid-connection bus
Wind active power observation	P_{W1}	Active power output of the monitored wind turbine
Candidate shedding buses	Load 39, Load 04, Load 23, Load 27, Load 20, Load 24, Load 16	Controllable load shedding locations
Shedding depths	0%, 20%, 35%, 50%, 65%, 80%, 100%	Discrete load shedding levels
Number of selected shedding locations	1–3	Number of buses selected in one action
Frequency security threshold	$f_{min} \geq 49.2$ Hz	Post-fault frequency security requirement
Voltage security range	0.90–1.10 p.u.	Renewable energy terminal voltage security range
Over-shedding penalty	−10,000	Penalty for infeasible load shedding
Frequency stability reward	3000	Reward for satisfying frequency security
Frequency instability penalty	−5000	Penalty for frequency instability
Lagrange multiplier	1.2	Weight of voltage security penalty
Discount factor	0.98	Reward discount factor
Learning rate	0.004010403089631225	PPO learning rate
Gradient clipping ratio	[0.7, 1.3]	PPO clipping range
Training steps	20,000	Total training steps
Rollout steps per update	128	Number of rollout steps per PPO update
Minibatch size	16	Minibatch size used in PPO optimization
PPO update epochs	10	Number of PPO optimization epochs per update
Number of environments	1	Single-environment PPO training setting
Random seeds	{0, 1, 2, 3, 4}	Fixed random seeds used for repeated runs
Number of independent runs	5	Number of repeated independent training runs
Evaluation episodes per run	20	Number of deterministic evaluation episodes for each trained policy
Convergence criterion	Reward plateau in the late training stage	Training is considered converged when the mean episode reward fluctuates within a narrow range without a sustained upward trend

As shown in Table 2, the agent observes the frequency of Bus 39, the voltage of Bus 34, and the active power output of the monitored wind turbine to evaluate the post-fault frequency response, renewable energy voltage security margin, and wind power response. The action space is discretized by predefined load shedding locations and shedding depths, while the reward and penalty coefficients are selected according to the security-priority principle described in Section 3.3. The PPO training settings summarized in Table 2 follow the implementation used in the original study. These hyperparameters were determined through comparative trial experiments over several candidate settings, and the final configuration was selected according to reward improvement and training stability. In the training process, the reward curve fluctuates significantly during the early exploration stage, increases gradually as the interaction experience accumulates, and enters a converged state after the later training stage. For reproducibility, Table 3 summarizes the main PPO training settings used in this study, including the rollout steps per update, minibatch size, number of update epochs, random seed setting, number of independent runs, and evaluation episodes per run.

Table 3 reports the repeated-run results for the representative disturbance scenario. The main performance metrics are presented as mean $\pm$ standard deviation over five independent runs with fixed random seeds {0, 1, 2, 3, 4}, and the corresponding variances are also listed.

Table 4 presents the comparison between the traditional contingency strategy and the CMDP-based strategy for the same representative disturbance scenario. For the CMDP-based strategy, the values reported in Table 4 correspond to the mean results over five independent runs summarized in Table 3. As can be seen from the table, compared with the traditional contingency-based emergency load-shedding measure, the CMDP-based strategy reduces the total load shedding amount while maintaining frequency stability and improving the voltage security performance of renewable energy grid-connection points.

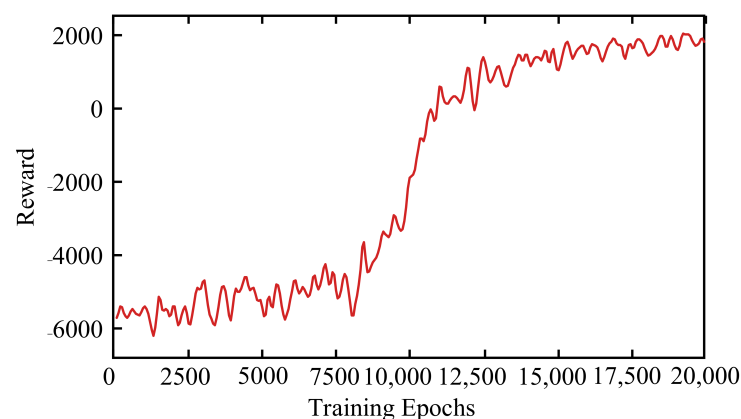
Table 3. Results of five independent runs and their statistical summary for the representative disturbance scenario.

Run/Statistic	Random Seed	Total Load Shedding Amount (MW)	Frequency Nadir (Hz)	Voltage W1 (p.u.)	Voltage P1 (p.u.)
Run 1	0	517.6	49.21	1.05	1.06
Run 2	1	522.4	49.23	1.06	1.06
Run 3	2	519.8	49.22	1.06	1.05
Run 4	3	515.9	49.21	1.05	1.06
Run 5	4	520.3	49.24	1.06	1.06
Mean	-	519.20	49.222	1.056	1.058
Standard deviation	-	2.53	0.013	0.0055	0.0045
Variance	-	6.40	1.70×10^{-4}	3.00×10^{-5}	2.00×10^{-5}

Table 4. Comparison between the traditional contingency strategy and the CMDP-based strategy.

