Learning Volt-VAR Droop Curves to Optimally Coordinate Photovoltaic (PV) Smart Inverters

Daniel Glover, *Graduate Student Member, IEEE*, and Anamika Dubey, *Senior Member, IEEE*

***Abstract*—Learning-based solutions for power systems operational tasks are earning more consideration as potential candidates to help overcome challenges brought upon by the aggressive integration of inverter-based resources (IBRs) in active distribution networks (ADNs). Despite achieving high evaluation accuracies, machine learning (ML) methods are not yet accepted at utility-scale primarily due to safety concerns and limited interpretability. This presents an opportunity for ML approaches which can satisfy both performance and regulatory requirements. In an effort to improve these shortcomings, this work proposes a robust Deep Reinforcement Learning (DRL) based model-free adaptive volt-VAR control (VVC) dispatch framework of solar photovoltaic (PV) smart inverters (SIs) for system-wide voltage regulation and loss reduction. The framework utilizes reward shaping with a barrier function (BF) filter to embed physical boundaries for Category B-type SIs specified by the IEEE 1547-2018 standard into the constrained *Markov Decision Process* (CMDP) formulation. Results carried out on the IEEE 123 bus test system show that the proposed method converges to a robust discrete policy offline, producing QV-droop curves compliant with IEEE 1547-2018, which outperform the baseline benchmark during overloaded conditions.**

***Index Terms*—Deep reinforcement learning, volt-VAR, distributed energy resource, smart inverter, voltage regulation.**

NOMENCLATURE

$B_{pen,i}$	Barrier function violation penalty.
$D_{pen,i}$	Barrier function breakpoint distance penalty.
$\mathbf{D}^{qv}_{DER,i}$	DER QV-droop curve.
$\mathbf{D}^{qv}_{DER,1547catB}$	DER QV-droop IEEE 1547-2018.
$\mathbf{D}^{qv,c_i}_{DER,i}$	DER QV-droop voltage breakpoint constraints.
$D^{qv_{inf}}_{DER,i}$	DER QV-droop curve infraction.
Eff_{xy}	DER Efficiency curve.
I_{Nort}	Norton equivalent current source.
$K^{qv}_{DER,i}$	DER QV-droop curve gain.
$K_{Q_{abs}}$	QV-droop curve gain absorbing reactive power.
$K_{Q_{inj}}$	QV-droop curve gain injecting reactive power.
P_{DC}	Solar array active power (kW pu).
$P_{DER,i}$	DER Active power (kW pu).
$P^{out}_{DER,i}$	DER Watt priority active power (kW pu).
$P^{Rated}_{DER,i}$	DER Nameplate real power (kW pu).
$P_{D,i}$	Active power demand (kW pu).
P_{gen}	Active power generation (kW pu).
P_{loss}	Active power losses (kW pu).
P_{MPP}	Maximum power-point tracking (pu).
$Q^{\max}_{pu_{avail}}$	DER Max available reactive power (kVAR pu).
$Q_{DER,i}$	DER Reactive power (kVAR pu).
$Q^{avail}_{DER,i}$	DER Available reactive power (kVAR pu).
$Q^{limit}_{DER,i}$	DER Reactive power limit (kVAR pu).
$Q_{D,i}$	Reactive power demand (kVAR pu).
$S^{Rated}_{DER,i}$	DER Nameplate apparent power (kVA pu).
V_{dev}	Voltage deviation from V_{ref} (V pu).
V_{ref}	Grid reference voltage (V pu).

Received 22 March 2024; revised 4 July 2024; accepted 6 August 2024. Date of publication 2 October 2024; date of current version 31 January 2025. Paper 2023-ESC-1651.R1, presented at the 2023 IEEE Industry Applications Society Annual Meeting, Nashville, TN, USA, Oct. 29–Nov. 02, and approved for publication in the IEEE TRANSACTIONS ON INDUSTRY APPLICATIONS by the Energy Systems Committee of the IEEE Industry Applications Society [DOI: 10.1109/IAS54024.2023.10406599]. This work was supported by National Science Foundation (NSF) under CPS Grant # 2208783. *(Corresponding author: Daniel Glover.)*

The authors are with EECS, Washington State University, Pullman, WA 99164-1009 USA (e-mail: daniel.glover1@wsu.edu; anamika.dubey@wsu.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIA.2024.3472655.

