Energy and AI 24 (2026) 100768

Contents lists available at ScienceDirect

Energy and AI

journal homepage: www.elsevier.com/locate/egyai

Physics-informed distributed reinforcement learning for privacy-aware voltage regulation using local smart meter data

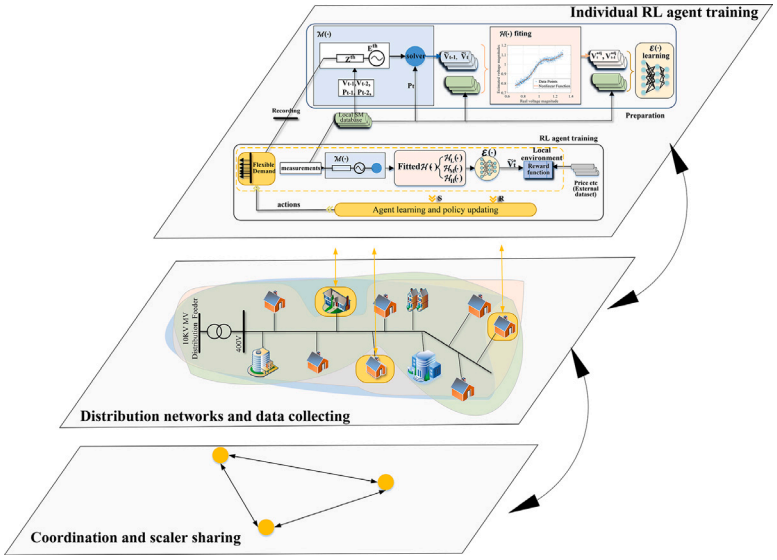
Dong Liu [a], Juan S. Giraldo [b], Peter Palensky [a], Pedro P. Vergara [a,*]

[a] Intelligent Electrical Power Grids Group, Delft University of Technology, 2628CD, The Netherlands
[b] Techno-Economic Energy Transition Studies, Netherlands Organisation for Applied Scientific Research, 2595 DA, The Netherlands

HIGHLIGHTS

- A physics-informed distributed learning method is proposed for voltage regulation.
- Local Thevenin models support privacy-aware voltage estimation at smart meters.
- A hybrid correction scheme improves voltage magnitude estimation accuracy.
- A coordination layer mitigates violations caused by excessive actions.
- Smart meter with intelligent decision-making ability support to power grid management.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Voltage regulation
Reinforcement learning
Distributed optimization
Privacy-aware

ABSTRACT

Centralized reinforcement learning-based voltage regulation in distribution networks is becoming increasingly difficult due to the growing penetration of distributed energy resources, high computational burden, repeated power flow calculations, and increasing privacy concerns. This paper proposes a physics-informed fully distributed reinforcement learning framework that enables autonomous voltage regulation using only local smart meter data. A Thevenin-equivalent-based local voltage estimation model and a hybrid correction mechanism are developed to support accurate local decision-making without synchronized global measurements or centralized power flow solvers. A lightweight coordination mechanism is further introduced to refine the actions of independently trained local agents. Case studies show that the proposed framework reduces voltage violations by approximately 80%, achieves performance close to that of power flow-based training environments, and achieves a training speedup of about 6×. The results also indicate that the relaxation

* Corresponding author.
E-mail address: P.P.VergaraBarrios@tudelft.nl (P.P. Vergara).

https://doi.org/10.1016/j.egyai.2026.100768
Received 30 March 2026; Received in revised form 26 April 2026; Accepted 2 May 2026
Available online 6 May 2026

factors in the reward function and the coordination scaler are critical to voltage regulation efficiency, whereas the discount factor has a smaller impact. These findings demonstrate the practicality of the proposed framework for privacy-aware fully distributed voltage regulation.

1. Introduction

1.1. Motivation

With the widespread adoption of smart electrical appliances and the increasing penetration of distributed energy resources (DERs), the operating conditions of distribution networks (DNs) have become more uncertain and time-varying, which poses significant challenges to voltage regulation [1]. Particularly, the stochasticity and variability of flexible household assets can aggravate voltage magnitude violations, especially in low-voltage distribution networks (LVDNs) with limited controllable resources [2]. Compared with medium-voltage distribution networks, where a larger proportion of aggregated demand and grid side observability can support centralized or optimization based voltage regulation, LVDNs are characterized by small scale, dispersed loads, limited dispatchable flexibility, and reduced system visibility at the household level. Under such conditions, distribution system operators (DSOs) increasingly rely on the active participation of end-users and household flexible assets to mitigate voltage violations during peak demand or highly congested periods [3].

However, enabling effective voltage regulation at the smart meter (SM) level remains challenging. Existing centralized and distributed approaches often rely on feeder topology information, repeated power flow (PF) calculations, or synchronized multi-agent interaction during training and control [4]. These requirements increase computational burden, communication dependence, and privacy exposure, and may reduce practical applicability in LVDNs where only local measurements are directly available. In addition, accurately estimating local voltage magnitudes using only SM level measurements remains difficult, particularly under topology missing and noisy data. These challenges motivate the development of a privacy-aware distributed learning framework that can support local voltage regulation using only local SM data.

1.2. Literature review

Traditional voltage-regulation approaches in distribution networks mainly rely on coordinated control of multiple resources, such as capacitor banks, photovoltaic (PV) systems, energy storage systems (ESSs), and electric vehicles (EVs), through centralized [5] or distributed optimization frameworks [6]. These methods have shown effectiveness in mitigating voltage violations by coordinately dispatching fast and slow control assets. However, they usually require feeder topology information, surrogate network models, or synchronized measurement collection, and often depend on repeated PF and optimization calculations during operation. As a result, their computational burden grows with network scale, while communication requirements and centralized data collection may raise privacy concerns, especially at the household level [7]. These challenges are further amplified in LVDNs, where individual loads and flexible resources are more uncertain and difficult to predict. In this context, reinforcement learning (RL) offers a promising alternative, since a trained agent can make decisions directly from observed states without solving PF equations online at each time step, thereby reducing online computational burden and mitigating the challenges of communication delays.

Unlike traditional data-driven approaches, RL does not rely on labelled or forecasting datasets, but instead uses a reward mechanism to guide the agent towards maximizing cumulative rewards over time [8]. Owing to this property, RL has shown strong potential for solving complex long-term sequential decision-making problems in power systems. For example, based on flexible network modelling, data augmentation, and tensor-based power flow computation, the deep reinforcement learning framework in [9] improves both training performance and computational efficiency for the optimal dispatch of ESSs in active distribution networks. A two-layer RL-based approach is introduced in [10] to coordinate slow-response discrete resources and fast-response continuous resources, effectively preventing voltage violations in DNs. Recent studies have further extended RL-based voltage regulation from the perspectives of interpretability, safety, and practical deployment. Some studies incorporate network topology or graph information into RL for improved coordination and resilience. For voltage control in PV-rich distribution grids, a graph convolutional network-based multi-agent meta-reinforcement learning framework is introduced in [11] by incorporating network topology information to improve voltage quality, mitigate environmental instability, and address uneven learning among agents. A graph-based deep RL framework is developed in [12] for voltage regulation in DNs, improving resilience to cyber contingencies by integrating cyber–physical graph information into the control policy under partial observability. In [13], an explainable RL-based approach is proposed to improve the practical applicability of RL in voltage regulation, where a trained multi-agent system coordinates multiple DERs, including ESSs and hydrogen storage systems, for voltage magnitude regulation. Since the reliability of artificial neural network (ANN)-based RL actions is not yet fully verified, human expertise is incorporated into the actor–critic algorithm in [14] to optimize suspicious actions and prevent extra voltage violations caused by PV power. Similarly, a physics-shielded RL algorithm is proposed in [15] to coordinate DERs while ensuring equipment safety. A safe multi-critic soft actor–critic framework is introduced in [16], enabling the separate learning of multiple operational objectives while enforcing real-time compliance with network constraints, including voltage magnitude limits.

In parallel, privacy-aware and communication efficient RL frameworks have also attracted attention. A spatial–temporal transformer based federated reinforcement learning framework is proposed in [17] for online voltage regulation in active distribution systems, enabling privacy-aware and uncertainty-aware control through decentralized learning and physics-informed feature aggregation. To mitigate both medium-voltage violations and low-voltage three-phase voltage unbalance in PV-rich dual-level DNs, the federated RL method in [18] is developed based on hierarchical multi-agent learning. RL has also been extended to more complex operating scenarios involving demand response, cyber attacks, and high EV penetration. In [19], a multi-agent RL-based voltage constrained incentive demand response framework is proposed to coordinate PV integrated household demand for effective voltage regulation and demand reduction in LVDNs. In [20], an attention based multi-agent soft actor–critic algorithm is developed to achieve robust and adaptive voltage regulation under false data injection attacks. For distribution systems with high EV penetration, a three-level voltage-regulation framework is proposed in [21], in which an unscented Kalman filter-based deep RL approach is integrated with a multi-robot adaptive charging network to enable real-time and flexible mitigation of voltage violations.

In addition to improving voltage regulation performance, existing studies have also explored RL-based approaches from the perspectives of coordination and scenario adaptation. To improve the coordinated control of heterogeneous DERs, the voltage-reactive power sensitivity is introduced in [22] to guide the convergence direction of the multi-agent training process and reduce the frequency of voltage magnitude violations. Beyond coordination design, several studies have sought to improve the adaptability and data efficiency of RL-based energy

management frameworks. Considering the similarity of appliance characteristics across different houses, transfer learning is incorporated in [23] to enable the transfer of a basic agent from one home to another without complete model retraining, while also mitigating privacy leakage. To support RL training under distributed datasets,federated learning is adopted in [24] to enable agent training without compromising household privacy. In [25], clustering algorithms are used to group training datasets and assign them to different agents, so that during online application, a given scenario can be handled by the agent trained on the same or a similar category. In addition, to address the uncertainty in load profiles caused by human activities, a recognition model is integrated into the RL-based household energy management system in [26] to facilitate more effective appliance scheduling. Under a model-free RL setting, the deep reinforcement learning-based scheduling framework in [27] is developed for ESSs in unbalanced LVDNs and applied to real-time voltage regulation, achieving near optimal day ahead voltage regulation under uncertainty.

Despite the growing potential of RL for voltage regulation, several critical barriers still hinder its practical deployment in active distribution networks. First, most existing centralized RL or multi-agent RL frameworks rely on synchronized data aggregation and repeated PF calculations, which impose considerable computational and communication burdens and may increase privacy exposure by requiring access to sensitive consumer-level measurements [28]. Although hierarchical and coordinated control architectures provide a structured way to manage voltage violations, they usually depend on global information exchange and accurate network models, making them vulnerable to communication bottlenecks, single point failure risks, and rapidly changing operating conditions caused by household DERs [29].

Second, while decentralized and data-driven alternatives have emerged, many of them still depend on purely data-driven PF surrogates or complete ANN-based operators to approximate grid physics, such as local voltage magnitude estimation. These models may suffer from estimation bias under noisy or missing measurements and often lack physical consistency [30]. A related example is given in [31], where decentralized RL is evaluated for residential ESS management using a data-driven simulation environment. Although the agents are trained and operated locally without direct information sharing, the main objective is cost reduction and peak mitigation rather than voltage regulation. As a result, that framework does not need to address feeder topology dependence, local voltage magnitude estimation, or the strong coupling of agent observations through network voltage interactions, which remain central challenges in distributed voltage regulation problems.

Third, even in distributed settings, many RL agents are still trained using complete or centrally constructed datasets and are only deployed locally afterwards. This limits true local autonomy and may lead to poorly coordinated behaviours among independently trained agents. In voltage magnitude regulation, such insufficient coordination may trigger simultaneous excessive charging or discharging actions, thereby aggravating voltage violations rather than mitigating them. Consequently, achieving a framework that simultaneously supports local data-based autonomy, coordination among distributed agents, and physically consistent voltage control remains an open challenge.

Recent studies in maintenance and reliability have highlighted the potential of multi-agent approaches for decentralized decision-making under partial observability and imperfect information. These include applications to hybrid maintenance optimization [32], optimal preventive maintenance policy design [33], and condition-based maintenance optimization based on multi-agent reinforcement learning [34]. However, these studies mainly focus on long-term maintenance scheduling and system health management, rather than real-time operational control in power systems. Therefore, extending such decentralized and uncertainty-aware learning ideas to privacy-aware, physics-informed voltage magnitude regulation in DNs remains a meaningful and largely unexplored research direction.