Method	Control Measure	Total Load Shedding Amount (MW)	Frequency Nadir (Hz)	Voltage (W1/P1) (p.u.)
CMDP	Load20, Load23	519.2 MW	49.22 Hz	1.06 p.u./1.06 p.u.
Contingency Strategy	Load20	678 MW	49.21 Hz	1.15 p.u./1.09 p.u.

Figure 7 shows the convergence curve of the episode reward during training. The agent is designed to ensure frequency stability while simultaneously maintaining the voltage security of renewable energy buses and optimizing the load shedding amount. In the early stage, the reward fluctuates significantly because the agent explores many actions that may lead to frequency instability or voltage-constraint violations. As training proceeds, the accumulated interaction experience increases, and the reward gradually improves. After approximately 8000 training steps, the reward curve rises more steadily under the trust region mechanism of PPO. After about 15,000 training steps, the mean reward converges, indicating that the agent has learned an effective load-shedding strategy. In this study, training was regarded as converged when the mean episode reward entered a plateau region in the late stage and fluctuated within a relatively narrow range without a sustained upward trend.

**Figure 7.** Convergence curve of reward values during the PPO training process.

This convergence process indicates that the PPO-based CMDP agent gradually learns to avoid actions that cause frequency instability or renewable energy voltage violations. Although the exploration stage is relatively long due to the additional voltage security constraint, the final reward convergence shows that the proposed method can obtain a stable security-constrained load shedding policy.

Figure 8 compares the frequency responses under the CMDP- and contingency-based load-shedding strategies. As shown in the figure, the CMDP-based strategy increases the frequency nadir by 0.07 Hz while simultaneously optimizing the load shedding amount. This indicates that, compared with the contingency-based strategy, the proposed CMDP-based strategy achieves better overall performance in both frequency stability and load shedding economy.

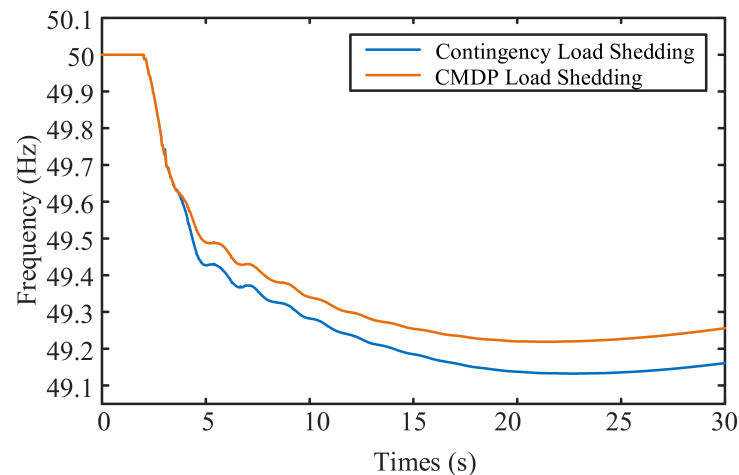


Figure 8. CMDP load shedding and contingency load shedding frequency chart.

Figures 9 and 10 compare the terminal voltage and active power responses of wind turbine W1 under the two load-shedding strategies. As shown in the figures, under the CMDP-based strategy, the steady-state terminal bus voltage of W1 remains below 1.10 p.u. and within the secure range, so the wind turbine can maintain its normal active power output. This demonstrates that CMDP can explicitly account for the voltage security of renewable energy units during strategy formulation.

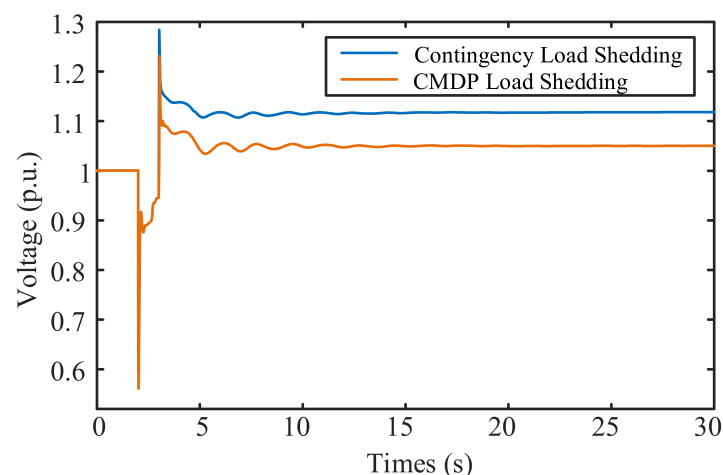


Figure 9. Comparison diagram of busbar voltage at the end of fan W1.