Digital Object Identifier 10.1109/TIA.2024.3472655

I. INTRODUCTION

THIS work is an extension of [1], focusing on improving reinforcement learning techniques for autonomous centralized Volt-VAR control (VVC) grid operations through informative reward design. Distributed energy resource (DER) integration across all domains of the power grid continues to accelerate as environmental policies and carbon-free emissions goals remain at the forefront of the energy agenda worldwide. In the U.S., the Energy Information Administration (EIA) reports that nearly $11.2GW_{ac}$ of solar photovoltaics (PV) was installed in 2023 H1, up nearly 44% from the previous year. The percentage of electric capacity additions from solar PV is expected to grow in the U.S. from approximately $31GW_{ac}$ in 2023 (56%), to nearly $41GW_{ac}$ in 2024 (62%) [2]. Meanwhile, surging DER installations at the distribution level are causing more concern for operators due to the wide range of capacities, locations, and the vast sizes of unbalanced multi-phase networks [3].

Active distribution networks (ADNs) are no longer operating as unidirectional, passive delivery systems, but are now managing bi-directional power and information flows among a massive network of interconnected devices. Higher penetrations of inverter-based resources (IBRs) have introduced faster timescale dynamics and various uncertainties, reducing the accuracy of solutions using deterministic models [4]. The coinciding Big Data boom and deployment of cloud-based data storage,

TABLE I
REVIEW OF SMART INVERTER VOLT-VAR ADAPTIVE DROOP CONTROL METHODS FOR PHOTOVOLTAIC SYSTEMS

Source	Formulation	Technique	Method	Description	Drawbacks
[9]	Multi-objective	Robust optimization	LVDI	Multiple SIs & timescales	Scalability
[10]	Two-stage	Distributed optimization	ADMM	Network partitioning	Communication
[11]	Distributed	Local optimization	LVC	Three-strategy proposal	Assumptions
[12]	Direct	TS Power Flow	AVR	Voltage setpoint tracking	Fix-and-forget
[13]	Rule-based	Robust minimization	RCC	Compute slope equation	User-defined set
[14]	Distributed	Optimal Power Flow	MISOCP	Dyn reactive current control	Model required
[15]	Distributed	Optimal Power Flow	MILP	QV scheduling model	Scalability
[16]	Distributed	Optimal Power Flow	Multi-period	PV with BESS	Convergence
[17]	Two-layer	Local optimization	EAC	Voltage flicker stability, SSE	Global visibility
[18]	Model-free DRL	Multi-agent	DQN	Multiple regulating devices	No safety analysis
[19]	Model-free DRL	Constrained Markov model	SAC	Decoupled neural network	Black-box design
[20]	Model-free DRL	Multi-agent	MADDPG	Sub-network partitioning	Information sharing
[21]	Model-free DRL	Markov model	SAC	Hierarchial droop control	Interpretability

wide-area monitoring systems, and microPMUs at the system-level, alongside advanced metering infrastructure (AMI) and measurement devices at the grid-edge, are supplying potential real-time data which can be leveraged by utilities to enhance grid operations [5].

A. *Motivating Adaptive Centralized Volt-VAR Control*

Optimal VVC, which utilizes reactive power support for voltage regulation, remains an extensively studied topic in ADNs considering the problem of rising voltage instability and smart inverter (SI) capabilities to provide grid services [6]. In response, IEEE 1547-2018 was established to provide a uniform technical standard for all DER grid interconnections and performance requirements, including provisions for local autonomous reactive power support from SIs [7]. Successful centralized DER coordination is critical to achieving global system-level objectives. Central and distributed optimal power flow (OPF) formulations for VVC attempt to maximize power delivery and utility profits, while reducing system losses and voltage limit violations using off-the-shelf non-linear solvers, but convergence suffers from forecasting errors, physical modeling inaccuracies, and noisy data distributions [8].