1.3. Contributions

The core idea of the proposed framework is to enable fully SM side RL training by constructing a locally usable voltage magnitude estimation environment through physics-based modelling and data-driven correction. To this end, this paper proposes a physics-informed distributed reinforcement learning framework with local voltage magnitude estimation, hybrid correction, and lightweight coordination for voltage magnitude regulation using only local SM data. To more clearly position the proposed framework relative to representative existing methods, Table 1 summarizes the key differences among these methods. The main contributions of this work are summarized as follows:

- A localized RL training framework is developed by integrating a dynamic Thevenin equivalent model into the construction of the local environment. This provides each SM with an initial local voltage magnitude estimation capability based only on its own measurements, enabling RL training to be shifted from the DSO side to the SM side without requiring repeated centralized PF calculations, synchronized global data polling, or topology-based online computation.
- A hybrid voltage magnitude correction mechanism is further proposed to improve the accuracy of the locally constructed voltage estimation environment. Specifically, piecewise correction functions are introduced to compensate for structured estimation errors caused by the simplified Thevenin model, while the ANN-based correction module is designed to capture the nonlinear sensitivity between agent actions and voltage responses. Together with the Thevenin-based local model, this correction strategy enables the RL training environment to be constructed entirely at the SM side, thereby avoiding dependence on the DSO or centralized PF-based computation during training.
- A lightweight coordination mechanism is introduced for independently trained agents to mitigate excessive actions during online deployment. By exchanging only limited coordination variables, mainly for determining the coordination scaler, neighbouring SMs can locally refine agent outputs without sharing raw measurements. Since this coordination is not designed as a communication intensive real-time control layer, it imposes only a limited communication burden and does not occupy continuous real-time communication channels. This design helps prevent excessive simultaneous actions, such as synchronized ESS charging/discharging, thereby improving the practicality of distributed voltage regulation under unknown topology.

2. Preliminaries

2.1. ESSs-based voltage magnitude regulation

Consider a DN with a set of nodes $\mathcal{N}$, where a subset of nodes $\mathcal{B} \subseteq \mathcal{N}$ are equipped with ESSs. Voltage magnitude regulation approach via ESSs can be performed by dynamically dispatching their power outputs, which can be formulated as a nonlinear optimization problem. The typical objective function of this problem is to use the minimal active power to avoid voltage magnitude violation during the period $\mathcal{T}$ [35], which is expressed as (1). The constraints are given by (2)–(8). The main notations used in this paper are summarized in Table 2.

$$\min_{P^b_{m,t}} \quad \sum_{t\in\mathcal{T}} \sum_{m\in\mathcal{B}} C^b_t |P^b_{m,t}| \tag{1}$$

s.t.

$$\sum_{km\in\mathcal{L}} P_{km,t} + P^b_{m,t} - \sum_{mn\in\mathcal{L}} P_{mn,t} - \sum_{mn\in\mathcal{L}} \frac{P^2_{mn,t} + Q^2_{mn,t}}{V^2_{m,t}} \mathrm{R_{mn}} = \mathrm{P^D_{m,t}} \quad \forall m \in \mathcal{N}, \forall t \in \mathcal{T} \tag{2}$$

Table 1
Comparison of the proposed framework with representative RL-based approaches in terms of control architecture, communication dependence, training interaction, topology requirement, local voltage magnitude estimation capability, coordination design, deployment location, and extensibility to other controllable devices with Markovian dynamics.

Ref.	Cent.	Dist.	Topology	Central comm.	Agent comm.	Train. inter.	Local V est.	Simple coord.	Sync.	Explainability	Deployment	Other Markovian devices
[7]	✗	✓	✓	✓	✗	✗	✗	✓	✓	✗	Aggregator/station	✗
[9]	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗	Distributed aggregator	✗
[11]	✓	✗	✓	✓	✗	✗	✗	✓	✓	✗	Distributed aggregator	Limited
[12]	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗	Stations & DSO	✓
[13]	✓	✗	✓	✓	✗	✗	✗	✓	✓	Limited	Aggregator & DSO	✗
[16]	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗	Aggregator & DSO	✗
[17]	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗	Aggregator	✗
[18]	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗	Aggregator	✗
[20]	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗	Aggregator & DSO	✗
[21]	✓	✗	✓	✓	✗	✗	✗	✓	✓	✗	Aggregator & DSO	✗
[22]	✓	✗	✓	✓	✗	✗	✗	✓	✓	✗	Distributed aggregator	✗
[27]	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗	DSO	✗
[30]	✗	✓	✗	✗	✗	✓	✗	✗	✓	✗	SM	✓
This work	✗	✓	✗	Limited	Limited	✓	✓	✓	✓	Limited	SM	✓

Table 2
Main notations used in the optimization, Thevenin-equivalent modelling, and reinforcement learning framework for ESS-based voltage magnitude regulation in the DN.

Symbol	Description	Symbol	Description
$\mathcal{N}$	Set of node index	$\mathcal{B}$	Subset of nodes equipped with ESSs
$\mathcal{L}$	Set of connection lines	$\mathcal{T}$	Scheduling/control time horizon
m, n	Indices of nodes in $\mathcal{N}$	t	Time step index
mn, km	Indices of lines in $\mathcal{L}$.	R_{mn}, X_{mn}	Resistance and reactance of line mn
$P^b_{m,t}/Q^b_{m,t}$	Active/reactive power output of the ESS at node m and time t.	C^b_t	Compensation price paid to ESS owners at time t
$V_{m,t}$	Voltage magnitude at node m and time t	$\underline{V}$, $\overline{V}$	Lower and upper voltage magnitude limits
$P_{km,t}/P_{mn,t}/\ Q_{km,t}/Q_{mn,t}$	Active/reactive power flowing into/out of node m at time t.	$P^D_{m,t}$, $Q^D_{m,t}$	Net active and reactive power demand at node m and time t
$SOC_{m,t}$	State of charge of ESSs at node m and time t.	$\underline{SOC}$, $\overline{SOC}$	Lower and upper SOC limits of the ESS
$\underline{P}^b/\overline{P}^b$	Lower and upper active power limits of ESSs	E^{rated}_m	Rated energy capacity of the ESS at node m
Δt	Time interval between two consecutive steps	$\dot{E}^{th}_m$	Thevenin equivalent voltage at node m
$\dot{Z}^{th}_m$	Thevenin equivalent impedance at node m	r^{th}_m, x^{th}_m	Resistive and reactive parts of $\dot{Z}^{th}_m$
$\dot{V}_{m,t}$	Voltage at node m and time t	$\dot{I}_{m,t}$	Current from the network to node m at time t
P_m, Q_m	Active and reactive power at node m	y_m	Auxiliary variable defined as $y_m = V^2_m$
$\hat{V}_{m,t}$	Estimated voltage magnitude derived from the Thevenin equivalent model	a, b, c	Coefficients of the quadratic equation with respect to V^2_m
α	Learning rate in RL agent training	$\cos\theta$	Power factor
S_t	State of the RL environment at time t	π	Policy of the RL agent
a_t	Action selected by the RL agent at time t.	R_t	Reward obtained by the RL agent at time t
G_t	Target return/cumulative reward from time t	γ	Discount factor in RL.
V^{adj}_t	Adjusted estimated voltage magnitude	$\tilde{V}_t$	Final estimated voltage magnitude from V^{adj}_t
β	Coordination scaling factor for household ESS	$\tilde{a}_m$	Actual action applied to regulate the charging/discharging of the ESS at node m
w_λ	Weighting coefficient for the coordination penalty term	w_p	Weighting coefficient for ESS active power usage in the reward function
$\overline{\varsigma}$, $\underline{\varsigma}$	Relaxing factors for upper and lower voltage magnitude to trigger the penalty region	λ	Coordination penalty/coordination scalar used in the reward design

$$\sum_{km\in\mathcal{L}} Q_{km,t} + Q^b_{m,t} - \sum_{mn\in\mathcal{L}} Q_{mn,t} - \sum_{mn\in\mathcal{L}} \frac{P^2_{mn,t}+Q^2_{mn,t}}{V^2_{m,t}}\mathrm{X_{mn}} = \mathrm{Q^D_{m,t}} \quad \forall m\in\mathcal{N}, \forall t\in\mathcal{T} \tag{3}$$

$$V^2_{m,t} - V^2_{n,t} = 2(\mathrm{R_{mn}}P_{mn,t} + \mathrm{X_{mn}}Q_{mn,t}) - \frac{P^2_{mn,t}+Q^2_{mn,t}}{V^2_{m,t}}(\mathrm{R^2_{mn}} + \mathrm{X^2_{mn}}) \quad \forall m,n\in\mathcal{N}, \forall mn\in\mathcal{L}, \forall t\in\mathcal{T} \tag{4}$$

$$\underline{\mathrm{V}} \le V_{m,t} \le \overline{\mathrm{V}} \quad \forall m\in\mathcal{N}, \forall t\in\mathcal{T} \tag{5}$$

$$\underline{\mathrm{P}}^b \le P^b_{m,t} \le \overline{\mathrm{P}}^b, \quad \forall m\in\mathcal{B}, \forall t\in\mathcal{T} \tag{6}$$

$$\underline{\mathrm{SOC}} \le SOC_{m,t} \le \overline{\mathrm{SOC}}, \quad \forall m\in\mathcal{B}, \forall t\in\mathcal{T} \tag{7}$$

$$SOC_{m,t+1} = SOC_{m,t} + \frac{P^b_m \Delta t}{\mathrm{E}^{rated}_m}, \quad \forall m\in\mathcal{B}, \forall t\in\mathcal{T} \tag{8}$$

where C^b_t denotes the compensation price paid to the ESS owners for participating in voltage magnitude regulation. Indices m and n represent nodes in $\mathcal{N}$, while mn and km denote lines in $\mathcal{L}$.

$V_{m,t}$ is the voltage magnitude at node m at time t, which is bounded by the lower limit $\underline{\mathrm{V}}$ and upper limit $\overline{\mathrm{V}}$. The power balance is modelled by constraints (2) and (3). The fourth term in constraints (2) and (3) represents the active power loss and reactive power loss in the line connecting nodes n and m. Constraint (4) formulates the voltage magnitude drop across the lines. Parameters $\mathrm{R_{mn}}$ and $\mathrm{X_{mn}}$ are the resistance and reactance of the line mn. Variables $P_{km,t}$ and $P_{mn,t}$ represent the active power flowing into and out of node m at time t, respectively. $Q_{km,t}$ and $Q_{mn,t}$ represent the reactive power flowing into and out of node m at time t, respectively.

$P^b_{m,t}$ and $Q^b_{m,t}$ denote the active and reactive power outputs of the ESS at node m and time t, respectively. The two variables are assumed to be interrelated via the power factor $\cos\theta$. A positive value indicates charging behaviour, whereas a negative value represents discharging

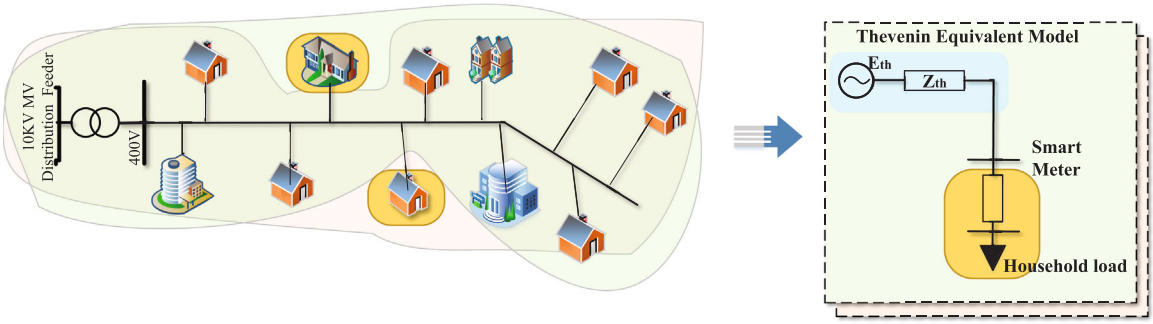


Fig. 1. Illustration of the distribution network seen from the SM side and its corresponding Thevenin-equivalent representation in an LVDN with two ESSs.

behaviour. When an ESS is not installed at m, the value is set to 0. The active $P^b_{m,t}$ of the batteries are bounded by their power limits $\overline{P}^b, \underline{P}^b$, reads as (6). $P^D_{m,t}$ and $Q^D_{m,t}$ represent the net demand at node m at time t. The State of Charge (SOC) of the ESS must remain within its allowable range and is updated based on the ESS power output, which is expressed as expression (7) and (8). E^{rated}_m denotes the capacity of the ESS at node m. In this study, the ESS charging/discharging efficiency is idealized as 100% to avoid introducing additional model complexity and to focus on the proposed distributed learning and coordination framework.

2.2. Thevenin theorem and its dynamic representation

According to the Thevenin theorem, a linear circuit comprising multiple voltage sources and impedances can be reduced to an equivalent circuit consisting of a single voltage source Eth in series with an equivalent impedance $Z^{th} = r^{th} + jx^{th}$. This simplified representation preserves the electrical characteristics of the original network at the terminals of interest, thereby facilitating the analysis of system behaviour at specific nodes [36]. The electrical perspective observed from each SM differs, as illustrated in Fig. 1.