In summary, compared with the contingency-based strategy, the CMDP-based deep reinforcement learning method exhibits clear advantages. In particular, it optimizes the load shedding amount while increasing the frequency nadir, thereby more effectively avoiding unnecessary under-frequency load shedding and contributing to improved power supply reliability and system stability.

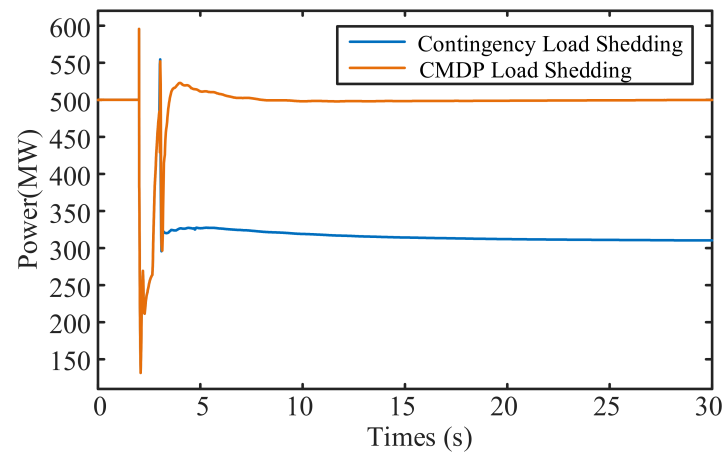


Figure 10. Comparison diagram of input power of fan W1.

5.2. Comparison of Emergency Load-Shedding Strategies Under Different Reinforcement Learning Algorithms

This section verifies the effectiveness of reinforcement learning for emergency load shedding by comparing the strategies obtained under different reinforcement learning formulations. The conventional MDP framework is also implemented using an actor–critic network and optimized with the PPO algorithm. Its network structure, number of neurons, activation function, gradient clipping ratio, and discount factor are kept consistent with those used in the CMDP framework. The MDP agent is trained for 20,000 steps, with a learning rate of 0.0013301571717463502.

Under the above basic scenario settings, the emergency load-shedding strategy obtained by the MDP-based agent after 20,000 PPO training steps is compared with that of the CMDP-based agent, as shown in Table 5. The conventional MDP achieves a higher average reward and a slightly better frequency nadir than CMDP. However, its load shedding performance is not superior to that of CMDP; moreover, because renewable energy voltage security constraints are not explicitly considered, it is more likely to cause terminal voltage violations at wind turbines. By contrast, although CMDP is slightly inferior in terms of frequency nadir, it provides better overall load shedding performance while ensuring the voltage security of renewable energy grid-connection points.

Table 5. CMDP and MDP strategy tables.

Method	CMDP	MDP
Control measure	Load 20, Load 23	Load 16, Load 23, Load 04
Total load shedding amount (MW)	519.2 MW	555.6 MW
Frequency nadir (Hz)	49.22 Hz	49.28 Hz
Voltage W1/P1 (p.u.)	1.06 p.u./1.06 p.u.	1.12 p.u./1.04 p.u.
Average reward	2480	2444

It should be emphasized that the conventional MDP benchmark uses the same PPO optimizer and similar actor–critic network settings as the proposed method, but it does not explicitly include renewable energy voltage security constraints in the decision-making process. Therefore, the comparison between MDP and CMDP mainly reflects the effect of the proposed safety-constrained formulation rather than the effect of changing the reinforcement learning optimizer. Therefore, the current benchmark comparison is intended to verify the effect of the proposed safety-constrained formulation under a consistent

simulation setting, rather than to represent an exhaustive comparison with all recent advanced benchmark methods.

Figure 11 shows the reward convergence curve of the conventional MDP-based strategy. Since the MDP framework focuses mainly on frequency stability and load shedding optimization, the reward fluctuation during the initial exploration stage is relatively small. As training proceeds, the agent gradually learns through interaction with the environment and begins to approach the optimal policy after approximately 2000 steps. Compared with CMDP, the MDP framework converges faster, and the mean reward has already converged after about 4000 steps, indicating that an effective load-shedding strategy has been obtained.

