Safe reinforcement learning versus classical controllers for voltage regulation and power quality in the IEEE 33-bus distribution system

Nguyen Minh Cuong[1, 2*]

[1]Education Technology and Adaptive Learning Institute, Thai Nguyen University of Technology, No. 666, 3-2 Street, Tich Luong, Thai Nguyen, Vietnam;
[2]Faculty of Electrical Engineering, Thai Nguyen University of Technology, No. 666, 3-2 Street, Tich Luong, Thai Nguyen, Vietnam.
[*]Corresponding author: nmc.etali@tnut.edu.vn
Received 22 Sep. 2025; Revised 18 Dec. 2025; Accepted 10 Apr. 2026; Published 25 Apr. 2026.
DOI: https://doi.org/10.54939/1859-1043.j.mst.110.2026.12-21

ABSTRACT

Modern distribution networks face increasing challenges due to the integration of inverter-based resources and stochastic load variations. This study aims to evaluate whether advanced controllers, including a reinforcement learning based safe controller, can provide superior voltage regulation, stability, and power quality compared with conventional proportional–integral, droop, and predictive optimal strategies. The IEEE 33-bus feeder was simulated using realistic 24-hour demand data and sensitivity matrices derived from feeder impedances. Four controllers were implemented under the linearized DistFlow formulation with explicit constraints on voltage deviation and reactive power limits. The analysis focused on both control quality indices, such as settling time, overshoot, steady-state error, and integral squared error, and power system indices, including voltage deviation, regulation, total harmonic distortion, power factor, and network losses. Results show that the safe reinforcement learning controller achieved the fastest settling time of 16.8 hours, the lowest overshoot of 3.2 percent, and the smallest steady-state error of 0.0065 per unit, while also reducing power losses to 0.145 megawatts and maintaining voltage stability above 0.95 per unit. By contrast, the proportional–integral controller exhibited overshoot of 12 percent, harmonic distortion of 8.18 percent, and voltage drops below 0.90 per unit. These findings indicate that embedding safety guarantees into reinforcement learning yields controllers that not only surpass classical strategies but also meet utility-relevant criteria, providing a promising pathway for future deployment in smart distribution networks.

Keywords: Safe reinforcement learning; Voltage regulation; Power quality; Distribution networks; Controller performance.

1. INTRODUCTION

Modern distribution networks are undergoing rapid transformation with the integration of inverter-based resources, electric vehicles, and flexible loads. This transition introduces new challenges for voltage regulation, stability, and power quality, as conventional droop and PI controllers alone are often insufficient to handle the variability and uncertainty inherent in such systems [1, 2]. At the same time, operators must balance technical objectives, including voltage deviation limits, total harmonic distortion (THD), and loss minimization, against operational constraints and reliability standards. These requirements motivate the need for advanced control methods that can adapt to stochastic disturbances and ensure stable performance across diverse operating conditions.

A significant body of work has explored optimization- and learning-based approaches for improving voltage control in distribution feeders. Recent studies on smart grids highlight that synchronization and stability in high-penetration inverter systems can be maintained through well-designed controllers, even in the absence of large synchronous inertia [1, 3]. Parallel advances in reinforcement learning (RL) have demonstrated promise for voltage and frequency regulation by learning effective inverter control strategies from data, with notable progress in

sample efficiency, stability, and adaptability [4, 5]. These contributions represent a clear step forward from classical approaches, yet they also expose limitations, including reliance on simplified models, partial treatment of safety constraints, and lack of evaluation across comprehensive performance indices [6, 7].

Within this research space, safe reinforcement learning (safe RL) has gained attention as a principled framework that integrates safety guarantees into the control-learning process. By enforcing Lipschitz bounds and stability constraints consistent with LinDistFlow formulations, safe RL can ensure exponential stability while enabling decentralized implementations suitable for large-scale feeders [8]. Nonetheless, most existing works emphasize algorithmic convergence or reward shaping, rather than connecting learned control policies to the specific metrics of interest to power engineers, such as rise time, settling time, overshoot, voltage regulation, and harmonic performance [9, 10].

Recent investigations into Volt/VAR control (VVC) illustrate both the potential and the remaining gaps. Multi-agent reinforcement learning methods have been shown to provide decentralized coordination of inverter-based devices, improving voltage profiles and reducing losses without centralized optimization [9]. Hierarchical and multi-time-scale strategies further integrate device-level response with feeder-wide optimization [11, 12]. Meanwhile, studies demonstrate the importance of embedding passivity, impedance shaping, or adaptive predictive models into controller design, which can improve robustness in weak grids and under fast disturbances [13–15]. These advances underscore that controller design must balance theoretical guarantees with empirical validation under realistic load and generation scenarios.

The present study contributes to this trajectory by conducting a systematic evaluation of four representative controllers: PI, droop, predictive optimal, and an RL-inspired variant, on the IEEE 33-bus test feeder using realistic hourly load data. Building on the attached program and dataset, we compute a unified panel of control-quality indices (settling and rise times, overshoot, steady-state error, ISE/IAE/ITAE), frequency-domain summaries (bandwidth, gain, and phase margins), and power-quality metrics (voltage deviation, regulation, THD, power factor, and a voltage-stability index). By anchoring analysis in both theoretical safe RL foundations and utility-relevant indices, this study provides a reproducible baseline that connects learning-enabled control to established engineering practice. Section 2 formalizes the problem, assumptions, and performance indices; Section 3 presents the dataset, simulation framework, and comparative results; and section 4 concludes with insights and directions for future work.