In particular, the nodal voltage at bus m at time steps t and $t+1$ can be described using a Thevenin equivalent formulation, which expresses the voltage as a function of the equivalent impedance and the injected current [37], as shown in expressions (9) and (10). Consequently, determining the Thevenin equivalent parameters requires two consecutive samples corresponding to distinct operating points.

$$\dot{E}^{th}_{m,t+1} = \dot{Z}^{th}_{m,t+1}\dot{I}_{m,t+1} + \dot{V}_{m,t+1} \tag{9}$$

$$\dot{E}^{th}_{m,t} = \dot{Z}^{th}_{m,t}\dot{I}_{m,t} + \dot{V}_{m,t}, \tag{10}$$

where $\dot{I}_{m,t}$ is the current from the network to node m at time t. When the two sampling times are very close, $\dot{E}^{th}$ and $\dot{Z}^{th}$ are assumed to be constant during this period. The approximate values of the Thevenin equivalent parameters can be inferred as follows:

$$\dot{Z}^{th}_m = \frac{\dot{V}_{m,t} - \dot{V}_{m,t+1}}{\dot{I}_{m,t+1} - \dot{I}_{m,t}} \tag{11}$$

$$\dot{E}^{th}_m = \frac{\dot{V}_{m,t}\dot{I}_{m,t+1} - \dot{V}_{m,t+1}\dot{I}_{m,t}}{\dot{I}_{m,t+1} - \dot{I}_{m,t}}. \tag{12}$$

Therefore, by online monitoring voltage magnitudes and current values, the overall dynamics of the entire network can be observed through a node. This feature provides a reference for local control on the SM side. Assuming that the two parameters $\dot{E}^{th}$ and $\dot{Z}^{th}$ remain unchanged between two samples,[1] it is possible to calculate whether the load at the next moment will cause the voltage magnitude to exceed the limit based on these two values, which provides the possibility of pre-management at the same time.

Specifically, based on the Thevenin equivalent voltage $\dot{E}^{th}$, impedance $\dot{Z}^{th}$, and load power P_m at the next time step, the voltage magnitude V_m can be calculated based on expression (4) [37]. Assuming $cos\theta$ is the power factor of node m and reactive power Q_m in (13) is $P_m tan\theta$. The expression in (4) can be re-written as in (13).

$$V_m^4 + V_m^2\left[-2\left(r^{\text{th}}_m P_m + x^{\text{th}}_m Q_m\right) - \left(E^{\text{th}}_m\right)^2\right] + \left(P_m^2 + Q_m^2\right)\left[\left(r^{\text{th}}_m\right)^2 + \left(x^{\text{th}}_m\right)^2\right] = 0. \tag{13}$$

Eq. (13) can be viewed as a quadratic equation with respect to V_m^2. By defining $y_m = V_m^2$, (13) is rewritten as $ay_m^2 + by_m + c = 0$. Applying the quadratic formula yields $V_m^2 = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$. Since the voltage magnitude must be real and non-negative, the physically meaningful solution is selected as $V_m = \sqrt{\frac{-b+\sqrt{b^2-4ac}}{2a}}$. Thus, the local PF calculation enables local users to estimate the voltage magnitude without relying on synchronized data collection or the centralized PF model.

Unlike purely data-driven black-box models, such as standard ANN-based surrogates for voltage-magnitude estimation, the Thevenin equivalent model provides a physics-grounded mechanism for relating local measurements to voltage variations. This makes the estimation process more interpretable and physically consistent, and helps maintain reasonable voltage estimates under changing operating conditions. However, because the simplified Thevenin representation cannot fully capture network coupling and measurement imperfections, its raw estimates may still exhibit noticeable errors. This limitation motivates the subsequent correction stage introduced in the proposed framework.

2.3. Centralized RL algorithm

The dispatch of ESSs can be formulated as a Markov decision process, in which the ESS state evolves according to its previous state of charge and charging/discharging action. Specifically, the SOC_{t+1} at time $t+1$ depends on the previous SOC_t and the action a_t, as described in (6). The general deployment process of a single-agent RL framework for centralized scheduling of multiple ESSs is summarized below [38].

Initialization

- *Environment construction*: the centralized environment represents the LVDN, including its operating constraints and dynamic behaviour, and is typically implemented using PF-based models.
- *State definition*: the state is defined using the nodal voltage magnitudes and the ESS state of charge, i.e.,

  $S_t = [SOC_t, V_{1,t}, V_{2,t}, \ldots, V_{N,t}]$.

- *Reward function*: DSOs aim to minimize the active power used for voltage regulation while penalizing voltage magnitude violations. Accordingly, the reward is formulated as

  $R_t = -C^b_t \sum_{m\in\mathcal{B}} \left|P^b_{m,t}\right|$

[1] $\dot{Z}_{\text{th}}$ is assumed constant impedance because no topological changes are expected between two samples, and no tap changes occur. $\dot{E}_{\text{th}}$ refers to the open-circuit voltage measured or calculated across the two specific terminals of the original circuit when the load is removed.

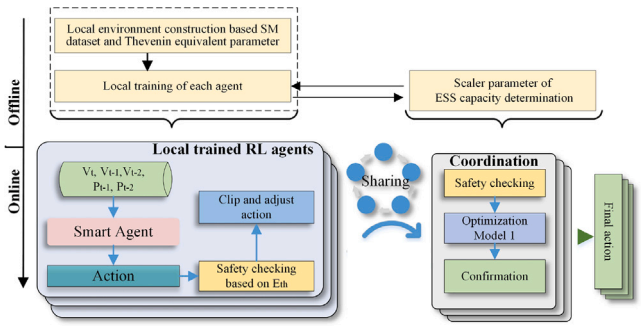


Fig. 2. Overall coordinated-distributed framework of the proposed physics-informed RL approach for voltage magnitude regulation using SM data.

$$+\sum_{m\in\mathcal{N}}\min\left\{0,\ \frac{\overline{V}-\underline{V}}{2}-\left|V_{m,t}-\overline{V}\right|\right\}. \tag{14}$$

The reward increases when less active power is used, and the voltage magnitudes remain within the allowable range.

- *Policy initialization*: the policy and value networks of the RL agent are initialized.

Agent observation and action selection

At time t, the agent observes the current state S_t and selects an action a_t according to the policy π, subject to the ESS operating constraints in (6)–(8). The action corresponds to the charging/discharging power of the ESS.

Environment interaction

After receiving the action, the environment updates the system state by solving the PF model subject to (2)–(5). The resulting reward R_t and the next state S_{t+1} are then returned to the agent:

$$S_{t+1}=f_{PF}(a_t,S_t,P_t^D,Z_R,Z_X,SOC_t) \tag{15}$$

where $f_{PF}(\cdot)$ denotes the state transition function based on PF calculation, and Z_R and Z_X are the line resistance and reactance parameters, respectively.

Policy update

The PPO algorithm is employed to update the policy π and the value function using the collected state–action–reward trajectories. The discounted return is defined as

$$G_t=\sum_{k=0}^{\infty}\gamma^k R_{t+k} \tag{16}$$

where $\gamma\in[0,1]$ denotes the discount factor, which determines the relative importance of immediate and future rewards. The interaction between the agent and the centralized environment is repeated until the stopping criterion is satisfied, after which the trained policy is used for ESS charging/discharging control in voltage regulation.

3. Physics-informed distributed RL framework

A distributed RL algorithm is proposed to enable RL agents deployed at households to perform voltage magnitude regulation in LVDNs without requiring centralized training or synchronized data collection for PF calculation. The overall framework of the proposed algorithm is illustrated in Fig. 2, which consists of independent training of local RL agents and online coordination. In this paper, household ESSs are treated as the only dispatchable resources. Nevertheless, the proposed algorithm is not limited to ESS dispatch and can also be extended to other controllable household appliances.

To facilitate the implementation of the proposed coordinated-distributed framework, the following assumptions are made in our work. First, SMs are assumed to be equipped with basic communication capabilities. Second, each SM is assumed to be aware of its neighbouring nodes, although the complete network topology is not available. Third, each SM is assumed to have sufficient local computational capability to process its own measurements and incorporate limited information received from neighbouring nodes for action refinement.

3.1. Privacy-aware distributed training environment

3.1.1. Voltage magnitude estimation via Thevenin equivalents

The voltage magnitude estimation process described in Section 2.2 is embedded within the local environment. In this environment, the values of E^{th} and Z^{th} are dynamically updated via (11) and (12) according to the SM data sampling frequency. Based on (12), the three inferred coefficients are as (18).

$$\begin{aligned} a &= 1, && \forall m\in\mathcal{B}\\ b &= -2(r_m^{\text{th}}P_m+x_m^{\text{th}}Q_m)-(E_m^{\text{th}})^2, && \forall m\in\mathcal{B}\\ c &= (P_m^2+Q_m^2)\left[(r_m^{\text{th}})^2+(x_m^{\text{th}})^2\right] && \forall m\in\mathcal{B}. \end{aligned} \tag{18}$$

Given the latest values of E^{th}, Z^{th}, a, b, and c, the voltage magnitude at bus m is then estimated by Eq. (17) in Box I, which is implemented in the local environment (i.e., SM)

However, the Thevenin equivalent parameters alone are not always sufficient to support accurate local voltage magnitude estimation. According to (11) and (12), the Thevenin equivalent parameters are estimated from two sampled operating points and then used for subsequent voltage estimation. Therefore, any estimation error in Eth and Zth will be directly transferred to the estimated voltage magnitude. In practice, such errors mainly arise from three sources. First, SM data are usually sampled at relatively low frequencies (e.g., every 5 or 15 min), which limits the capability of the identified Thevenin parameters to reflect fast changes in operating conditions. Second, measurement noise may further degrade the accuracy of the estimated voltage magnitude $\hat{V}_t$. Third, when the sampled current values $\dot{I}_{m,t}$ and $\dot{I}_{m,t+1}$ are very close, the estimated Thevenin parameters become numerically sensitive, causing the voltage magnitude computed from (11)–(13) to deviate from the expected operational range. These inaccuracies distort the local environment for independent RL training and may hinder the convergence of the RL agent to an effective voltage regulation policy.

On the one hand, the Thevenin equivalent model only represents the network around the current operating point and therefore cannot fully reproduce the post-action voltage evolution driven by the agent's decisions. Consequently, it is insufficient to provide an accurate action-to-voltage mapping for RL training. On the other hand, when the voltage estimated from the Thevenin parameters is inaccurate, the RL agent cannot determine whether the observed voltage variation is induced by its action, by estimation errors in the Thevenin parameters, or by measurement noise. Such ambiguity obscures the causal relationship between action and reward, which reduces training reliability and may slow down or even prevent convergence to an effective voltage regulation policy. This phenomenon is partially reflected in Fig. 5 and motivates the introduction of the subsequent hybrid correction strategy based on polynomial functions and neural networks.

3.1.2. Hybrid error–correction strategy

To reduce the impact of inaccurate Thevenin parameter estimation on local voltage reconstruction, a piecewise adjustment function $\mathcal{H}(\cdot)$ is introduced to correct the preliminary voltage estimate. The purpose of this correction is to constrain the estimated voltage magnitude to a physically reasonable range (e.g., 0.95 p.u. to 1.05 p.u., with voltage magnitude expressed in per unit based on a 400 V base voltage) and to suppress abnormal deviations caused by errors in Eth and Zth. In particular, the correction is designed to account for the fact that the estimation bias may differ across voltage regions, making a single global fitting function less effective. Therefore, a piecewise polynomial structure is adopted to provide greater flexibility and improved fitting

$$\hat{V}_{m,t} = \left(\frac{-b + \sqrt{b^2 - 4ac}}{2a} \right)^{\frac{1}{2}} = \left(\frac{2(r_m^{\text{th}} P_m + x_m^{\text{th}} Q_m) + (E_m^{\text{th}})^2 + \sqrt{\left[-2(r_m^{\text{th}} P_m + x_m^{\text{th}} Q_m) - (E_m^{\text{th}})^2\right]^2 - 4(P_m^2 + Q_m^2)\left[(r_m^{\text{th}})^2 + (x_m^{\text{th}})^2\right]}}{2} \right)^{\frac{1}{2}} \tag{17}$$

Box I.

accuracy under different operating conditions. The proposed piecewise adjustment function $\mathcal{H}(\cdot)$ is formulated as follows:

$$V_t^{\text{adj}} = \mathcal{H}(\hat{V}_{m,t}) = \begin{cases} \mathcal{H}_{\text{L}}(\hat{V}_{m,t} - V_{m,t}), & \hat{V}_{m,t} < \tilde{V}_{\min}, \\ \mathcal{H}_{\text{M}}(\hat{V}_{m,t} - V_{m,t}), & \tilde{V}_{\min} \le \hat{V}_{m,t} \le \tilde{V}_{\max}, \\ \mathcal{H}_{\text{H}}(\hat{V}_{m,t} - V_{m,t}), & \hat{V}_{m,t} > \tilde{V}_{\max}. \end{cases} \tag{19}$$

where $\mathcal{H}_{\text{L}}(\cdot)$, $\mathcal{H}_{\text{M}}(\cdot)$, and $\mathcal{H}_{\text{H}}(\cdot)$ denote high-order polynomial fitting functions defined over different voltage intervals, whose parameters are identified from historical local data. The thresholds $\tilde{V}_{\min}$ and $\tilde{V}_{\max}$ determine the boundaries of these intervals. In this way, the proposed piecewise structure provides a coarse but physics-consistent correction of the voltage estimate by compensating for systematic errors in the Thevenin-based reconstruction. The output V_t^{adj} represents the adjusted voltage magnitude at time t.