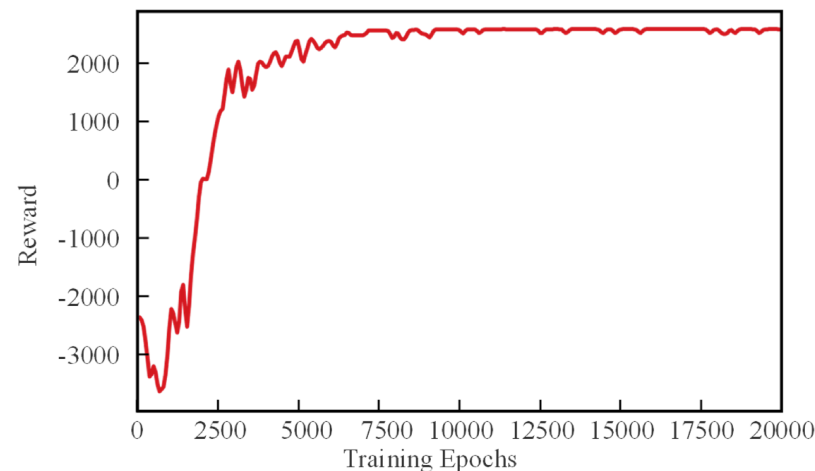


Figure 11. MDP reward value convergence change graph.

Compared with the CMDP-based training process, the conventional MDP converges faster because it only considers frequency stability and load shedding amount optimization, without explicitly incorporating renewable energy voltage security constraints. However, as shown in Table 5 and Figures 12 and 13, the faster convergence of MDP does not guarantee a safer control policy, since the obtained strategy may still cause renewable energy terminal voltage violations and active power reduction. Therefore, the convergence performance should be evaluated together with the final security performance, rather than only by the number of training steps required for reward stabilization.

Figure 14 compares the system frequency responses under the CMDP-based and MDP-based load-shedding strategies. As shown in the figure, the MDP-based strategy achieves a slightly higher frequency nadir than CMDP, indicating a certain advantage in frequency recovery. However, CMDP further optimizes the load shedding amount while simultaneously accounting for voltage security constraints, making it more suitable for coordinated security-constrained control.

Figure 12 compares the terminal voltage responses of wind turbines W1 and W2 under the two strategies. The conventional MDP does not explicitly consider the voltage security of renewable energy units; therefore, after load shedding, the terminal bus voltage remains relatively high. In particular, during 4–5 s, the voltage of W1 stays above 1.10 p.u., causing the unit to enter the high-voltage ride-through state, and its active power output correspondingly decreases, as shown in Figure 13. This indicates that, compared with conventional MDP, the CMDP-based strategy is safer because it keeps wind turbines within a secure voltage range after fault clearing and reduces the risk of cascading disturbances caused by wind turbine disconnection.

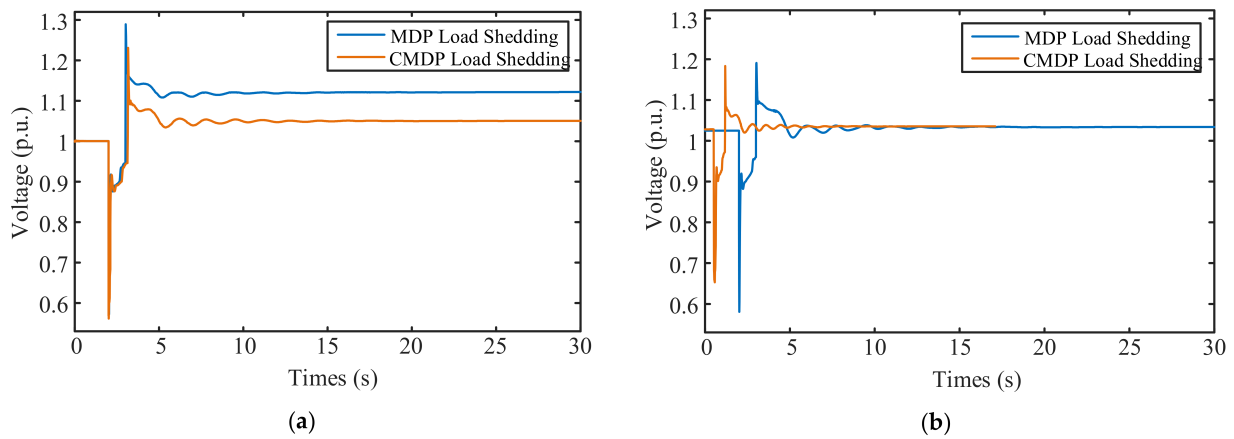


Figure 12. (a) Comparison diagram of busbar voltage at the end of fan W1; (b) comparison diagram of busbar voltage at the end of fan W2.