Achieving fast optimal VVC dispatch solutions is becoming intractable using existing approaches that may only provide limited input data, making them very difficult to solve [22]. Optimizing under uncertainty requires robust, stochastic techniques [9] which pose scalability and accuracy issues under extremely varying data. Decentralized optimization methods have also been explored in the literature, where a two-stage distributed optimization approach designates real and reactive power rules for SIs with alternating direction method of multipliers (ADMM) to set legacy voltage regulation devices [10]. However, distributed optimization still requires some sort of centralized oversight and assumes the existence of a communication network between agents, which may not be feasible in ADNs [11]. To address these issues, our work combines the data-driven approach with centralized management and decentralized local autonomous SI controls to coordinate multiple DERs for improved system observability under various dynamic uncertainties, eliminating limited local blindness to global network information.

Regarding constrained optimization methods for VVC control, a significant drawback of optimizing SIs under the 1547-2018 standard lies directly within the recommended QV-droop curve settings established in Section V. Firstly, although the default settings are adequate during normal operating conditions, they lack adaptability to local environment dynamics and may prove suboptimal over time as the network evolves [12], [13], requiring constant re-calibration. For example, [12] suggests using an updated default setting by eliminating the dead-band parameter to utilize the complete inverter gain capabilities while simultaneously adjusting the setpoint reference voltage tracking mechanism to avoid mitigation of the setpoint sensitivity. Although these works provide corrective measures for improving baseline droop curves, the strategies are not customized to each inverter, indicating that all VVC responses should be similar regardless of DER location, impacting adaptability over time and optimality of the solution.

The authors in [14], [15] propose droop scheduling through periodic adjustments, modeled as a mixed-integer linear programming (MILP) problem, utilizing distributed optimization techniques to solve the resulting mixed-integer second-order cone programming (MISOCP) problem by enabling SI dynamic reactive current control (DRCC). However, this method may suffer from scalability challenges in larger networks with many SIs and settings, specifically when all model information is required. Incorporation of the mathematical constraints to model the IEEE 1547-2018 standard requirements also poses significant complexities in [16], where authors model a multi-time period distributed OPF (DOPF) to optimize for the volt-VAR and volt-WATT droop settings, represented as piecewise linear constraints into the optimization formulation. This process introduces a large number of integer variables into the problem, which significantly increases the computational cost. In [17], a two-layer real-time adaptive droop control is used to overcome instability due to voltage oscillations (flicker) under fast-changing conditions. However, the local control is blind to global information and less useful in meeting system-wide optimization objectives. Unlike optimization-based approaches, which can suffer from lack of model-based information and convergence issues, our method learns to optimize within a

feasible solution space without requiring a network model, while guaranteeing safe operation.

B. Deep Reinforcement Learning and Smart Inverter VVC

Deep reinforcement learning (DRL) algorithms have proven capable of performing a variety of power systems operational tasks under model-based uncertainty [23], [24]. Unfortunately, proposed solutions often lack safety guarantees required for real-world adoption of these algorithms. Learned policies suffer from limited transferability and policy degradation when exposed to unseen data distributions online, known as the *simulation-to-reality* gap [25], [26]. Authors in [27] discuss *safeRL* and *worst-case* analysis, translating hard constraints onto learning spaces to enforce safety boundaries, but at the expense of creating conservatism in exploration leading to poor performance.

Regarding DRL and VVC, [18] discusses a multi-agent DRL framework utilizing SIs and legacy devices using model-free deep Q-networks (DQN), but safe exploration is neglected. A constrained *soft actor-critic* (CSAC) off-policy VVC is implemented in [19], employing a device decoupled neural network structure to improve efficiency in learning VVC among dual timescale devices. Unfortunately, the authors mention the notion of safety, but do not appropriately evaluate it. In [20], a multi-agent (MADRL) approach to real-time local droop curve adjustments is proposed. However, this approach requires a neighboring agent communication infrastructure and does not appropriately consider safe learning practices. A model-free approach to VVC in [21] emphasizes hierarchical droop control with fast acting response for real time loss reduction, but lacks interpretability. Although these model-free DRL methods produce promising results, they do not acknowledge safety criterion in evaluative fashion, nor do they learn complete function sets. Our model, however, embeds safe learning into the training process to approximate compliant droop curves and provides a measure of interpretability.

The fundamental problem with model-free RL is that safety must be learned through environmental interaction, which presumes unsafe exploration in the initial stages of learning. Some encouraging work in [28] discusses reward shaping for incorporating domain constraint knowledge into the learning process to improve safety and the speed of policy learning. Promising work in [29], [30] discusses utilization of *control barrier functions* (CBFs) to improve safe learning under unknown system dynamics by establishing safe learning sets through boundary conditionals. To our knowledge, little progress has been made in these areas with DRL in power systems applications, and is thus a focus of this work.