Although the Thevenin-based estimation, combined with the piecewise correction, can significantly reduce extreme deviations and improve the physical consistency of the reconstructed voltage magnitude, the resulting local model remains a simplified approximation of the actual network and cannot fully reproduce the true voltage sensitivity to agent actions. In particular, when voltage fluctuations are observed, it may be unclear whether they are caused by the control action itself or by inaccuracies in the Thevenin-based approximation. Such ambiguity can weaken the quality of the reward signal and may consequently slow down or destabilize the learning process.

To mitigate this issue, a nonlinear sensitivity function $\mathcal{E}(\cdot)$ is introduced into the voltage magnitude estimation process of the local environment. In this work, a Transformer-based correction network is employed to approximate the nonlinear relationship between agent actions and the corresponding network voltage response. The Transformer architecture consists of three main components:

- *Input embedding layer*: The input variables are first projected into a latent feature space, providing a suitable representation for subsequent Transformer-based learning.
- *Transformer encoder*: The encoder captures nonlinear dependencies and interactions among the input features, enabling the model to adaptively extract the information most relevant to voltage estimation.
- *Regression head*: The encoded features are passed through a final linear layer to produce a scalar output corresponding to the estimated voltage magnitude $\tilde{V}_t$.

Accordingly, the final voltage magnitude estimated from local SM data and the corrected voltage V_t^{adj} is formulated as

$$\tilde{V}_t = \mathcal{E}(a_t, P_t^D, V_{t-1}, V_{t-2}, P_{t-1}, P_{t-2}, E^{th}, Z^{th}, V_t^{adj}) \tag{20}$$

where the inputs to $\mathcal{E}(\cdot)$ include the control action a_t, local SM measurements, the Thevenin equivalent parameters Eth and Zth, and the corrected voltage estimate V_t^{adj}. These inputs are selected to preserve physical interpretability while improving the accuracy of voltage estimation in the locally constructed training environment. The overall voltage-estimation process based on local SM data is illustrated in Fig. 3(a).

3.1.3. Distributed environment construction

The SM in the household equipped with an ESS is considered an individual agent. The agent has access to local datasets, including active power and voltage magnitude measurements, while the power factor is assumed to be known. The distributed environment is constructed to estimate the voltage magnitude based on the latest SM data and the updated actions from agents, and return the corresponding state S_t and reward R_t. In the distributed environment, each agent has one local environment, where the local voltage magnitude estimation process in Section 3.1.1 is integrated for voltage magnitude estimation using local SM data. All agents are trained independently within their respective local environments. The general training process of a single RL agent in a distributed environment is illustrated in Fig. 3(b).

The action a_t represents the charging and discharging power of the ESS, modelled as variable P_t^b. The action space is constrained by the ESS charging/discharging limits, as formulated in expression (6). The state vector S_t at time step t is defined as:

$$S_t = [P_t^D, SOC_t, \tilde{V}_t, V_{t-1}, V_{t-2}, P_{t-1}^D, P_{t-2}^D] \tag{21}$$

where P_t^D represents the active power of the household at time t, SOC_t is the state of charge of the ESS at time t, and $\tilde{V}_t$ is the estimated voltage magnitudes. The inclusion of past voltage and power values $(V_{t-1}, V_{t-2}, P_{t-1}^D, P_{t-2}^D)$ allows the agent to capture temporal dependencies and are used to calculate Eth and Zth.

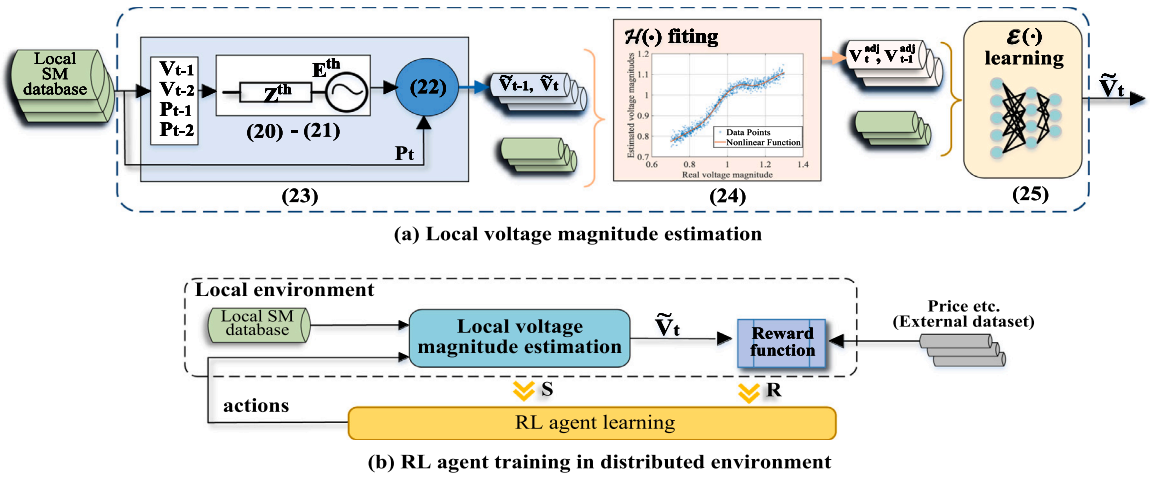
The reward function R_t is defined using a step function to penalize voltage magnitude violations beyond acceptable limits. It is updated as follows:

$$R_t = \begin{cases} -w_p C_t^b \sum |P^b| - w_\lambda \lambda, & \tilde{V}_t > \overline{V} \text{ or } \tilde{V}_t < \underline{V}, \\ -w_p \sum |P^b|, & \underline{V} \le \tilde{V}_t \le \overline{V}. \end{cases} \tag{22}$$

where w_p and w_λ are weight coefficients. λ is a large positive penalty when there is a voltage magnitude violation in DNs. This reward function ensures that the agent prioritizes actions that maintain voltage magnitude within permissible limits.

Compared with centralized RL algorithms, the distributed RL model for voltage control faces two key challenges. First, effective convergence becomes difficult when positive rewards are sparse, such as in operating conditions where voltage violations occur infrequently. Since voltage magnitudes in DNs usually remain within acceptable limits, the RL agent receives only weak training signals, regardless of whether a ramp-type or step-type reward function is used. In particular, moderate control actions under normal conditions often do not trigger voltage magnitude violations and thus provide limited guidance for policy improvement. Second, voltage violation mitigation cannot be achieved solely from the perspective of an individual user, because the local voltage magnitude depends not only on its own demand but also on the demand of other users connected to the same feeder. To address these challenges, relaxing factors $\underline{\varsigma}$ and $\overline{\varsigma}$ are introduced into the reward function. To ensure that the resulting voltage interval remains non-empty (i.e., $\overline{V} - \overline{\varsigma} > \underline{V} + \underline{\varsigma}$) and physically reasonable, these factors are selected from $[0, 0.5]$ p.u.

$$R_t = \begin{cases} -w_p C_t^b \sum |P_m^b| - w_\lambda \lambda, & \tilde{V}_t > \overline{V} - \overline{\varsigma} \quad \text{or} \\ & \tilde{V}_t < \underline{V} + \underline{\varsigma}, \\ -w_p \sum |P_m^b|, & \underline{V} \le \tilde{V}_t \le \overline{V}. \end{cases} \tag{23}$$



(a) Local voltage magnitude estimation

(b) RL agent training in distributed environment

Fig. 3. Core workflow of the proposed distributed RL method: (a) local voltage estimation based on the Thevenin-equivalent model, and (b) RL agent training in the constructed local environment.

The SOC of the ESS at time t, denoted as SOC_t, is updated according to (7) and (8). The remaining state variables are updated sequentially as follows:

$$V_{t-2} = V_{t-1}, \tag{24}$$

$$V_{t-1} = V_t, \tag{25}$$

$$P^b_{t-2} = P^b_{t-1}, \tag{26}$$

$$P^b_{t-1} = P^b_t. \tag{27}$$

Each household agent is trained locally based on the local environment, which updates the state S_t and rewards R_t. In this situation, R_t is obtained by $\tilde{V}_t$ and the expression (23), guiding the agent towards an approximate convergence direction.

Nevertheless, independently trained agents face two major challenges when deployed in multi-agent voltage regulation systems. To keep the local voltage magnitude within acceptable limits, an agent may either (i) depend on redundant ESS capacity or (ii) learn overly aggressive control actions that are locally effective but globally excessive. The voltage magnitude at a given node depends not only on its own power injection but also on the load dynamics and control actions of adjacent nodes. Thus, when neighbouring agents are not coordinated, a local agent may apply stronger corrective actions than would be necessary from a network perspective. Since each agent reacts only to its own voltage condition, independently trained agents tend to charge when local voltage is high and discharge when it is low. Such synchronous behaviour can produce an excessive aggregate response, which may shift rather than mitigate load peaks and may even worsen voltage magnitude violations.

3.2. Online action coordination mechanism

To mitigate this challenge, two coordination strategies are considered, namely a consensus-free coordination scheme and an optimization-based action-refinement scheme. The first strategy is communication-light and can be implemented offline, and the second enables more explicit online coordination among neighbouring agents.

3.2.1. Consensus-free coordination

A simple coordination scaler, denoted by β, is introduced to modify agent actions. During online deployment, the scaler is used to adjust both the action magnitude and the corresponding available ESS capacity. The overall online deployment process of the proposed distributed RL framework is summarized in Algorithm 1. The coordinated action applied to the ESS at node m is formulated as:

$$\tilde{a}_m = \beta a_m \tag{28}$$

Algorithm 1: Online deployment of locally trained RL agents

Input: $\underline{V}, \overline{V}, \underline{P}^b, \overline{P}^b, \overline{SOC}, \underline{SOC}, \beta, ANN^0$, trained agents
Output: Optimized actions for each agent
for *each agent* $m \in \mathcal{B}$ **do**
  Recall recent voltage measurements V_{t-1}, V_{t-2} and load measurements P_{t-1}, P_{t-2};
  Generate local action $a_t \leftarrow$ Trained agent(S_t);
  Estimate Z^{th}_m and E^{th}_m via (11) and (12);
  Compute $\hat{V}_t$ via (17);
  Obtain $V^{adj}_t \leftarrow \mathcal{H}(\cdot)$ via (19);
  Obtain $\tilde{V}_t \leftarrow \mathcal{E}(\cdot)$ via (20);
  if $\tilde{V}_t < \underline{V}$ ***or*** $\tilde{V}_t > \overline{V}$ **then**
    Adjust a_t to prevent severe voltage magnitude violations;
  else
    Share $\tilde{V}_t$ and a_t with neighbouring agents;
  end
end
Update coordinated actions $\tilde{a}_m \leftarrow \beta a_m, \ \forall m \in \mathcal{B}$ via (28);
Update local actions a^* by solving the coordination model in (23)–(33);
Apply actions a^* to the ESSs locally;

$$\tilde{E}^{rated}_m = \beta E^{rated}_m. \tag{29}$$

Here, $\tilde{a}_m$ denotes the coordinated action used for ESS charging/discharging control. Since this strategy only requires the determination of a scaler and does not rely on iterative real-time interaction, its communication burden is low, and its sensitivity to communication delay is limited.

3.2.2. Optimization-based action-refinement

An optimization-based action-refinement model is introduced to coordinate independently trained agents by dynamically adjusting their actions during online operation. During online deployment, each agent exchanges limited information with its neighbouring agents Ω, including its action a_t and the estimated voltage magnitude $\tilde{V}_t$.