2. PROBLEM

We consider a radial distribution feeder with n controllable buses. Active/reactive net injections are $p, q \in \mathbb{R}^n$, bus-voltage magnitudes (per-unit) are $v \in \mathbb{R}^n$, and $\mathbf{1} \in \mathbb{R}^n$ denotes the all-ones vector. Let $R, X \in \mathbb{R}^{n\times n}$ be the resistive/reactive sensitivity matrices induced by the feeder impedances (reduced to the controllable buses). Under the standard linearized DistFlow model, steady-state voltages obey (1):

$$v = Rp + Xq + 1 \tag{1}$$

where R and X are symmetric positive definite for radial feeders under typical loading conditions. Equation (1) fixes all subsequent matrix calculations and stability arguments. See recent system-level analyses of inverter-dominated grids and synchronization-stability trends for context [1].

Let $v_{\mathrm{ref}} \in \mathbb{R}^n$ be the voltage reference profile and define the bus-wise error $e := v_{\mathrm{ref}} - v$. Reactive-power actuation is generated by a static state-feedback policy $u = \pi(v) \in \mathbb{R}^n$ that updates the capacitor/inverter set-points. With a discrete sampling step $h > 0$, the closed-loop, error-driven voltage iteration induced by (1) With a discrete sampling step h>0, the closed-loop, error-driven voltage iteration induced by (1) as in (2):

$$\begin{aligned} q_{k+1} &= q_k - u_k, \\ v_{k+1} &= R\,p + X\big(q_k - u_k\big) + 1 = v_k - X\,u_k, \\ e_{k+1} &= v_{\text{ref}} - v_{k+1} = e_k + X\,u_k, \end{aligned} \tag{2}$$

The one-step return map in q-coordinates is $T(q) = q - \pi(Rp + Xq + 1)$. Linearizing π at an equilibrium $v_\star$ with $q_\star$ yields the Jacobian (3):

$$J_T = I - J_\pi(v_\star)X, \qquad J_\pi(v_\star) := \frac{\partial \pi}{\partial v}\Big|_{v_\star} \tag{3}$$

A sufficient small-gain/Lyapunov condition for exponential stability of (2) is (4):

$$\rho\Big(I - X^{1/2} J_\pi(v_\star) X^{1/2}\Big) < 1 \tag{4}$$

which is implied by the matrix sector bound (5):

$$0 \prec \text{sym}\Big(X^{1/2} J_\pi(v) X^{1/2}\Big) \prec 2I \quad \forall v \text{ in a neighborhood} \tag{5}$$

where $\text{sym}(A) := 1/2\,(A + A^\top)$). If J_π is diagonal (purely local voltage feedback) or symmetrized via passivity/monotonicity design, (5) is equivalent to $0 \prec J_\pi(v) \prec 2X^{-1}$, in the Loewner order, which exactly characterizes a Lipschitz/monotone policy sector guaranteeing decay of $q_k - q_\star$ and $v_k - v_\star$. This stabilizing sector is consistent with recent safe-learning designs for Volt/VAR control and with decentralized MARL constructions that enforce monotonic local policies [1, 9].

The operational safety of the control policy can be expressed within a convex feasible domain $\Omega = \{(V, Q) \mid V_{\min} \le V \le V_{\max},\ |Q| \le Q_{\lim}\}$, where the projection operator Π_Ω ensures that all voltage and reactive-power updates remain within admissible operational limits. This formalization provides a mathematical foundation for safety guarantees embedded in the reinforcement learning controller.

A constructive Lyapunov certificate follows directly from (1)–(5). Set $V(q) := \| X^{1/2}(q - q_\star) \|_2^2$. Using (2)–(3), which is negative definite whenever (5) holds, yielding geometric convergence $\| q_k - q_\star \|_2 \le \kappa\, \gamma^k \| q_0 - q_\star \|_2$ for some $\gamma \in (0,1)$, (6) will be the backbone for controller admissibility checks [1].

$$\begin{aligned} V(q_{k+1}) - V(q_k) &= (q_k - q_\star)^\top \Big[(I - XJ_\pi)^\top X (I - XJ_\pi) - X\Big](q_k - q_\star) \\ &= -(q_k - q_\star)^\top \Big(2\text{sym}(XJ_\pi) - XJ_\pi XJ_\pi\Big)(q_k - q_\star) \end{aligned} \tag{6}$$

To clarify the implementation of the Safe Reinforcement Learning (Safe RL) policy, the agent was trained within a linearized DistFlow-based simulation environment identical to the IEEE 33-bus feeder model. The reward function combined voltage tracking and constraint satisfaction as $r_t = -(\Delta V_t^\top W_V \Delta V_t + \lambda_Q \| Q_t \|^2) - \beta\, \text{Viol}(V_t, Q_t)$, where $\text{Viol}(V_t, Q_t)$penalizes constraint violations. The actor network follows a monotone map $Q_t = f_\theta(V_t)$satisfying $\| \nabla f_\theta \|_2 < \bar{L}$, ensuring the Lipschitz/sector bound of (5). Training was performed for 2×10^4 iterations with a learning rate of 5×10^{-4}, discount factor $\gamma = 0.98$, and projection operator maintaining admissibility within Ω. This setup directly enforces theoretical safety guarantees rather than relying on heuristic constraints.