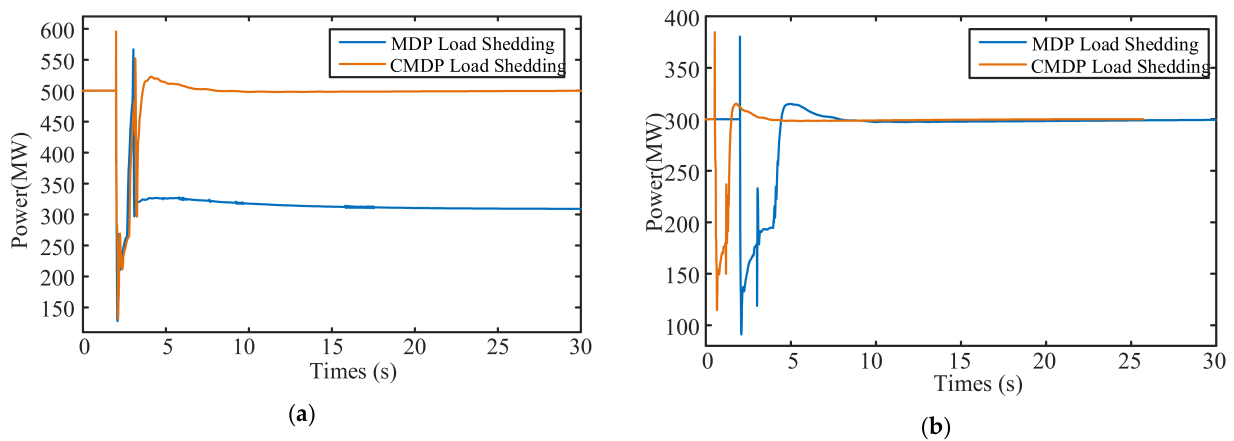


Figure 13. (a) Comparison diagram of input power of fan W1; (b) comparison diagram of input power of fan W2.

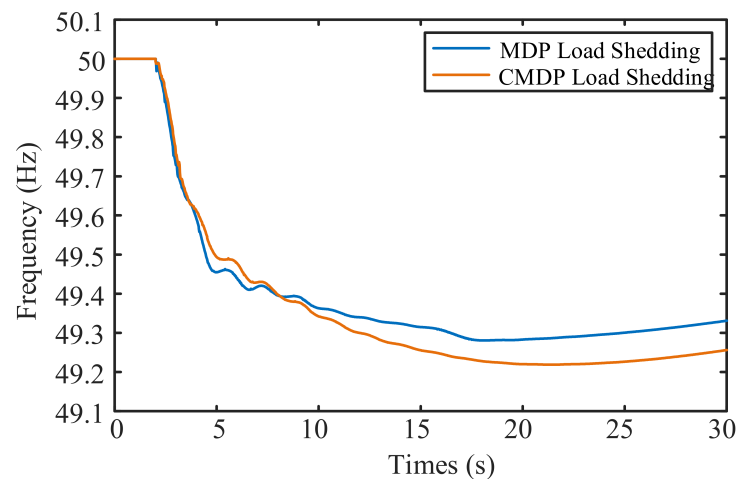


Figure 14. CMDP load shedding and MDP load shedding frequency chart.

In summary, although the conventional MDP shows a certain advantage in terms of frequency nadir, it does not account for the voltage security of renewable energy units, and its overall load shedding performance is inferior to that of CMDP. By contrast, although CMDP is slightly less effective in the frequency metric, it can simultaneously coordinate

frequency stability, voltage security, and load shedding optimization, thereby achieving better overall control performance.

To further strengthen the robustness-related discussion, the revised manuscript provides a clearer quantitative comparison of the main performance indicators, including frequency nadir, post-fault voltage security performance, and load shedding amount. In the representative disturbance scenarios considered in this study, the proposed method consistently maintains post-fault frequency security and improves the voltage security performance of renewable energy grid-connection points while avoiding excessive load shedding. These observations support the conclusion that the proposed safety-constrained strategy exhibits favorable control robustness with respect to the coordinated objectives considered in this work. The robustness-related discussion in the revised manuscript is therefore based on case study evidence that can be directly confirmed from the available project materials.

5.3. Additional Validation Under High Renewable Energy Penetration

To further evaluate the applicability of the proposed CMDP-based emergency load-shedding strategy under different operating conditions, an additional high-renewable-penetration operating condition is considered in this subsection. Compared with the base operating condition in Sections 5.1 and 5.2, the renewable generation share is increased and the corresponding generation/load distribution is adjusted accordingly, while the modified IEEE 39-bus system topology is retained. The modified 10-generator IEEE 39-bus system under the high-renewable-penetration operating condition is shown in Figure 15.