The objective function in (30) is designed to minimize action adjustments during online coordination, so that the coordinated actions remain as consistent as possible with the original actions generated by the local agents. This helps avoid unnecessary disruption to the subsequent ESS charging/discharging schedule while still enforcing voltage magnitude regulation. The optimization problem is subject to constraints on voltage magnitude limits as well as bounds on action

adjustments. The estimated voltage profile is represented as a vector $\tilde{V}_{\text{est}} = [\tilde{V}_1, \ldots, \tilde{V}_{|\Omega|}]$. The optimization-based action-refinement problem is formulated as:

$$\min_{\mathbf{a}} \quad \sum_{m=1}^{\mathcal{B}} (a_{m,adj} - \tilde{a}_m)^2 \tag{30}$$

s.t.

$$\tilde{V}_{\text{est}} = \mathcal{F}_v(\tilde{a}_1, \ldots, \tilde{a}_{|\Omega|}, \tilde{V}_1, \ldots, \tilde{V}_{|\Omega|}), \qquad \forall m \in \mathcal{B} \tag{31}$$

$$\underline{V} \le \tilde{V}_{\text{est}} \le \overline{V}, \qquad \forall m \in \mathcal{B} \tag{32}$$

$$\underline{P}^b \le a_{m,adj} \le \overline{P}^b, \qquad \forall m \in \mathcal{B} \tag{33}$$

The function $\mathcal{F}_v(\cdot)$ is introduced to characterize the voltage sensitivity to the actions of multiple agents and is used to estimate the coordinated voltage magnitudes. This estimation depends not only on the local agent's action and voltage magnitude, but also on the corresponding information received from neighbouring agents. In addition, the feasible action adjustment is constrained by the ESS state of charge SOC_t and the charging/discharging power limits.

When active power variations are relatively small, the relationship between voltage magnitude variation and active power injection (i.e., a_t) can be approximately described by a sensitivity-based model, as discussed in [39]. Motivated by this observation, a high-order polynomial function $\mathcal{F}_v(\cdot)$ is adopted to approximate the nonlinear relationship between action variation and voltage magnitude variation. Its parameters are identified using historical data. Although the exchanged coordination variables (e.g., actions) may reveal limited information about agent parameters, they do not expose the original SM data.

3.3. Deployment and reproducibility

To improve the reproducibility of the proposed physics-informed distributed RL framework, this subsection summarizes its implementation procedure, including discussion of local training, online deployment, coordination, communication requirements, and privacy considerations. Algorithm 2 summarizes the overall implementation procedure of the proposed framework.

The proposed framework consists of two stages: local offline training and online deployment. During the offline training stage, each SM first prepares its local dataset using only locally collected historical measurements. Specifically, two consecutive local samples are used to estimate the Thevenin-equivalent parameters. Based on these estimates, the corresponding voltage magnitudes are calculated, and then the locally measured voltage magnitudes at the same SM and time step are collected, which are used to fit the polynomial correction function. After obtaining function $\mathcal{H}(\cdot)$, additional locally constructed samples, including corrected voltage estimates, local actions, and the corresponding measured voltage responses, are used to train the Transformer-based correction model. Based on these locally prepared data, each SM constructs its own local environment and trains its RL agent independently. Therefore, the proposed framework does not require repeated centralized PF-based label generation or synchronized feeder-wide data exchange during training.

Regarding data preprocessing, filtering was explored in preliminary experiments to suppress abnormal fluctuations in the estimated Thevenin parameters. Although filtering improved smoothness, it also introduced temporal delays, which weakened the immediate correspondence between control actions and observed voltage responses. Since such a delay may interfere with RL training, no additional filtering stage is included in the final framework. For missing SM samples, simple interpolation using adjacent samples is used to preserve data continuity. Alternatively, the latest estimated parameter can be reused as a fallback strategy. In addition, the input data are normalized before training the Transformer neural network and RL agents to improve numerical stability and facilitate model convergence. During the online deployment stage, each trained agent determines its local ESS action based on the current local state. The Thevenin-equivalent parameters are then updated using recent measurements, and the voltage magnitude is estimated and corrected accordingly. The two coordination strategies considered in this work have different communication requirements. The scaler-based coordination strategy in Section 3.2.1 can be implemented in an offline way, as agents only need to exchange limited action-related or scaler-related information after local actions are employed in order to determine appropriate scaling factors. These scaling factors can then be applied over a subsequent time

Algorithm 2: Overall implementation procedure of the proposed physics-informed distributed RL framework

Input: $\underline{V}, \overline{V}, \underline{P}^b, \overline{P}^b, \overline{SOC}, \underline{SOC}, \beta, \gamma, w_p, w_\lambda, \lambda, \overline{\varsigma}, \underline{\varsigma}$, local historical measurements, normalization rules, neighbouring node information

Output: Coordinated local actions for voltage magnitude regulation

Stage 1: Offline local training

for *each agent* $m \in \mathcal{B}$ **do**
- Prepare local historical samples from voltage, load, and actions measurements;
- Estimate Thevenin-equivalent parameters via (11)–(12) in Section 2.2;
- Construct the local voltage estimation and correction model via (17)–(20);
- Build the local RL environment according to the state, action, and reward design in Section 3.1;
- Train the local RL agent independently and store the learned policy;

end

Stage 2: Online local decision making

for *each time step* t **do**
- **for** *each agent* $m \in \mathcal{B}$ **do**
  - Observe the current local state S_t;
  - Generate local action a_t using the trained policy in Section 3.1;
  - Update the Thevenin-equivalent parameters using recent local measurements via (11)–(12);
  - Estimate and correct the local voltage magnitude via (17)–(20);
  - **if** $\tilde{V}_t < \underline{V}$ ***or*** $\tilde{V}_t > \overline{V}$ **then**
    - Adjust the local action to avoid severe voltage magnitude violations;
  - **end**
- **end**
- **Stage 3: Coordination**
- **if** *the scaling-based coordination strategy in Section 3.2.1 is adopted* **then**
  - Exchange limited coordination variables with neighbouring agents;
  - Update coordinated actions via $\tilde{a}_m = \beta a_m$ in (28);
- **else**
  - Exchange action-related variables with neighbouring agents;
  - Refine local actions by solving the coordination model in (23)–(33);
- **end**
- **Stage 4: Execution and update**
- Apply the coordinated action locally to the controllable device;
- Store new local measurements for optional local model refinement;

end

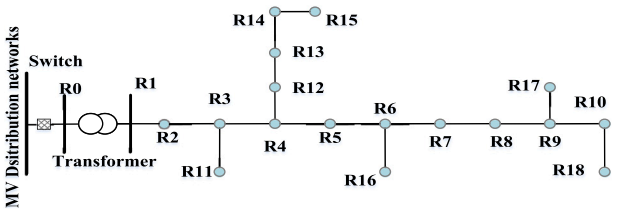


Fig. 4. Topology of the 18-bus CIGRE low-voltage distribution network.

Table 3
Main parameter settings for RL agent training.

Parameter	Value
RL algorithm	PPO
Policy network	MLP
Batch size	64
Learning rate	1×10^{-4}
Discount factor γ	0.5
Parameter λ	1000
ESS installation nodes	R9, R14, R16
ESS rated power	60 kW
ESS capacity	3 MWh
Voltage operating range	[0.95, 1.05] p.u.
SOC constraint	[10%, 90%]

Table 4
Main settings of the piecewise polynomial functions.

Parameter	Value
Number of segments in $\mathcal{H}(\cdot)$	3
Segment 1 interval	[0, 390 V]
Segment 2 interval	[390, 400 V]
Segment 3 interval	[400 V, ∞)
Order of $\mathcal{H}_L(\cdot)$, $\mathcal{H}_M(\cdot)$, $\mathcal{H}_H(\cdot)$	3,3,3
Order of $\mathcal{F}_v(\cdot)$	4

period, which reduces the communication burden and coordination frequency. Therefore, this strategy is relatively insensitive to communication delays, packet losses, and asynchronous updates. In contrast, the real-time action correction strategy in Section 3.2.2 requires action exchange at each time step for online refinement and is therefore more sensitive to non-ideal communication conditions. Its performance may degrade in the presence of delay, packet loss, or asynchrony. From a privacy perspective, no raw SM data are shared during either training or deployment. Historical SM data are only used for local model construction and refinement. For the scaling-based coordination strategy, only scaling-related parameters are exchanged, which can also be implemented through privacy-aware techniques if required. For the real-time action correction strategy, only action-related variables are shared for local optimization. Although these variables may reveal limited information about local agent behaviour, they may be used to infer the structure of the agent, but cannot be used to reconstruct the original inputs or private user data.

4. Simulation analysis

The CIGRE low-voltage distribution network is adopted as the test system, in which each node is connected to specific loads, as illustrated in Fig. 4. The corresponding load profiles are taken from [40], with a time resolution of 15 min. The PF results obtained using Pandapower are used as the benchmark for centralized calculations. The main parameter settings for RL training, the piecewise polynomial functions and the Transformer-based correction model are summarized in Tables 3, 5, and 4, respectively. All simulations and model training were implemented in Python on a laptop equipped with an 11th Gen Intel Core processor (4 cores and 8 logical processors) and 16 GB RAM.

Table 5
Hyperparameters and experimental settings of the Transformer model.

Parameter	Configuration/Value
Total samples	16,000
Training set size	12,000 (75%)
Validation set size	2000 (12.5%)
Test set size	2000 (12.5%)
Attention heads (*nhead*)	4
Encoder layers (*num_layers*)	2
Feedforward dimension	128
Dropout	0.1
Optimizer	AdamW
Learning rate	0.001
Weight decay	1×10^{-4}
Batch size	64
Maximum epochs	200
Early stopping patience	20
Learning-rate scheduler	ReduceLROnPlateau
Normalization	MinMaxScaler
Default loss function	MSE
Model selection criterion	Lowest validation loss

4.1. Accuracy analysis of voltage magnitude estimation

4.1.1. Piecewise adjustment function-based voltage magnitude correction

The performance of the function $\mathcal{H}(\cdot)$ is evaluated using RMSE, MAE, and the 95th percentile of the absolute error (i.e., donated as $P95$). The $P95$ indicator reflects the upper-tail error behaviour and is therefore useful for assessing robustness against infrequent but large deviations. To jointly evaluate average accuracy and tail-error suppression capability, a composite metric was defined as

$$Score = RMSE + 0.2 \times MAE + 0.1 \times P95. \tag{34}$$

As shown in Fig. 5 and Table 6, all error indicators decrease monotonically as the polynomial degree of the function $\mathcal{H}(\cdot)$ increases from 1 to 8. The first-order model exhibits the poorest accuracy, with an RMSE of 4.3002, an MAE of 2.8198, and a $P95$ value of 6.7676, indicating that linear compensation is insufficient to describe the nonlinear and complex relationship in voltage magnitude correction. A substantial reduction in all metrics is shown when the polynomial degree is increased to 2 and 3, suggesting that the dominant nonlinear characteristics can already be captured by low-order polynomial terms. When the degree is increased beyond 5, the gains become progressively smaller. Specifically, the RMSE decreases from 2.1276 at Degree 6 to 2.0967 at Degree 8, while the composite $Score$ decreases only from 2.7816 to 2.7366. In addition, the error density plots and boxplots show that higher-order models produce a more concentrated error distribution around zero and a narrower dispersion range. The decreasing $P95$ values further validate that the proposed Piecewise adjustment function-based voltage magnitude estimation strategy effectively reduces large error events in the tail region. Among all tested cases, the eighth-order piecewise polynomial model achieves the best overall performance. However, the limited improvement from Degrees 4 to 8 indicates that the model accuracy tends to saturate at high polynomial orders. Therefore, Degrees 4–8 can be regarded as the most competitive configurations, with Degree 8 providing the highest accuracy and Degrees 2–4 offering a better balance between complexity and performance.

4.1.2. Transformer-based voltage magnitude correction

Table 5 summarizes the main hyperparameters and training settings of the Transformer neural network. To further clarify the training process and assess the robustness of the adopted configuration, additional case studies were conducted by varying the input scaler for normalization and loss function, as summarized in Table 7. The maximum number of training epochs was set to 200. However, an early stopping constraint based on the validation loss is employed, and therefore some cases converged before reaching the maximum epoch.

Table 6
Quantitative comparison (unit is V) of error metrics for the function (19) with different polynomial degrees.

Degree	RMSE	MAE	P95	Score
1	4.3002	2.8198	6.7676	5.5409
2	2.9188	1.3575	5.0461	3.6950
3	2.3439	1.1751	4.7447	3.0534
4	2.3345	1.1364	4.8029	3.0421
5	2.2500	1.0394	4.8074	2.9386
6	2.1276	1.0049	4.5307	2.7816
7	2.1238	0.98846	4.5498	2.7765
8	2.0967	0.97852	4.4416	2.7366

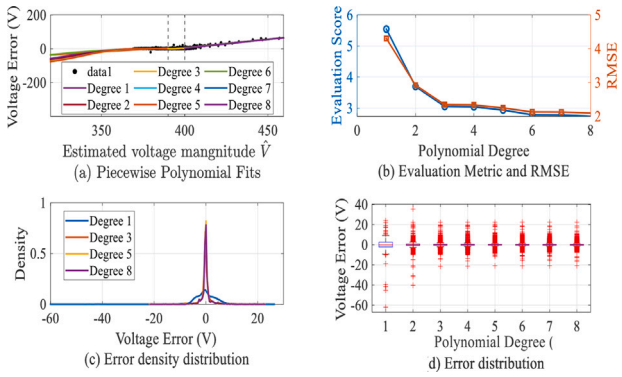


Fig. 5. Performance evaluation of the proposed piecewise polynomial compensation model under different polynomial degrees, including fitting curves, error metrics, density distributions, and boxplots.