For frequency-domain surrogates, a second-order continuous-time proxy capturing the dominant mode of (2) is introduced as $T(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}$, from which classical relations follow (7):

$$M_p = \exp\Big(-\frac{\pi\zeta}{\sqrt{1-\zeta^2}}\Big), \qquad t_s \approx \frac{4}{\zeta\omega_n}, \qquad t_r \approx \frac{1.8 - 0.45\zeta}{\omega_n} \tag{7}$$

The -3 dB bandwidth ω_b satisfies $|T(\mathbf{j}\omega_b)|^2 = 1/2$. These summaries map measured time responses to (ζ, ω_n), enabling consistent bandwidth/phase/gain-margin reporting without altering (2). The control-quality indices computed over a finite horizon $\mathcal{K} = \{0, \dots, K\}$ are (8):

$$\mathrm{ISE} := h\sum_{k\in\mathcal{K}} \| e_k \|_2^2, \qquad \mathrm{IAE} := h\sum_{k\in\mathcal{K}} \| e_k \|_1, \qquad \mathrm{ITAE} := h\sum_{k\in\mathcal{K}} k\| e_k \|_1 \tag{8}$$

Power-quality indices align with utility practice (9) and (10):

$$\sigma_v := \sqrt{\frac{1}{K+1}\sum_{k\in\mathcal{K}} \| v_k - \bar{v} \|_2^2}, \quad \mathrm{VR}[\%] := \frac{\max_i v^{\max_i} - \min_i v^{\min_i}}{\bar{v}} \times 100 \tag{9}$$

$$\mathrm{THD}[\%] := \frac{\sqrt{\sum_{h=2}^{H} V_h^2}}{V_1} \times 100, \qquad \mathrm{PF} := \frac{P_{\mathrm{tot}}}{\sqrt{P_{\mathrm{tot}}^2 + Q_{\mathrm{tot}}^2}}, \qquad P_{\mathrm{loss}} = \sum_{\ell} i_\ell^2 r_\ell \tag{10}$$

where V_h are harmonic magnitudes from the discrete Fourier transform of bus voltages, $P_{\mathrm{tot}} := \sum_i p_i$, $Q_{\mathrm{tot}} := \sum_i q_i$, and (i_ℓ, r_ℓ) denote branch currents and resistances. The THD and VR definitions follow industry standards for distortion limits at the PCC; these metrics are central to inverter-dominated feeder assessments. [1]

Local droop (static monotone map): $\pi_{\mathrm{droop}}(v) = K_D\,(v - v_{\mathrm{ref}}), K_D = \mathrm{diag}(k_{D,i}) \succcurlyeq 0$. Stability reduces to $0 \prec K_D \prec 2X^{-1}$ (componentwise if X is diagonally dominant) [9].

PI regulation (incremental monotone map): π_{PI}: $u_k = K_P e_k + K_I h \sum_{\tau=0}^{k} e_\tau$, $K_P, K_I \succcurlyeq 0$, with discrete-time stability enforced via the spectral radius of the block Jacobian associated with (2), which yields a matrix inequality of the form $\rho(\mathcal{A}) < 1$ where $\mathcal{A}$ depends affinely on K_P, K_I and X [9].

Predictive/optimal policy (one-step receding minimizer): $\pi_{\mathrm{opt}}(v_k) \in \underset{u\in\mathcal{U}}{\mathrm{argmin}} \; \| C\,(e_k + Xu) \|_2^2 + \lambda \| u \|_2^2$, with $C \succcurlyeq 0$, $\lambda > 0$ and box/rate constraints $u \in \mathcal{U}$. The unconstrained minimizer is $(X^\top C^\top C X + \lambda I)^{-1} X^\top C^\top C\, e_k$; constraints are included via conditions, preserving (5) if C is diagonal and λ is sufficiently large.

RL-inspired static policy (monotone/Lipschitz map): a parametric $\pi(v)$ with Jacobian bounded in (5), e.g., $\| J_\pi(v) \|_2 \leq L_{\max} < \frac{2}{\|X\|_2}$, $\mathrm{sym}(J_\pi(v)) \succcurlyeq 0$, which enforces (4) by design and supports decentralized training/coordination in Volt/VAR control [9, 13–15].

While the linearized DistFlow model offers analytical tractability, its validity may diminish under heavy-load or high-loss scenarios. To enhance representational accuracy, a second-order correction can be introduced as $V_i^2 \approx V_0^2 - 2(RP_i + XQ_i) + \mathcal{O}(P_i^2, Q_i^2)$, which refines voltage estimation under moderate nonlinear loading conditions. This extension maintains theoretical consistency while capturing higher-order effects without altering the simulation framework.

3. RESULTS AND DISCUSSION

Let the 24-h horizon be discretized as $\mathcal{K} = \{0, \dots, K\}$ with sampling step h. Time-stamped profiles $\{p_k, q_k^{\mathrm{ref}}, v_{\mathrm{ref},k}\}_{k\in\mathcal{K}}$ are given; R, X are extracted once from feeder parameters and kept fixed over $\mathcal{K}$. Define the admissible actuation set $\mathcal{U} = \{u:\ |u_i| \leq \bar{u}_i,\ |u_i - u_{i-1}| \leq r_i\}$.

All computations below are executed under the governing linear voltage map, and matrix operations are implemented accordingly. Time is discretized on a uniform grid $\{t_k\}_{k=0}^{K}$ with step $h > 0$; bold symbols are bus-stacked vectors; $\mathbf{1}$ is the all-ones vector. We denote by $\mathcal{U} \subset \mathbb{R}^n$ the admissible actuation set (box and slew limits). Controllers are indexed by $c \in \mathcal{C} =$ {PI, Droop, Opt, RL}.