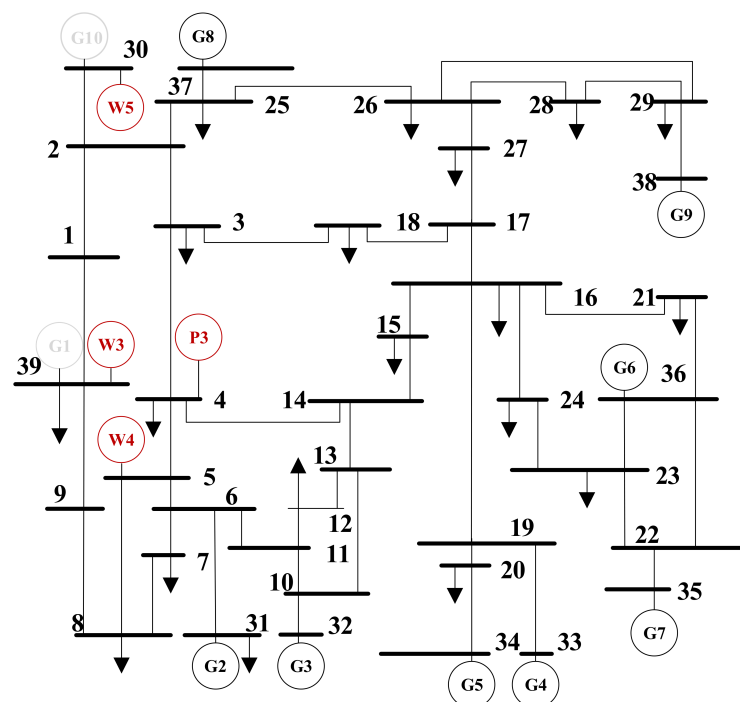


Figure 15. Modified IEEE 39-bus system under the high-renewable-penetration operating condition.

Under this operating condition, multiple fault scenarios are tested to further examine the control performance of the proposed strategy. The quantitative results of the proposed CMDP-based strategy and the conventional MDP-based strategy under different fault scenarios in the high-renewable-penetration operating condition are summarized in Tables 6 and 7.

Table 6. Effectiveness of the proposed CMDP-based strategy under different fault scenarios in the high-renewable-penetration operating condition.

Fault	Control Measure	Total Load Shedding Amount (MW)	Frequency Nadir (Hz)	Voltage (W3/W4/W5/P3)
G3	Load 04, Load 23, Load 15	280.00	49.22	1.04/1.01/0.99/1.01
G4	Load 20, Load 16, Load 15	418.10	49.21	1.05/1.03/1.01/1.01
G5	Load 20, Load 16, Load 15	127.70	49.24	1.05/1.00/1.00/1.01
G6	Load 20, Load 24	343.24	49.22	1.04/1.02/1.00/1.02
G8	Load 23	111.38	49.22	1.04/1.02/1.00/1.01
G9	Load 24, Load 16, Load 23	317.06	49.24	1.05/1.03/1.00/1.02

Table 7. Effectiveness of the conventional MDP-based strategy under different fault scenarios in the high-renewable-penetration operating condition.

Fault	Control Measure	Total Load Shedding Amount (MW)	Frequency Nadir (Hz)	Voltage (W3/W4/W5/P3)
G3	Load 04, Load 23, Load 20	294.75	49.25	1.08/1.04/1.00/1.02
G4	Load 16, Load 15, Load 24	439.85	49.26	1.11/1.06/1.02/1.04
G5	Load 20, Load 16, Load 23	136.95	49.27	1.09/1.04/1.01/1.03
G6	Load 20, Load 24, Load 23	359.44	49.26	1.10/1.05/1.01/1.04
G8	Load 23, Load 16	118.98	49.25	1.08/1.04/1.01/1.03
G9	Load 24, Load 23, Load 20	334.56	49.28	1.12/1.07/1.02/1.04

As shown in Tables 6 and 7 under the high-renewable-penetration operating condition, the conventional MDP-based strategy exhibits a slight advantage in terms of the frequency nadir for all tested fault scenarios. However, this improvement is achieved at the expense of a slightly higher total load shedding cost and a less satisfactory voltage security performance at renewable energy grid-connection points. By contrast, the proposed CMDP-based strategy maintains the monitored renewable energy terminal voltages within or closer to the allowable operating range while achieving slightly lower or comparable load shedding costs in all tested fault scenarios. These results further indicate that, although MDP may achieve slightly better frequency recovery, CMDP provides better overall control performance when frequency stability, voltage security, and load shedding economy are considered together.

Therefore, the additional validation under the high-renewable-penetration operating condition provides further support for the applicability of the proposed CMDP-based strategy beyond the original base operating condition.

Nevertheless, the present revision still does not include comparisons with a broader set of recent benchmark methods under a fully consistent simulation framework. Some recent methods are developed under different modeling assumptions and implementation settings, and some do not provide sufficiently detailed public implementations or fully matched simulation settings for direct reproduction in our framework. A broader comparison under a unified framework will therefore be an important direction for future work.

6. Conclusions

To satisfy the secure load-shedding requirements of new-type power systems, this paper proposes a CMDP-based emergency load-shedding-control strategy that simultaneously considers frequency security and voltage stability.

- (1) By introducing penalty terms, the CMDP-based decision-making model transforms complex security constraints into an unconstrained optimization problem. Combined with a PPO-optimized actor–critic agent, the proposed method improves both training convergence and decision-making efficiency.
- (2) Compared with the conventional contingency-based strategy, the proposed security-constrained intelligent load-shedding method can effectively reduce the total load shedding amount while maintaining frequency stability and preserving the voltage security of renewable energy units.
- (3) Compared with the conventional MDP-based strategy, the CMDP framework with explicit security boundaries provides better overall control performance in maintaining renewable energy voltage security while ensuring system frequency stability, thereby effectively mitigating the risks associated with voltage violations.