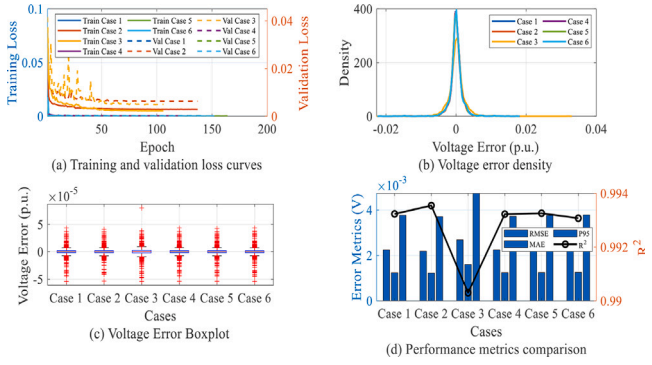


Fig. 6. Performance comparison of the six Transformer training cases, including loss curves in (a), voltage error distributions in (b), boxplots in (c), and quantitative evaluation metrics in (d).

Fig. 6 compares the six Transformer training cases in terms of convergence characteristics, error distribution, and prediction accuracy. As shown in Fig. 6(a), all cases converge stably, with both the training and validation losses decreasing rapidly during the initial stage and then gradually reaching convergence. The error distributions in Fig. 6(b) and the boxplots in Fig. 6(c) show that the error of voltage magnitude correction remains closely centred around zero with similar dispersions across all cases. This trend is further confirmed by the quantitative results in Fig. 6(d), where RMSE, MAE, P95, and R^2 exhibit only minor variations among the six cases. These results indicate that the accuracy of Transformer-based voltage magnitude correction is only weakly affected by the choice of scaler and loss function for the present voltage estimation task. Therefore, a standard training setting, such as MinMax scaling with MSE loss, is sufficient to ensure reliable correction accuracy.

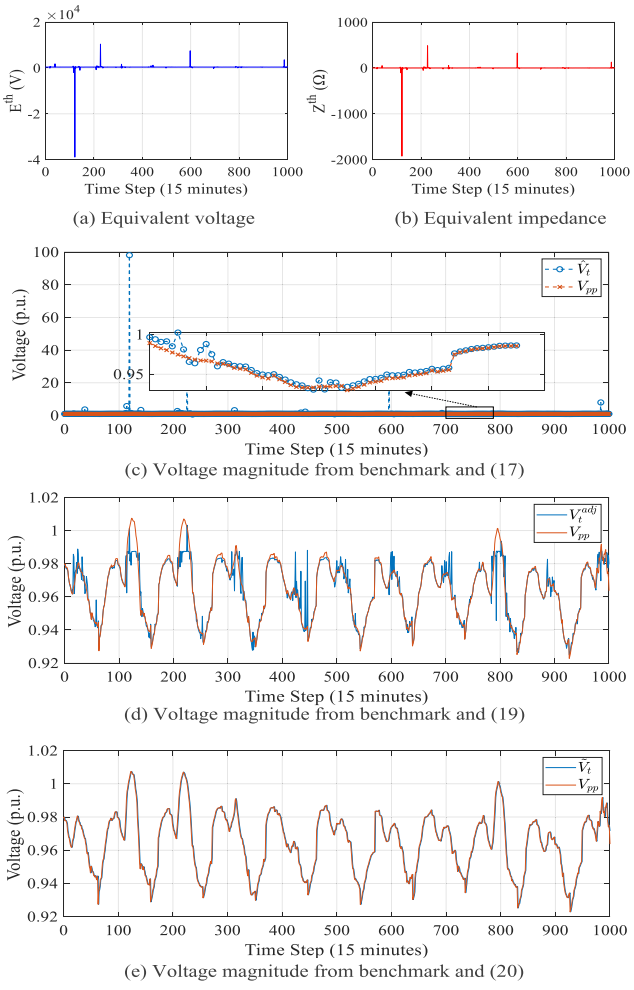


Fig. 7. Local voltage-magnitude estimation and error distribution: (a) the value of estimated Eth, (b) the value of estimated Zth, (c)–(e) estimated voltage magnitudes.

4.1.3. *Voltage magnitude estimation in the local environment*

The estimated voltage magnitude at node $R10$ from each sub-step are depicted in Fig. 7. V_{PP} represents the benchmark voltage magnitude obtained from Pandapower (i.e., taken as benchmark). Fig. 7(a) and (b) shown dynamic Eth and Zth. Fig. 7(c) depicts the estimated voltage magnitude $\hat{V}_t$ only using Eth and Zth, where there are multiple significant errors (e.g., 100 p.u. at time step 100) induced by the abnormal value of parameters Eth and Zth.

Based on Fig. 7(c), excluding significant errors, the relative error between the estimated voltage magnitude $\hat{V}_t$ and the true value V_{PP} ranges from 0.2% to 2500%. As shown in Fig. 7(d), given the piecewise function $\mathcal{H}(\cdot)$ (i.e., 3 3-order functions), the abnormal voltage magnitude is corrected into the reasonable region (i.e., 0.9–1.05 p.u.). Relative Error 1 is generally limited to a range of approximately ±4%. This demonstrates that the proposed piecewise function compensates for voltage magnitude estimation errors arising from Thevenin parameter equivalence, bringing the estimated voltage magnitude closer to true values. Nevertheless, the adjusted voltage magnitude exhibits abrupt high-frequency peaks, resulting from the inability of the piecewise function to cover all scenarios due to parameter inaccuracies. These abrupt high-frequency peaks in the estimated voltage magnitude prevent the agents from accurately distinguishing whether the voltage magnitude changes are caused by their current actions or by inherent estimation errors, thereby affecting convergence. In addition, although the voltage magnitude becomes more accurate after modification by the piecewise

Table 7
Case studies for the Transformer training process.

Item	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Scaler	MinMaxScaler	StandardScaler	RobustScaler	MinMaxScaler	MinMaxScaler	MinMaxScaler
Loss function	MSE	MSE	MSE	MSE	Huber	SmoothL1
Purpose	Baseline	Scaler analysis	Scaler analysis	Baseline	Loss comparison	Loss comparison

function, initial results indicate that the adjusted voltage magnitude still deviates from the true value when an action is applied. This is because the simplified Thevenin equivalent and the piecewise function cannot fully imitate the sensitivity of the node to the action. The voltage magnitude modified by the trained Transformer is shown in the third row in Fig. 7(e). The Relative Error 2 spans a broader range, typically within ±2%. Fig. 7 depicts that the further adjusted voltage magnitude $\tilde{V}_t$ not only aligns more closely with the true voltage magnitude V_{pp}, but also effectively mitigate high-frequency variations.

4.2. Convergence and efficiency of distributed agents under different training environments

This subsection presents a comparative analysis of the convergence behaviour and voltage magnitude regulation performance of the distributed agents under different training environments. Specifically, two environment configurations are considered for convergence evaluation, namely the centralized PF-based environment and the approximate local voltage magnitude estimation environment introduced in Section 3.1.1. In addition, a broader comparison is conducted among the proposed local environment, the pandapower-based environment, and a purely ANN-based environment. Three ESSs are installed at nodes R9, R14, and R16, and the corresponding agents are denoted as Agent 9, Agent 14, and Agent 16, respectively. To provide a representative and concise comparison, the quantitative performance assessment is conducted at node 9, as summarized in Table 8, while the convergence characteristics are shown in Fig. 8.

In Fig. 8(a), the dashed curves represent the evolution of cumulative rewards during the RL agent training process using the centralized PF model, whereas the solid curve corresponds to cumulative rewards under the distributed voltage magnitude estimation approach. Despite the presence of estimation errors inherent to the distributed approach, reflected in a slight deviation in the final convergence value, the overall learning dynamics exhibit comparable convergence trends. The final reward deviation across the three training cases is less than 10, corresponding to an approximately 30% relative error. This observation indicates that the distributed voltage magnitude estimation approach maintains sufficient fidelity to enable policy learning by the agent.

Fig. 8(b) further reveals a substantial difference in training duration between the two models. Specifically, the distributed approach achieves a convergence speedup of about 3× compared with the centralized approach. All training procedures were executed on an identical computing infrastructure, thereby eliminating hardware-induced variability and ensuring comparability. It should be emphasized that the evaluation excludes the latency associated with SM data acquisition in the centralized configuration. In practical scenarios, the process of aggregating distributed SM data is likely to introduce additional delays and computational burdens. By contrast, the decentralized approach relies on only locally available SM data, thereby obviating the need for external data transmission. This not only reduces the training time but also mitigates the privacy risks associated with centralized data collection and communication. The experimental results underscore that the proposed distributed approach constitutes a computationally efficient and privacy-aware alternative to centralized training algorithms. Despite its simplified formulation, it supports effective agent training with minimal degradation in performance, thereby demonstrating its potential for practical deployment in DNs.

To further evaluate the effectiveness of the proposed local training environment, an additional comparison was carried out using three

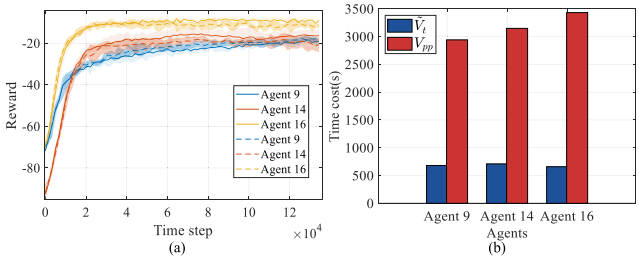


Fig. 8. Training rewards and time cost for the centralized PF-based model and the proposed distributed model. In (a), dashed and solid curves represent the reward trajectories of the two models, respectively. In (b), red and blue bars indicate the corresponding training time. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 8
Performance comparison of different training environments for voltage regulation at node 9.

Environment	Stable reward	Conv. step	VVR (%)	Max. dev. (p.u.)	Avg. dev. (p.u.)
PF-based	−19.52	83 673	5.71	0.0583	0.0398
Proposed	−17.10	92 517	7.86	0.0583	0.0412
Pure ANN-based	−15.02	86 395	9.29	0.0707	0.0424

different environment settings: the proposed local environment, the pandapower-based environment, and a purely Transformer ANN-based environment (i.e., totally black box). The comparison focuses on node 9 and considers both the training characteristics and the final voltage magnitude regulation performance. The stable reward and convergence step are used to describe the training process, while the voltage violation rate (VVR), the maximum voltage deviation, and the average voltage deviation from 1.0 p.u. are used to assess the control performance. The VVR is defined as:

$$VVR = \frac{N_{\text{viol}}}{N_{\text{total}}} \times 100\%, \tag{35}$$

where N_{viol} denotes the number of time steps at which the voltage magnitude violates the prescribed range, and N_{total} is the total number of evaluated time steps. Here, the stable reward is defined as the average reward over the final training window and is used to represent the converged training performance of each training environment.