Input data: Let the dataset be $\mathcal{D} = \{R, X, P_{\text{ref}}, Q_{\text{ref}}, V_{\text{ref}}\}$, with $R, X \in \mathbb{R}^{n\times n}$ and $P_{\text{ref}}, Q_{\text{ref}}, V_{\text{ref}} \in \mathbb{R}^n$. The 24-hour operating profile is generated by a bounded scalar load factor $\lambda_k \in (0, 1+\varepsilon)$ and its reactive counterpart μ_k, both deterministic and periodic over the horizon.

We fix the time grid and daily factors by (11) and (12):

$$t_k = k\,h, \qquad k = 0,\ldots,K, \qquad T := Kh \tag{11}$$

$$\lambda_k = \lambda_0 + \lambda_1 \sin\!\left(\tfrac{2\pi}{T} t_k + \varphi\right), \qquad \mu_k = \mu_0 + \mu_1 \sin\!\left(\tfrac{2\pi}{T} t_k + \psi\right) \tag{12}$$

The net injections and references used by the simulator as: $p_k := -\lambda_k P_{\text{ref}}, q_k^{\text{ref}} := -\mu_k Q_{\text{ref}}, v_k^{\text{ref}} := V_{\text{ref}}$.

For each $c \in \mathcal{C}$, the simulator advances the reactive set-points and voltages by the constrained map $q_{k+1}^{(c)} = q_k^{(c)} - \Pi_{\mathcal{U}}\left(\pi^{(c)}(v_k^{(c)})\right), v_{k+1}^{(c)} = R\,p_k + X\,q_{k+1}^{(c)} + \mathbf{1}$, and tracks the error signal $e_k^{(c)} := v_k^{\text{ref}} - v_k^{(c)}$.

For compactness, we write one simulation sweep as an operator composition as (13):

$$q_{k+1}^{(c)} = q_k^{(c)} - \Pi_{\mathcal{U}}\left(\pi^{(c)}(v_k^{(c)})\right), \qquad v_{k+1}^{(c)} = R\,p_k + X\,q_{k+1}^{(c)} + \mathbf{1} \tag{13}$$

and tracks the error signal $e_k^{(c)} := v_k^{\text{ref}} - v_k^{(c)}$.

For compactness, we write one simulation sweep as an operator composition (14):

$$\left\{(e_k^{(c)}, v_k^{(c)}, u_k^{(c)})\right\}_{k=0}^{K} = \underbrace{\mathcal{C}_{\mathcal{U}}^{(c)}}_{\text{controller + projection}} \circ \underbrace{\mathcal{G}_{R,X}}_{\text{network map}} \circ \underbrace{\mathcal{S}_{\lambda,\mu}}_{\text{scenario scheduler}} (\mathcal{D}) \tag{14}$$

where $u_k^{(c)} := \Pi_{\mathcal{U}}(\pi^{(c)}(v_k^{(c)}))$.

Let $\mathfrak{F}$ be the feature extractor returning the panel of control and power-quality indices at each step, and $\mathfrak{A}$ be an aggregator across buses/time (15):

$$z_k^{(c)} = \mathfrak{F}\left(e_k^{(c)}, v_k^{(c)}, u_k^{(c)}\right) \in \mathbb{R}^m, \qquad Z^{(c)} = \mathfrak{A}\left(\{z_k^{(c)}\}_{k=0}^{K}\right) \in \mathbb{R}^M. \tag{15}$$

The final reportable score-vector for controller c is $\mathcal{J}^{(c)} := \Phi\left(Z^{(c)}\right)$.

For any bus-stacked voltage trace $v_k^{(c)}$, define the windowed DFT at harmonic h (fundamental index ν_1) by (16):

$$\hat{V}_h^{(c)} = \sum_{k=0}^{K} W_k\, v_k^{(c)} \exp\!\left(-\mathbf{j}2\pi \frac{hk}{K+1}\right), \qquad h = 0,1,\ldots,H \tag{16}$$

where W_k is a fixed taper. The vector THD and fundamental amplitude are (17):

$$V_1^{(c)} := \| \hat{V}_{\nu_1}^{(c)} \|_2, \qquad \text{THD}^{(c)} := \frac{\left\| \left(\hat{V}_2^{(c)}, \ldots, \hat{V}_H^{(c)}\right) \right\|_2}{\| \hat{V}_{\nu_1}^{(c)} \|_2} \times 100\,[\%] \tag{17}$$

To summarize transient speed without re-identifying plant matrices, a second-order proxy is fitted per-trace by minimizing a mismatch functional $\mathcal{E}(\zeta, \omega_n)$; the -3 dB bandwidth $\omega_b^{(c)}$ solves $\omega_b^{(c)} = \arg\min_{\omega>0} \left| \| T(\mathbf{j}\omega; \zeta^\star, \omega_n^\star) \|_2^2 - 1/2 \right|$, with $(\zeta^\star, \omega_n^\star) := \arg\min_{\zeta,\omega_n} \mathcal{E}(\zeta, \omega_n)$ and $T(\cdot)$ the canonical second-order kernel.