C. Contributions

Specifically, targeting adaptive VVC droop control for improved centralized dispatch operations, this paper builds upon the aforementioned works, incorporating notions of safe learning and dispatch strategies through the use of constrained policy space reduction and reward shaping. A summary of proposed contributions of this work are as follows:

- We design a robust learning framework specifically for potential utilities adopting this technology, focusing on adaptive methods for DER coordination (see Fig. 1).
- We propose a centralized model-free VVC adaptive droop curve dispatch method using *Advantage Actor-Critic* (A2C) algorithm to simultaneously meet both regulatory [7] and global system objectives (voltage regulation and loss reduction) by safely learning a set of compliant, customized linear piecewise droop functions for dispatch to multiple Category B-type smart inverters.
- We model the learning framework as a constrained *Markov Decision Process* (CMDP) to incorporate explicit hard constraints from IEEE 1547-2018 into the learning process through informative reward design.
- We implement a barrier function (BF) filter into the reward design to enforce learning of compliant action sets offline, avoiding unsafe simultaneous system interaction, dispatching only safe droop curves to local SI controls.
- We show that the resulting method learns customized QV-droop curves for multiple distributed DERs which outperform the baseline for category B-type SIs designated in [7] in testing on an overloaded distribution system.
- We demonstrate the ability of the agent to successfully learn unique, compliant QV-droop curves in parallel with traditional voltage regulation.
- We include a measure of interpretability of the trained model using linear correlation metrics to provide insight into the resulting policy.

The remainder of this work proceeds as follows. Section II describes the problem formulation for a centralized VVC dispatch operation, covering device models and the standard QV-droop function based on IEEE 1547-2018. Section III covers the MDP modeling formulation and RL, DRL algorithm selection, and reward function purpose with proposed learning spaces implementation. Section IV details the simulation case studies with training and testing results, and Section V concludes with important takeaways and future work direction.

II. Problem Formulation

This section describes the systematic approach to develop the proposed DRL-based centralized VVC dispatch control in a distribution network from Fig. 1. First, we introduce the mathematical formulation for the centralized OPF with optimal VVC QV-droop dispatch from a distribution systems operator (DSO) to multiple SIs (or DERs). Next, we outline the solar PV system model and generic droop curve from [7]. This optimization problem motivates the DRL approach for VVC droop curve adaptive tuning and dispatch modeled as a constrained *Markov Decision Process* in the following section.

A. Centralized OPF Formulation and Distribution System

We model the distribution network as a graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$ containing a set of $\mathcal{N}$ nodes and $\mathcal{E}$ lines (edges) such that $\mathcal{N} = \{1 : N\}$, containing multiple solar PV SIs installed at $\mathcal{N}_{\mathcal{DER}}$ buses such that $|\mathcal{N}_{\mathcal{DER}}| \leq |\mathcal{N}|$ and $\mathcal{N}_{\mathcal{DER}} \subset \mathcal{N}$. Any two nodes i and j are connected by an edge (i, j) representing a