It should also be acknowledged that the present study mainly evaluates renewable energy voltage security using terminal voltage thresholds derived from grid-connection requirements. Although this treatment is suitable for emergency load-shedding decision making on the electromechanical time scale, it cannot fully represent all detailed dynamics of converter-dominated renewable power systems, such as converter inner-loop control, phase-locked-loop dynamics, current limiting, transient stability interactions, and interarea oscillations. Future work will further incorporate detailed converter-control models and electromagnetic transient or hybrid simulation platforms to improve the dynamic representation of converter-dominated power systems. A broader comparison with stronger recent benchmark methods under a unified experimental framework remains an important direction for future work.

Author Contributions: Q.F.: conceptualization, project administration. Z.H.: formal analysis, investigation. W.H.: data curation, writing—original draft. Y.J.: funding acquisition, methodology. Z.L. and W.C.: resources, software. M.L. and J.G.: supervision, visualization. R.Z.: validation, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the State Grid Liaoning Electric Power Co., Ltd. Economic and Technological Research Institute (SGLNJJY00GHJS2500051).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: Author Qiushi Fang, Zhentao Han, Yufei Jin, Mingxuan Lu and Jiawen Gao were employed by the company State Grid Liaoning Electric Power Co., Ltd. Economic and Technological Research Institute. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The authors declare that this study received funding from State Grid Liaoning Electric Power Co., Ltd. Economic and Technological Research Institute. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

References

1. Huang, G. Simulation Study on Suppression Measures for Commutation Failure in UHVDC Transmission System. Master's Thesis, Chongqing University of Technology, Chongqing, China, 2023. (In Chinese) [[CrossRef](#)]
2. Chen, W. Research on Emergency Load Shedding Strategy for Power System Frequency Stability Based on Safe Reinforcement Learning. Master's Thesis, Northeast Electric Power University, Jilin, China, 2024. (In Chinese) [[CrossRef](#)]
3. Zhang, L.; Ye, T.; Xin, Y.; Han, F.; Fan, G. Problems and measures of power grid accommodating large scale wind power. *Proc. CSEE* **2010**, *30*, 1–9. (In Chinese) [[CrossRef](#)]