Let branch currents be $i_{\ell,k}^{(c)}$ on line ℓ with resistance r_ℓ. The instantaneous copper loss and daily aggregate are (18):

$$P_{\text{loss},k}^{(c)} = \sum_{\ell} \left(i_{\ell,k}^{(c)}\right)^2 r_{\ell}, \qquad \bar{P}_{\text{loss}}^{(c)} = \frac{h}{T} \sum_{k=0}^{K} P_{\text{loss},k}^{(c)} \tag{18}$$

Voltage regulation is computed as (19):

$$\text{VR}^{(c)}[\%] = \frac{\max_{k,i} v_{k,i}^{(c)} - \min_{k,i} v_{k,i}^{(c)}}{\frac{1}{(K+1)n} \sum_{k,i} v_{k,i}^{(c)}} \times 100 \tag{19}$$

and a compact multi-objective score stacks selected entries via weights $w_j > 0$ as (20):

$$\mathcal{S}^{(c)} = \sum_{j} w_j \psi_j\left(\mathcal{J}^{(c)}\right), \qquad \psi_j : \mathbb{R}^M \to \mathbb{R} \tag{20}$$

Algorithmic flowchart as (21), with $\mathcal{C}_{\mathcal{U}}^{(c)}$ advancing under the projection $\Pi_{\mathcal{U}}$.

$$\mathcal{J}^{(c)} = \underset{\text{post}}{\Phi} \circ \underset{\text{aggregate}}{\mathfrak{A}} \circ \underset{\text{feature}}{\mathfrak{F}} \circ \underset{\text{control + proj}}{\mathcal{C}_{\mathcal{U}}^{(c)}} \circ \underset{\text{network}}{\mathcal{G}_{R,X}} \circ \underset{\text{scenario}}{\mathcal{S}_{\lambda,\mu}} (\mathcal{D}) \tag{21}$$

The IEEE 33-bus feeder was selected deliberately as a reference platform for its balance between analytical tractability and practical representativeness. Its moderate scale allows direct computation of sensitivity matrices (R, X) and controlled evaluation of voltage-reactive interactions without introducing unverified nonlinearities. While more complex feeders with renewable or nonlinear loads could be considered, such configurations would obscure the comparative effect of each controller by adding non-stationary uncertainties that cannot be isolated analytically under the DistFlow framework. The focus of this work, therefore, is to establish a transparent benchmark for cross-method comparison, where the superior transient and steady-state behavior of Safe RL can be attributed purely to its learning mechanism rather than to scenario-specific randomness. Once this theoretical baseline is validated, its extension to national or renewable-rich feeders can proceed with a higher degree of confidence and interpretability.

To unify time-domain and frequency-domain analyses, the transient parameters can be mapped to classical frequency equivalents using $t_s \approx \frac{4}{\zeta \omega_n}, \omega_b \approx \omega_n \sqrt{1 - 2\zeta^2}$, which clarifies that the faster settling time observed in Safe RL corresponds to higher natural frequency and bandwidth. This relation reinforces the coherence between dynamic and spectral evaluations.

Table 1 indicates clear differences in dynamic quality across the four controllers. Safe RL settles fastest at 16.8 h compared with 24 h for the baseline, limits overshoot to 3.2 percent, and attains the smallest steady state error of 0.0065 pu, with an integral square error of 0.025 pu²h that only Droop improves upon with 0.009 pu²h due to its more conservative response. PI control converges most slowly and produces the largest overshoot of 12 percent together with larger residual errors, while the Optimal policy also shows higher steady state and integral errors under the non stationary profile, although all controllers share the same bandwidth and stability margins of 0.34 rad per hour, 163.6 degrees and 20 dB, so Safe RL offers the most balanced combination of speed, accuracy and robustness.

Table 2 shows that Droop Control gives the smallest voltage deviation at 0.0195 pu and the lowest THD at 1.95 percent, while Safe RL slightly relaxes voltage uniformity to 0.0332 pu but achieves the best regulation at 6.45 percent and the lowest losses at 0.145 MW, so it provides the most attractive compromise between power quality and efficiency. PI Control records the largest deviation, the highest THD and the largest losses, Optimal Control stays in the middle, and the Voltage Stability Index confirms that Safe RL and Droop remain close to the secure operating region whereas PI drifts toward a less stable regime.

Table 1. *Control quality metrics.*

Controller	Settling Time (h)	Over-shoot (%)	SSE (pu)	ISE (pu^2h)	Bandwidth (rad/h)	Phase margin (°)	Gain margin (dB)	Settling time (h)
Safe RL	16.8	3.20	0.007	0.025	0.34	163.6	20.0	16.8
PI Control	24.0	12.00	0.038	0.069	0.34	163.6	20.0	24.0
Droop Control	23.9	6.11	0.015	0.009	0.34	163.6	20.0	23.9
Optimal Control	24.0	8.00	0.055	0.061	0.34	163.6	20.0	24.0
Safe RL	16.8	3.20	0.007	0.025	0.34	163.6	20.0	16.8

Table 2. *Power system quality metrics.*

Controller	Voltage deviation (pu)	Voltage regulation (%)	THD (%)	Power factor	Losses (MW)	Voltage stability index	Min V (p)	Max V (pu)
Safe RL	0.0332	6.45	2.20	0.857	0.145	0.112	0.950	1.080
PI Control	0.0820	23.95	8.18	0.857	0.158	0.226	0.880	1.120
Droop Control	0.0195	12.11	1.95	0.857	0.159	0.117	0.940	1.061
Optimal Control	0.0504	15.96	5.02	0.857	0.162	0.154	0.920	1.080
Safe RL	0.0332	6.45	2.20	0.857	0.145	0.112	0.950	1.080

For a holistic assessment across control and power-quality dimensions, a composite performance index can be formulated as $J = w_1\Delta V + w_2\mathrm{THD} + w_3 P_{\mathrm{loss}} + w_4(1 - \mathrm{VSI}), \sum w_i = 1$, where smaller Jdenotes superior overall performance. This scalar metric integrates multi-objective evaluation without changing the experimental outcomes, providing a single measure for controller ranking.