As summarized in Table 8, the PF-based environment yields the best overall performance in terms of VVR and voltage deviation, while the proposed local environment performs slightly worse but remains close to the PF-based benchmark. This indicates that the proposed local voltage estimation and correction mechanism can provide sufficiently informative training signals for RL-based voltage control without requiring a centralized PF solver. Compared with the purely ANN-based environment, the proposed method shows slightly better voltage magnitude regulation performance, suggesting that a fully data-driven approximation is feasible for voltage estimation and RL training, but may introduce additional bias into the learned policy. More importantly, the pure ANN-based environment is fully black-box, and its combination with RL further limits interpretability. In contrast, the proposed framework is physics-informed and partially interpretable, since its voltage estimation stage is explicitly grounded in the Thevenin-equivalent formulation and structured correction modules. It is also worth noting that

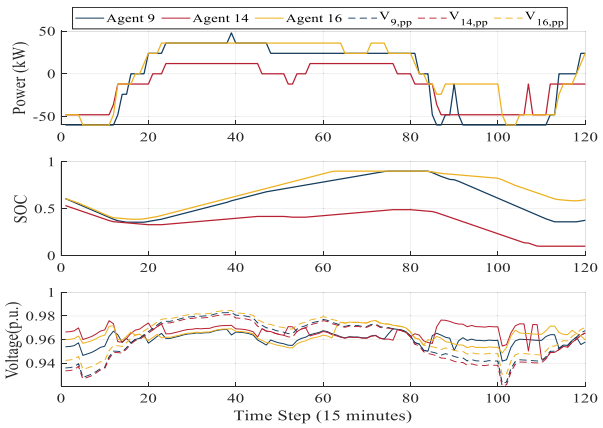


Fig. 9. Power, SOC of ESS, and voltage magnitudes at nodes R9, R14, and R16 with locally trained agents. Dashed lines denote the centralized-model case without control, and solid lines denote the proposed approach. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

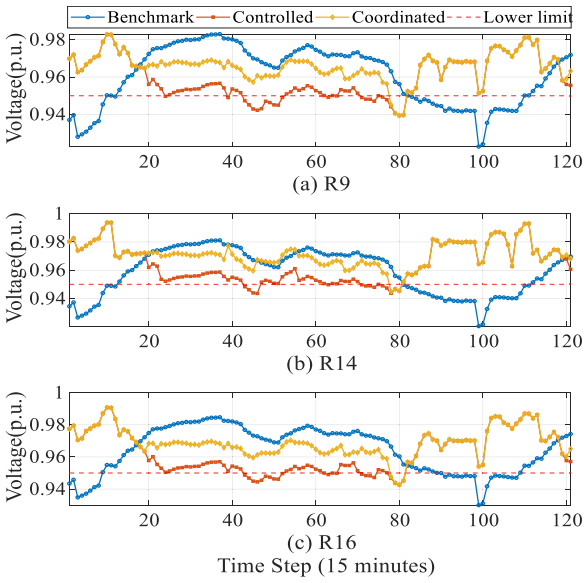


Fig. 10. Voltage magnitudes at node R9 in (a), node R14 in (b), and node R15 in (c) under three conditions: no regulation (Benchmark), individual control, and coordinated control. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the stable reward does not perfectly align with the final voltage magnitude regulation metrics, since the reward reflects the learning objective in the training environment, whereas VVR and voltage deviations directly characterize the physical control performance.

4.3. Online implementation and coordination performance

This section investigates the practical feasibility of the proposed method, coordination strategy, and extension possibilities. First, each RL agent is deployed sequentially within DNs and evaluated independently. The charging and discharging power of the ESSs, their corresponding SOC, and the voltage magnitude at the associated bus are depicted in Fig. 9. Dashed lines represent the benchmark voltage magnitudes (i.e., $V_{9,pp}$, $V_{14,pp}$, and $V_{16,pp}$) at nodes, obtained from the centralized model without any control strategy. Generally, the uncontrolled voltage profiles tend to violate the lower voltage magnitude limits more frequently than the upper limits. This is primarily attributed to high residential demand during certain periods. In response, the trained RL agents learn to discharge the ESSs to increase voltage magnitude, as evidenced by the SOC trajectories. Specifically, the distributed approach achieves a convergence speedup of about 3× compared with the centralized approach. As a result, in periods of low demand or when voltage magnitudes are within acceptable ranges, the agents proactively charge the ESS, which may lead to a slight decrease in voltage magnitude. Overall, compared to uncontrolled scenarios, the voltage profiles regulated by the agents exhibit reduced fluctuations and fewer violations of voltage limits. This indicates that locally trained RL agents, even when acting independently, are effective in voltage regulation and supporting the safe operation of the DNs. Although the agent mitigates voltage violations, the voltage magnitude at node 9 exhibits periods with $V_9 < 0.95$ because the ESS discharging is constrained by its maximum rated power limit.

In the initial scenario, three agents (i.e., located at nodes R9, R14, and R16) share information with one another. To demonstrate the scalability of the proposed coordination approach, this setup is extended by introducing additional agents at nodes R7, R15, and R18, which are neighbours of R9, R14, and R16, respectively. In this expanded scenario, each agent shares information only with its neighbours. The raw voltage magnitudes represent the baseline case without any voltage magnitude regulation. The controlled voltage magnitudes are obtained from agents trained independently, while the coordinated voltage magnitudes are obtained under the case in which independently trained agents operate under the proposed coordination strategy.

First, the coordination scaler β is set as 1, i.e., the coordination strategy relies solely on the optimization model. The results are depicted in Fig. 10 and Fig. 11(a). As expected, when multiple independently trained agents are simultaneously deployed for voltage magnitude regulation, their independent decisions may lead to excessive and uncoordinated charging or discharging of multiple ESSs, shown as the deep orange curves in Fig. 10 and light green curves in Fig. 11(a). This behaviour results in local voltage magnitudes exceeding operational limits (e.g., time steps 40–60), as each agent is trained solely on its local voltage profile, without considering system-wide voltage magnitude interactions. This phenomenon is verified and illustrated when individually trained agents at nodes R9 to R16 are applied in parallel, a simultaneous charging event causes the voltage magnitude to drop below the lower operational threshold of 0.95 p.u.. For coordination, agents exchange both their intended actions and estimated voltage magnitude. Using this shared information, each agent adjusts individual actions to ensure coordinated behaviour, thereby mitigating adverse system level effects. The effectiveness of this approach is demonstrated by the yellow curve in Fig. 10, where the adjusted control strategy significantly reduces voltage magnitude violations. However, the purple curves in Fig. 11(a) show that the optimization model fails to improve the coordination. This is because the adjusted space (i.e., constraint (33)) does not support sufficient coordination, which is verified as follows.

The residual voltage violations are mainly caused by the physical limits of the ESSs, including limited available energy and rated power, which restrict the corrective capability of the coordinated control. Even so, the proposed method improves steady-state voltage regulation performance and confirms the effectiveness of coordinated multi-agent control. Compared with optimization-based coordination, the scaling-factor-based mechanism is simpler to implement, since no online interaction is required once β is specified, which reduces communication burden. In addition, the framework only requires basic local communication and computation, and is therefore suitable for distributed deployment.

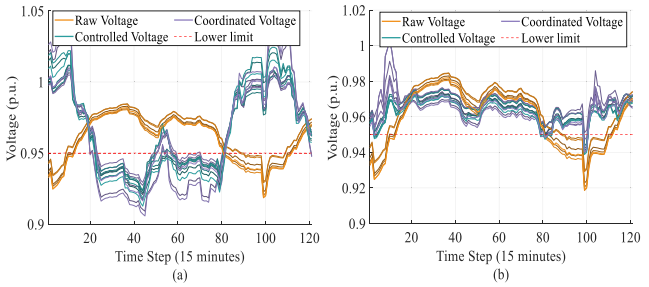


Fig. 11. Voltage magnitudes at six nodes under coordinated control: (a) with a scaling factor of 1.0 (i.e., relying solely on optimization models), and (b) with a scaling factor β of 0.2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

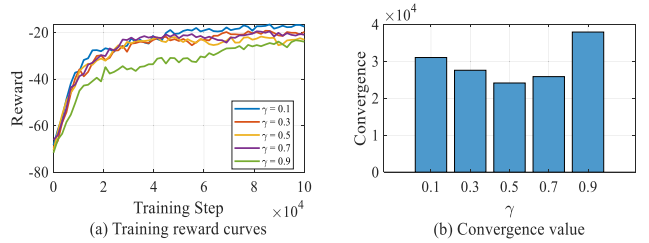


Fig. 12. Sensitivity analysis of the discount factor γ for the proposed RL-based local voltage control framework. (a) Training reward curves obtained under different values of γ. (b) Comparison of the corresponding convergence values, where a smaller value indicates faster convergence and higher training efficiency. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.4. Evaluating the sensitivity of parameter settings

This section evaluates the sensitivity of the proposed voltage regulation scheme with respect to three critical parameters: the discount factor γ, the relaxation factor ς and the scaler β.

First, to investigate the influence of the discount factor on the proposed RL-based local voltage control framework, a sensitivity study was conducted with $\gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, while keeping all other training settings unchanged. The convergence is defined by the relative improvement of the smoothed reward curve.

$$R_{\text{target}} = R_{\text{start}} + \eta \left(R_{\text{final}} - R_{\text{start}} \right), \tag{36}$$

where R_{start} and R_{final} are the average rewards over the initial and final windows, and $\eta = 0.95$. The convergence index t_{conv} is the earliest point at which the reward reaches and stays above R_{target} over a predefined window. It is then mapped to the full training horizon by:

$$\text{convergence} = \frac{t_{\text{conv}} - 1}{N - 1} T, \tag{37}$$

where N is the reward-curve length and T is the total number of training steps.

As shown in Fig. 12(a), all cases exhibit a similar overall learning trend, but noticeable differences remain in both the final reward and the convergence behaviour. The quantitative comparison in Fig. 12(b) shows that $\gamma = 0.5$ achieves the smallest convergence value, indicating the best training efficiency among the tested cases. By contrast, $\gamma = 0.9$ results in the slowest convergence and the least favourable overall performance. These results suggest that, unlike conventional centralized RL settings where $\gamma = 0.9$ is often used to emphasize long-term return, the present local voltage control task is better suited to a moderate discount factor. Therefore, $\gamma = 0.5$ is selected as the default setting in this study.

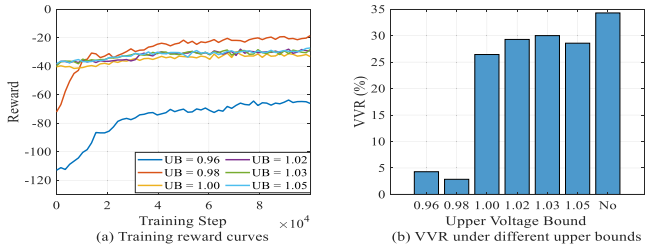


Fig. 13. Sensitivity analysis of the upper voltage bound used in the reward function. (a) Training reward curves under different upper-bound settings. (b) VVR under different upper bounds, where the first bar corresponds to the case without regulation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

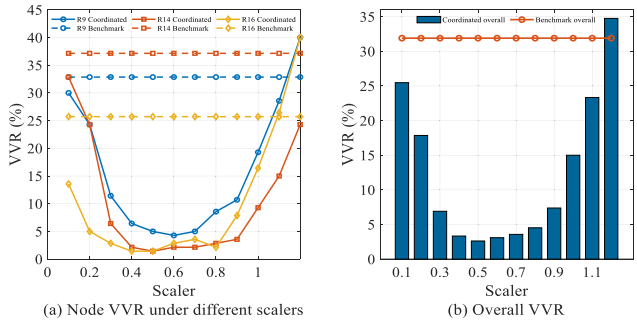
To further investigate the influence of the relaxing factors in the reward function, a sensitivity study was conducted by varying the upper voltage bound during training while keeping the lower bound unchanged. Given that the voltage distribution in the investigated LVDN is biased towards the lower limits, the upper voltage bound is critical for guiding the training of agents by impacting the reward values. Specifically, different upper-bound settings were considered during training, and the resulting agents were evaluated using both the training reward curves and the VVR is adopted as an indicator.

Compared to the discount factor γ, Fig. 13 indicates that the upper voltage bound used in the reward function has a substantial impact on both learning behaviour and voltage regulation efficiency. As shown in Fig. 13(a), smaller upper bounds lead to more evident reward improvement during training, whereas the cases with larger upper bounds exhibit only limited reward growth. This suggests that a stricter upper bound is more effective in shaping the learning process. The quantitative results in Fig. 13(b) further validate this observation: the configuration with a 0.98 upper bound attains the lowest voltage violation rate, demonstrating a significant improvement over the other upper-bound settings.

This behaviour can be explained by the role of the reward value in policy learning. When the upper bound is relatively small, voltage regulation is penalized more frequently. The reward sequence contains clearer distinctions between desirable and undesirable operating conditions. Thus, the agent receives more informative feedback and is more likely to learn an effective voltage regulation policy. In contrast, when the upper bound is too large, the penalty becomes less active during training, which weakens the guidance provided by the reward value. Although the training reward may still increase, the learned policy is less effective in practice. This can also be observed from the battery operation pattern, where the agent mainly learns discharging behaviour but fails to develop an effective charging strategy. Therefore, a relatively small upper voltage bound is preferable for the proposed framework.

Third, to further investigate the influence of the coordination scaler β on the proposed multi-agent voltage regulation approach, a case study was conducted by varying the coordination scaler β from 0.1 to 1.1. For each setting, the VVR was calculated at nodes R9, R14, and R16, together with the overall VVR of the coordinated case. The corresponding benchmark VVR without voltage regulation strategies was also included for comparison. This analysis aims to examine how the β affects the ability of multiple local agents to regulate the network voltage within the admissible operating range.

Fig. 14 indicates that the coordination scaler β plays a critical role in the proposed multi-agent coordination framework. As shown in Fig. 14(a), the VVRs at nodes R9, R14, and R16 decrease significantly as β increases from a small value of 0.1, but start to increase again once β exceeds an intermediate range 0.5–0.7. The same pattern is reflected in Fig. 14(b), where the overall VVR reaches its minimum around



(a) Node VVR under different scalers (b) Overall VVR

Fig. 14. Sensitivity analysis of the coordination scaler. (a) Node VVR under different scalers. (b) Overall VVR comparison between the coordinated and benchmark cases.