4. Chan, M.L.; Dunlop, R.D.; Schweppe, F. Dynamic Equivalents for Average System Frequency Behavior Following Major Disturbances. *IEEE Trans. Power Appar. Syst.* **1972**, PAS-91, 1637–1642. [\[CrossRef\]](#)
5. Anderson, P.M.; Mirheydar, M. A low-order system frequency response model. *IEEE Trans. Power Syst.* **1990**, 5, 720–729. [\[CrossRef\]](#)
6. Xu, T.; Li, C.; Liu, Y.; Liu, C.; Su, D.; Xu, C.; Wu, H. Optimization of emergency load shedding based on cultural particles warm optimization algorithm. In *2019 IEEE Congress on Evolutionary Computation (CEC)*, Wellington, New Zealand, 10–13 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1208–1212. [\[CrossRef\]](#)
7. Wang, Z.; Wang, Z.; Liang, Z.; Chai, Y.; Yu, R.; Bian, J. Research on emergency load shedding strategy for transmission and distribution networks with high proportion of distributed photovoltaic generation. *Power Syst. Clean Energy* **2024**, 40, 40–51+61. (In Chinese)
8. Chen, Y.; Liao, S.; Xu, J. Emergency load-shedding optimization control method based on reinforcement learning assistance. *Energy Rep.* **2022**, 08, 1051–1061. [\[CrossRef\]](#)
9. Chen, C.; Cui, M.; Li, F.; Yin, S.; Wang, X. Model-Free emergency frequency control based on reinforcement learning. *IEEE Trans. Ind. Electron.* **2020**, 17, 2336–2346. [\[CrossRef\]](#)
10. Huang, Q.; Huang, R.; Hao, W.; Tan, J.; Fan, R.; Huang, Z. Adaptive power system emergency control using deep reinforcement learning. *IEEE Trans. Smart Grid* **2020**, 11, 1171–1182. [\[CrossRef\]](#)
11. Chen, H.; Zhuang, J.; Zhou, G.; Wang, Y.; Sun, Z.; Levron, Y. Emergency load shedding strategy for high renewable-energy penetrated power systems based on deep reinforcement learning. *Energy Rep.* **2023**, 9, 434–443. [\[CrossRef\]](#)
12. Hu, Z.; Shi, Z.; Zeng, L.; Yao, W.; Tang, Y.; Wen, J. Knowledge-enhanced deep reinforcement learning for intelligent event-based load shedding. *Int. J. Electr. Power Energy Syst.* **2023**, 148, 108978. [\[CrossRef\]](#)
13. Song, S.; Jung, Y.; Jang, G.; Jung, S. Proximal policy optimization through a deep reinforcement learning framework for remedial action schemes of VSC-HVDC. *Int. J. Electr. Power Energy Syst.* **2023**, 150, 109117. [\[CrossRef\]](#)
14. Li, B.; Wang, H.; Li, J.; Lu, G. Integrating model-driven and data-driven methods for under-frequency load shedding control. *Electr. Power Syst. Res.* **2025**, 238, 111103. [\[CrossRef\]](#)
15. Xie, J.; Sun, W. Distributional deep reinforcement learning-based emergency frequency control. *IEEE Trans. Power Syst.* **2022**, 37, 2720–2730. [\[CrossRef\]](#)
16. Vu, T.L.; Mukherjee, S.; Yin, T.; Huang, R.; Tan, J.; Huang, Q. Safe reinforcement learning for emergency load shedding of power systems. *arXiv* **2020**, arXiv:2011.09664. [\[CrossRef\]](#)
17. Vu, T.L.; Mukherjee, S.; Huang, R.; Huang, Q. Barrier function-based safe reinforcement learning for emergency control of power systems. In *2021 60th IEEE Conference on Decision and Control (CDC)*; IEEE: Piscataway, NJ, USA, 2021. [\[CrossRef\]](#)
18. Yu, P.; Zhang, H.; Song, Y.; Wang, Z.; Dong, H.; Ji, L. Safe reinforcement learning for power system control: A review. *Renew. Sustain. Energy Rev.* **2025**, 223, 116022. [\[CrossRef\]](#)
19. Wachi, A.; Shen, X.; Sui, Y. A survey of constraint formulations in safe reinforcement learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence; IJCAI*: Darmstadt, Germany, 2024; pp. 8262–8271. [\[CrossRef\]](#) [\[PubMed\]](#)
20. GB/T 36995-2018; Test Procedure of Fault Voltage Ride-Through Capability for Wind Turbines. Standards Press of China: Beijing, China, 2019. Available online: <https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=A9D6F07EB8A70DEFA2974C513C6F3325> (accessed on 4 June 2026). (In Chinese)
21. GB/T 19939-2005; Technical Requirements for Grid Connection of Photovoltaic Systems. Standards Press of China: Beijing, China, 2005. Available online: <https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=6AE22AC243342D4698C74F1A5BEDC4C2> (accessed on 4 June 2026). (In Chinese)
22. Xu, Z.; Xiao, H.; Zhang, Z.; Xue, Y.; Liu, G. *Flexible HVDC Transmission Systems*, 2nd ed.; China Machine Press: Beijing, China, 2017. (In Chinese)
23. Tang, Y.; Li, B.; Zeng, R.; Jiang, Q.; Liu, T. Analysis of influencing factors of VSC-HVDC on short-circuit current of AC systems. *Electr. Power Eng. Technol.* **2022**, 41, 40–47. (In Chinese) [\[CrossRef\]](#)
24. Zhu, L.; Kou, L.; Fan, Z.; Gu, H. Influence of VSC-HVDC integration into urban power grids on short-circuit current near converter stations. *Glob. Energy Interconnect.* **2023**, 6, 139–148. (In Chinese) [\[CrossRef\]](#)
25. Sun, Y.F.; Xiong, X.; Wang, Z.H.; Wang, W.R. Analysis method of fault current in flexible DC power grid considering AC infeed and steady-state operating conditions. *South. Power Syst. Technol.* **2021**, 15, 1–11. [\[CrossRef\]](#)
26. Zhang, Q.; Xin, C.; Li, F.; Yin, C.; Liu, J.; Li, Y. Multi-DC coordinated transient overvoltage suppression strategy for high-proportion renewable-energy sending-end areas. *Electr. Power Eng. Technol.* **2023**, 42, 98–106. (In Chinese) [\[CrossRef\]](#)
27. Zhou, J.; Guo, Y.; Yang, L.; Shi, J.; Zhang, Y.; Li, Y.; Guo, Q.; Sun, H. A review on frequency management for low-inertia power systems: From inertia and fast frequency response perspectives. *Electr. Power Syst. Res.* **2024**, 228, 110095. [\[CrossRef\]](#)

28. Su, L.; Qin, X.; Zhang, S.; Zhang, Y.; Jiang, Y.; Han, Y. Fast frequency response of inverter-based resources and its impact on system frequency characteristics. *Glob. Energy Interconnect.* **2020**, *3*, 475–485. [[CrossRef](#)]
29. Sun, Z. Research on Online Decision-Making Method for Distributed-Centralized Low-Voltage Load Shedding. Master's Thesis, Harbin Institute of Technology, Harbin, China, 2017. (In Chinese)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.