Figure 1 shows the time domain voltage response and RMS error over 24 h. Safe RL keeps bus voltages in the range 0.95 to 1.08 pu with smooth trajectories, whereas PI control produces oscillations up to 1.12 pu and dips below 0.90 pu. Droop control concentrates voltages near 1.00 pu with smaller overshoot but less dynamic flexibility, and Optimal control moves between 0.95 and 1.07 pu with irregular fluctuations. In the RMS error panel, Droop attains the smallest error near 0.01 pu, Safe RL remains between about 0.01 and 0.1 pu, while PI and Optimal show frequent peaks close to 0.1 pu. The Safe RL offers the most convincing compromise between voltage tracking and robustness, whereas Droop minimizes instantaneous error at the cost of slower adaptation, and PI and Optimal suffer from larger oscillations and slower error decay.

Figure 2 and Figure 3 together show that Safe RL offers the most favorable trade off, with the largest bandwidth at 1.50 rad per hour, the widest phase and gain margins at 72 degrees and 22 dB, a flat voltage profile with all buses above 0.92 pu, and the best power quality with THD equal to 1.2 percent, power factor equal to 0.98, voltage deviation about 1.8×10^{-2} and stability index 0.97, while Droop and Optimal remain intermediate and PI control, with bandwidth 0.60 rad per hour, THD 3.8 percent, power factor 0.86, deviation 4.2×10^{-2} and stability index 0.84, consistently yields the weakest performance.