$\beta = 0.6$. In all cases near this region, the coordinated strategy clearly outperforms the benchmark case without coordination. These results suggest that when β is too small, the coordination among RL agents is not sufficient to induce effective voltage regulation actions from the local agents. Conversely, if β is too large, the combined responses of multiple agents may become excessive, resulting in over-regulation and degraded voltage performance. Hence, an intermediate coordination scaler is preferable for voltage regulation.

4.5. Discussion and limitations

The proposed physics-informed distributed RL framework provides a practically relevant alternative to centralized voltage regulation approaches by reducing the dependence on repeated global PF calculations and centralized data collection. By constructing local training environments from SM measurements, the framework supports decentralized agent training while limiting raw data sharing. These features are particularly relevant for LVDNs, where system visibility is limited and end-user flexibility is increasingly important. While this work focuses on household ESSs, the proposed framework is general and can be extended to other controllable devices whose operation can be formulated as a Markov decision process. Furthermore, although the present validation is based on a specific use case and a dataset from a single country, the methodology itself is not restricted to that national context and is readily applicable to other regional and international settings.

At the same time, several limitations should be noted. First, the accuracy of the local voltage magnitude estimation stage depends on the quality of the dynamic Thevenin model and the subsequent correction modules. When feeder conditions evolve significantly, such as under topology changes, changing load composition, or increasing DER penetration, the fitted polynomial and ANN-based correction models may gradually lose accuracy, making periodic refitting or retraining necessary. Second, the present study assumes the availability of controllable flexibility resources, such as household ESSs, and adopts an idealized ESS model with 100% charging/discharging efficiency. While this simplification helps isolate the effect of the proposed learning and coordination framework, it does not fully reflect practical battery behaviour and may limit applicability in less flexible settings. Third, the coordination strategy is evaluated only under a limited range of feeder conditions and agent densities. As the number of independently trained agents increases, coordination becomes more complex, and its scalability under denser and more heterogeneous systems remains to be established. Finally, the framework should be understood as privacy-aware rather than formally privacy-aware, since the current study does not provide a formal threat model, leakage analysis, or cryptographic privacy guarantee.

These limitations also point to several directions for future research. In particular, further work is needed on scalable coordination under denser multi-agent settings, robustness under imperfect data and communication conditions, and validation under more realistic deployment scenarios, including non-ideal storage behaviour and evolving feeder characteristics. In addition, the integration of explainability tools may help improve the transparency of learned control behaviours and the role of different voltage magnitude correction modules in the overall framework.

5. Conclusion

This paper presents a physics-informed distributed reinforcement learning framework for voltage regulation in distribution networks, introducing a fully distributed decision making intelligence to smart meters. By constructing local training environments from smart meter measurements, the proposed method enables effective voltage regulation using only local smart meter data, thereby avoiding repeated centralized power flow calculations and reducing the need for real-time data sharing. The results show that the proposed hybrid correction mechanism, combining a piecewise function and a Transformer ANN-based sensitivity model, significantly improves local voltage estimation accuracy, while the coordination strategy enables effective coordination among independently trained agents with limited information exchange. The case studies further indicate that the proposed framework enables fully distributed deployment, reduces voltage violations by approximately 80%, achieves performance close to that of power flow-based training environments, and provides about a 6× training speedup. Sensitivity analyses further show that the relaxation factors in the reward function and the coordination scaler are more critical to voltage magnitude regulation performance. Nevertheless, several limitations should also be acknowledged. The framework depends on the accuracy of the local estimation and correction modules, assumes idealized ESS behaviour, and has been validated under a limited range of feeder conditions and agent densities. Future work will therefore focus on improving scalability and robustness under more realistic deployment scenarios.

CRediT authorship contribution statement

Dong Liu: Writing – original draft, Validation, Software, Methodology, Conceptualization. **Juan S. Giraldo:** Writing – review & editing, Conceptualization. **Peter Palensky:** Funding acquisition. **Pedro P. Vergara:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the China Scholarship Council (CSC) under Grant No. 202206130017.

Data availability

Data will be made available on request.

References

[1] Wang X, Huang Y, Liu Y, Liang D, Zhou Y. Deep reinforcement learning-based coordinated optimization between distribution networks and microgrids towards demand response uncertainty. Energy AI 2026;100717.

[2] Yu X, Jiang R, Jin X, Jia H, Mu Y, Wei W, et al. Method for compensating multi-time measurement data of distribution network based on alternating minimization matrix completion combined with VMD-ARIMA-LSTM. Energy AI 2025;100563.

[3] Henry R, Ernst D. Gym-ANM: Reinforcement learning environments for active network management tasks in electricity distribution systems. Energy AI 2021;5:100092.

[4] Liu D, Giraldo JS, Palensky P, Vergara PP. Model-free privacy preserving power flow analysis in distribution networks. IEEE Trans Smart Grid 2025;16(6):5446–58.

[5] Vergara PP, Giraldo JS, Salazar M, Panda NK, Nguyen PH. A mixed-integer linear programming model for defining customer export limit in PV-rich low-voltage distribution networks. J Mod Power Syst Clean Energy 2023;11(1):191–200.

[6] Hu J, Ye C, Ding Y, Tang J, Liu S. A distributed MPC to exploit reactive power V2G for real-time voltage regulation in distribution networks. IEEE Trans Smart Grid 2022;13(1):576–88.

[7] Yang Y, Tang R. Cooperative multi-agent reinforcement learning for grid-aware EV charging management with cross-site redirection. Energy AI 2025;100664.

[8] Hassouna M, Holzhüter C, Lytaev P, Thomas J, Sick B, Scholz C. Graph reinforcement learning for power grids: A comprehensive survey. Energy AI 2026;100671.

[9] Hou S, Gao S, Xia W, Duque EMS, Palensky P, Vergara PP. RL-ADN: A high-performance deep reinforcement learning environment for optimal energy storage systems dispatch in active distribution networks. Energy AI 2025;19:100457.

[10] Liu H, Wu W, Wang Y. Bi-level off-policy reinforcement learning for two-timescale volt/VAR control in active distribution networks. IEEE Trans Power Syst 2023;38(1):385–95.

[11] Ge L, Li J, Hou L, Lai J. Autonomous voltage regulation for smart distribution network with high-proportion PVs:A graph meta-reinforcement learning approach. IEEE Trans Sustain Energy 2025;16(4):2768–81.

[12] Zhao Y, Liu J, Liu X, Yuan K, Ding T. Enhancing the tolerance of voltage regulation to cyber contingencies via graph-based deep reinforcement learning. IEEE Trans Power Syst 2024;39(2):4661–73.

[13] Zhang T, Yu L, Yue D, Dou C, Xie X, Shi T. Explainable deep reinforcement learning approach for smart voltage regulation of high renewable-penetrated distribution networks considering hydrogen-storage system. Electr Power Syst Res 2025;246:111654.

[14] Sun X, Xu Z, Qiu J, Liu H, Wu H, Tao Y. Optimal volt/var control for unbalanced distribution networks with human-in-the-loop deep reinforcement learning. IEEE Trans Smart Grid 2024;15(3):2639–51.

[15] Zheng X, Yu S, Cao H, Shi T, Xue S, Ding T. Sensitivity-based heterogeneous ordered multi-agent reinforcement learning for distributed volt-var control in active distribution network. IEEE Trans Smart Grid 2025;16(3):2115–26.

[16] Kim HJ, Kim D. Safe multi-critic reinforcement learning framework for coordinated energy management and voltage regulation in active distribution networks. Sustain Energy Grids Netw 2026;102179.

[17] Wu H, Xu Z. Prototype federated reinforcement learning for voltage regulation in distribution systems with physics-aware spatial-temporal graph perception. IEEE Trans Sustain Energy 2026;17(1):697–708.

[18] Liu X, Liu Y, Chen Y, Tang Z, Gao H, Li Z. Federated reinforcement learning based dual-level voltage regulation for PV-rich distribution grids. Int J Electr Power Energy Syst 2026;175:111492.

[19] Ahmed F, Arshad A, Rehman AU, Hussain GA, Lehtonen M. A multi-agent reinforcement learning framework for voltage-constrained incentive demand response in PV-rich low-voltage distribution systems. IEEE Access 2026;14:45410–22.

[20] Wei X, Qiu W, Wing Chan K, Yao W, Chung CY. Visual-based reinforcement learning for voltage regulation and attack mitigation in distribution networks. IEEE Trans Smart Grid 2026;17(1):132–43.

[21] An S, Qiu J, Lin J, Sun X, Liu B, Yuan Z. Planning and integration of a multi-robot adaptive charging network for voltage regulation via unscented Kalman filter-based deep reinforcement learning. IEEE Trans Smart Grid 2026;17(1):518–36.

[22] Chen P, Liu S, Wang X, Kamwa I. Physics-shielded multi-agent deep reinforcement learning for safe active voltage control with photovoltaic/battery energy storage systems. IEEE Trans Smart Grid 2023;14(4):2656–67.

[23] Khan M, Silva BN, Khattab O, Alothman B, Joumaa C. A transfer reinforcement learning framework for smart home energy management systems. IEEE Sens J 2023;23(4):4060–8.

[24] Lee S, Choi D-H. Federated reinforcement learning for energy management of multiple smart homes with distributed energy resources. IEEE Trans Ind Inform 2022;18(1):488–97.

[25] Zenginis I, Vardakas J, Koltsaklis NE, Verikoukis C. Smart home's energy management through a clustering-based reinforcement learning approach. IEEE Internet Things J 2022;9(17):16363–71.

[26] Wu Z, Chen X, Lin Y, Wen J, Chen Y. A smart home energy management system based on human activity recognition and deep reinforcement learning. Energy Build 2024;325:114951.

[27] Wang S, Du L, Fan X, Huang Q. Deep reinforcement scheduling of energy storage systems for real-time voltage regulation in unbalanced LV networks with high PV penetration. IEEE Trans Sustain Energy 2021;12(4):2342–52.

[28] Razghandi M, Zhou H, Erol-Kantarci M, Turgut D. Smart home energy management: VAE-GAN synthetic dataset generator and Q-learning. IEEE Trans Smart Grid 2024;15(2):1562–73.

[29] Zhang Z, Zhang Y, Yue D, Dou C, Ding X, Zhang H. Economic-driven hierarchical voltage regulation of incremental distribution networks: A cloud-edge collaboration based perspective. IEEE Trans Ind Inform 2022;18(3):1746–57.

[30] Cibaku E, Gama F, Park S. Boosting efficiency in state estimation of power systems by leveraging attention mechanism. Energy AI 2024;16:100369.

[31] Kaspar K, Hussain S, Menon RP, Ouf MM, Eicker U. Reinforcement learning for residential energy storage management at the neighborhood scale: A multi-benchmark evaluation. Energy AI 2026;100697.

[32] Wang J, Ma X, Yang L. Hybrid maintenance optimization for multi-state competing failure systems under inspection uncertainties and spare dynamics. Reliab Eng Syst Saf 2026;112524.

[33] Wu D, Gao K, Peng R, Sun J, Yang L, Wei F. Optimal preventive maintenance policy considering technical-innovation-empowered renewals. Proc Inst Mech Eng O 2026;1748006X251412931.

[34] Tan L, Wei F, Ma X, Peng R, Xiao H, Yang L. Systemic condition-based maintenance optimization under inspection uncertainties: A customized multiagent reinforcement learning approach. IEEE Trans Reliab 2025.

[35] Fu A, Cvetković M, Palensky P. Distributed cooperation for voltage regulation in future distribution networks. IEEE Trans Smart Grid 2022;13(6):4483–93.

[36] Smon I, Verbic G, Gubina F. Local voltage-stability index using tellegen's theorem. In: 2007 IEEE power engineering society general meeting. 2007, 1–1.

[37] Giraldo JS, Castrillon JA, Castro CA. Network-free voltage stability assessment of power systems using phasor measurements. In: 2015 IEEE Eindhoven PowerTech. 2015, p. 1–5.

[38] Vergara PP, Salazar M, Giraldo JS, Palensky P. Optimal dispatch of PV inverters in unbalanced distribution systems using reinforcement learning. Int J Electr Power Energy Syst 2022;136:107628.

[39] Chen YC, Wang J, Domínguez-García AD, Sauer PW. Measurement-based estimation of the power flow Jacobian matrix. IEEE Trans Smart Grid 2016;7(5):2507–15.

[40] Hou S, Fu A, Duque EMS, Palensky P, Chen Q, Vergara PP. DistFlow safe reinforcement learning algorithm for voltage magnitude regulation in distribution networks. J Mod Power Syst Clean Energy 2025;13(1):300–11.