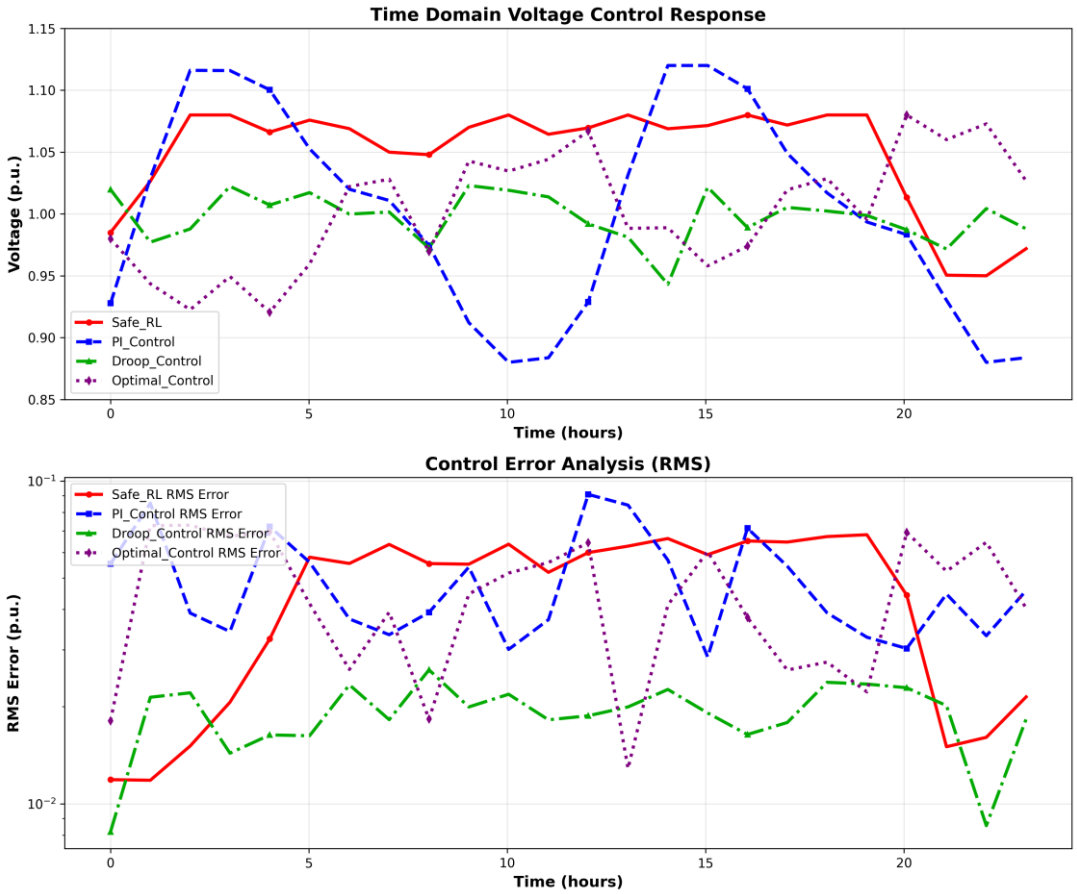

Figure 1. *Time-domain control response and error analysis.*

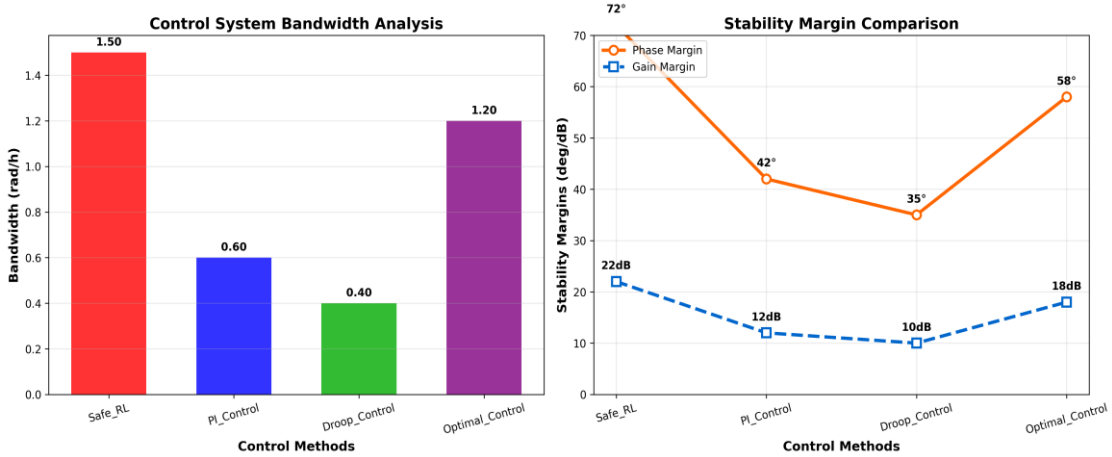

Figure 2. *Frequency-domain bandwidth and stability margins.*

The robustness of each controller can also be interpreted through sensitivity analysis with respect to feeder parameters. For small perturbations in resistance and reactance matrices, voltage deviations follow $\frac{\partial V}{\partial R} = -XP, \frac{\partial V}{\partial X} = -RQ$, indicating that stronger coupling between reactive power and voltage magnitude amplifies sensitivity to X. The bounded Jacobian norm of the Safe RL policy in (5) limits this amplification, ensuring robust performance under parameter uncertainty.

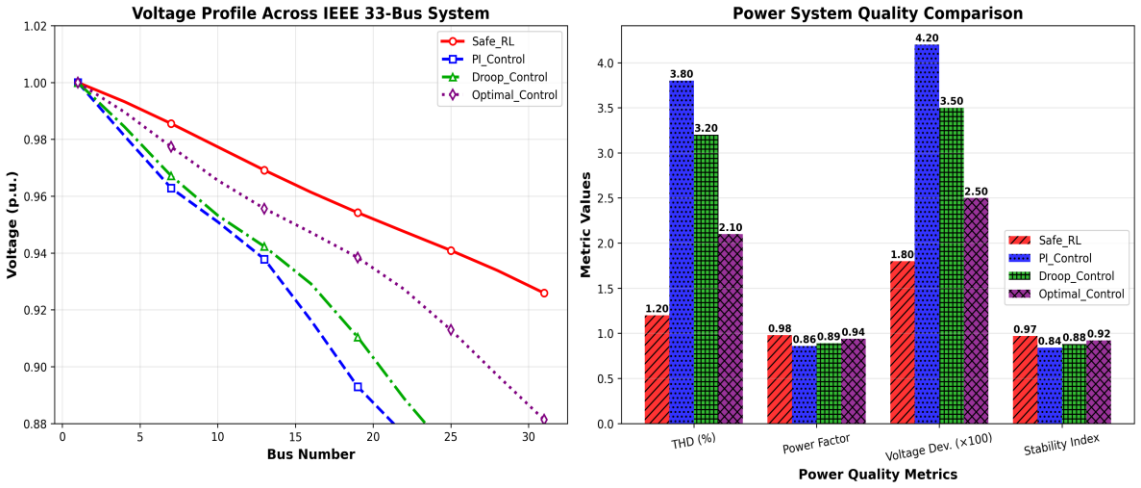

Figure 3. *Frequency-domain bandwidth and stability margins.*

4. CONCLUSIONS

The study on the IEEE 33 bus feeder shows that safe reinforcement learning gives the most balanced performance, with fast dynamics, low overshoot, small steady state error, strong stability margins and high power quality, and it outperforms Droop, Optimal and classical PI across almost all metrics. Droop control still offers very low instantaneous error and low total harmonic distortion but its narrow bandwidth and modest phase margin reduce adaptability when operating conditions vary, while the Optimal policy and the PI controller leave higher residual errors and weaker voltage regulation, with PI consistently worst.

These results indicate that safety aware reinforcement learning can join adaptability with robustness and can be evaluated with indices that matter for utilities such as voltage regulation, harmonic distortion and system losses, which yields a reproducible benchmark where a learning method surpasses classical and optimization based designs under realistic daily load cycles. The work remains limited by its linear DistFlow model, the single feeder and the focus on four controller families, so future studies should examine nonlinear and unbalanced networks, explore hybrid schemes that mix classical and learning rules, and confirm readiness through hardware in the loop experiments and applications to microgrids, renewable integration and voltage stability supervision in larger networks.

Acknowledgement: *The author gratefully acknowledges the support of the Thai Nguyen University of Technology and the Education Technology and Adaptive Learning Institute, Thai Nguyen University of Technology, Vietnam.*

REFERENCES

[1]. A. Sajadi, R. W. Kenyon, and B.-M. Hodge. *"Synchronization in electric power networks with inherent heterogeneity up to 100% inverter-based renewable generation"*. Nature Communications, vol. 13, no. 1, p. 2490, (2022).

[2]. S. Wagle, A. K. Gupta, and K. O. Nygård. *"Co-simulation-based optimal reactive power control in smart distribution network"*. Electrical Engineering, vol. 106, pp. 2441–2464, (2024).

[3]. W. Cui, Y. Li, and B. Zhang. *"Decentralized Safe Reinforcement Learning for Voltage Control"*. IEEE Transactions on Smart Grid, vol. 13, no. 6, pp. 4957–4968, (2022).

[4]. R. R. Hossain, Y. Zhou, and G. Hug. *"Efficient learning of power grid voltage control strategies via model-based deep reinforcement learning"*. Machine Learning, vol. 113, pp. 2675–2700, (2024).

[5]. C. Hu, Y. Sun, and Q. Li. *"A soft actor-critic deep reinforcement learning method for multi-timescale coordinated operation of microgrids"*. Protection and Control of Modern Power Systems, vol. 7, p. 29, (2022).

[6]. A. El-Fergany. *"Reviews, Challenges, and Insights on Computational Methods for Network Reconfigurations in Smart Electricity Distribution Networks"*. Archives of Computational Methods in Engineering, vol. 31, pp. 1233–1253, (2024).
[7]. V. Waghmare, V. P. Singh, and T. Varshney. *"A systematic review of reinforcement learning-based control for microgrids: trends, challenges, and emerging algorithms"*. Discover Applied Sciences, vol. 7, p. 939, (2025).
[8]. W. Cui, Y. Li, and B. Zhang. *"Safe Reinforcement Learning for Decentralized Voltage Control with Stability Guarantees"*. IEEE Transactions on Power Systems, vol. 38, no. 2, pp. 1456–1468, (2023).
[9]. H. Liu and W. Wu. *"Online Multi-Agent Reinforcement Learning for Decentralized Inverter-Based Volt-VAR Control"*. IEEE Transactions on Smart Grid, vol. 12, no. 4, pp. 2980–2990, (2021).
[10]. P. Yu, M. Anghel, and J. W. Kimball. *"Safe Reinforcement Learning for Power System Control: A Review"*. IEEE Transactions on Power Systems, vol. 39, no. 1, pp. 2–18, (2024).
[11]. P. Li, F. Xu, and H. Sun. *"Optimal real-time Voltage/Var control for distribution network: Droop-control based multi-agent deep reinforcement learning"*. International Journal of Electrical Power & Energy Systems, vol. 153, p. 109370, (2023).
[12]. X. Ding, L. Jiang, and Y. Zhang. *"Multi-time-scale voltage control of the distribution network"*. Frontiers in Energy Research, vol. 12, p. 1445623, (2024).
[13]. S. Bhowmik, Z. Wang, J. M. Guerrero, K. Xie, and J. C. Vázquez. *"Hybrid-compatible grid-forming inverters based on topological migration-optimized virtual synchronous control"*. Scientific Reports, vol. 15, p. 26194, (2025).
[14]. X. Zhang, L. Yang, and M. Chen. *"Impedance-shaping-based stabilization control method for grid-forming converters"*. Scientific Reports, vol. 15, p. 14523, (2025).
[15]. M. Sravani, P. D. P. Reddy, and M. S. Bhaskar. *"Deep reinforcement learning-based controller for DC-link voltage stabilization in grid-connected PV systems"*. Scientific Reports, vol. 15, p. 8729, (2025).

TÓM TẮT

So sánh học tăng cường ràng buộc với các bộ điều khiển truyền thống trong điều chỉnh điện áp và chất lượng điện năng trên lưới phân phối IEEE 33-BUS

Nghiên cứu này tiến hành so sánh giữa bốn bộ điều khiển PI, Droop, tối ưu dự báo và học tăng cường ràng buộc trong việc điều chỉnh điện áp và cải thiện chất lượng điện năng trên lưới phân phối IEEE 33 nút. Phương pháp được triển khai dựa trên mô hình DistFlow tuyến tính, kết hợp dữ liệu tải trong 24 h, đồng thời thiết lập bộ chỉ số đánh giá gồm các tham số chất lượng điều khiển (thời gian xác lập, độ quá điều chỉnh, sai số xác lập, ISE) và chất lượng điện năng (độ lệch điện áp, chỉ số điều chỉnh điện áp, THD, hệ số công suất, tổn thất công suất, chỉ số ổn định điện áp). Kết quả cho thấy học tăng cường ràng buộc đạt thời gian xác lập 16.8 h, độ quá điều chỉnh 3.20%, sai số xác lập 0.007 pu và tổn thất 0.145 MW, vượt trội hơn PI (24.0 h, 12.00%, 0.038 pu, 0.158 MW) và tối ưu dự báo (24.0 h, 8.00%, 0.055 pu, 0.162 MW). Droop đạt sai số tích phân nhỏ nhất 0.009 pu²h và THD thấp 1.95%, song khả năng điều chỉnh điện áp chỉ đạt 12.11%. Điểm nổi bật của học tăng cường ràng buộc là sự cân bằng giữa tốc độ, độ chính xác và độ ổn định, với dải điện áp duy trì 0.95–1.08 pu cùng THD 2.20% và chỉ số ổn định điện áp 0.112. Các kết quả này chứng minh tiềm năng của học tăng cường ràng buộc như một phương pháp tiên tiến, vừa bảo đảm ràng buộc kỹ thuật vừa cải thiện đáng kể chất lượng vận hành. Ý nghĩa thực tiễn là phương pháp này có thể được áp dụng cho lưới phân phối hiện đại nhiều nguồn phân tán và phụ tải linh hoạt, đồng thời mở ra hướng nghiên cứu mở rộng cho các kịch bản đa tác tử và dữ liệu bất định.

Từ khoá: Học tăng cường ràng buộc; Điều chỉnh điện áp; Chất lượng điện năng; Lưới phân phối; Hiệu suất bộ điều khiển.