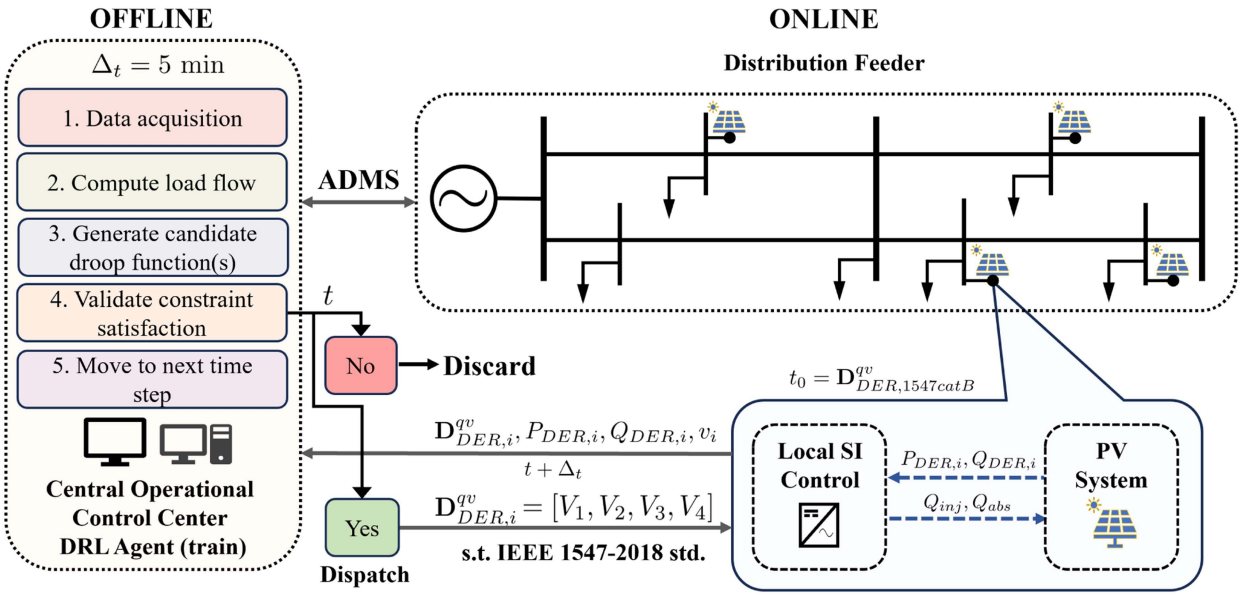

Fig. 1. DRL Framework: Centralized volt-VAR Control and Dispatch via Advanced Distribution Management System.

physical line connection, where node i is the parent of child node j. The objective of the central controller is to dispatch optimal QV-droop curves to selected DERs for system-wide voltage regulation via locally automated SI VVC control based on centrally collected measurement data at each time step t; t is omitted in the formulation for brevity.

The objective function in (1) minimizes voltage deviations $V_{dev} = \sum_{i=1}^{N}(v_i - 1)^2$ from 1.0 per unit at all buses i under observation, total system real power losses $P_{loss} = \sum P_{gen} - \sum P_{load}$, and provides a metric $D^{qv_{inf}}_{DER,i}$ to indicate a violation of VVC QV-droop curve rules at the ith SI location from those stated in IEEE 1547-2018 [7]. The operation must also abide by the physical, operational and technical constraints of the distribution power system and its components (1a)–(3).

$$\min \sum_{i=1}^{N} V_{dev} + P_{loss} + D^{qv_{inf}}_{DER,i} \tag{1}$$

subject to

$$P_i^{inj} - P_i^{spec} = V_i \sum_{j=1}^{N} V_j Y_{ij} \cos(\delta_i - \delta_j - \theta_{ij}) - (P_{DER,i} - P_{D,i}) \tag{1a}$$

$$Q_i^{inj} - Q_i^{spec} = V_i \sum_{j=1}^{N} V_j Y_{ij} \sin(\delta_i - \delta_j - \theta_{ij}) - (Q_{DER,i} - Q_{D,i}) \tag{1b}$$

$$P_{ij}^{\min} \le P_{ij} \le P_{ij}^{\max}, \quad \forall (i,j) \in \mathcal{E} \tag{1c}$$

$$Q_{ij}^{\min} \le Q_{ij} \le Q_{ij}^{\max}, \quad \forall (i,j) \in \mathcal{E} \tag{1d}$$

Constraints (1a)–(1d) delineate the power flow mismatch between real and reactive injected power P_i^{inj}, Q_i^{inj} at bus i and the specified powers P_i^{spec}, Q_i^{spec}, equivocal to generation minus demand $P_{DER,i} - P_{D,i}$, $Q_{DER,i} - Q_{D,i}$ considering real and reactive instantaneous penetration at each DER_i, with line flow limits P_{ij} and Q_{ij}. The operational line voltage limits for all buses are defined in (2) per the ANSI Standard [31] with $V_i^{\max} = 1.05$ pu and $V_i^{\min} = 0.95$ pu in (2). Each DER is bound by its nameplate apparent power rating, $S^{Rated}_{DER,i}$ in (3).

$$V_i^{\min} \le V_i \le V_i^{\max}, \quad \forall i \in \mathcal{N} \tag{2}$$

$$0 \le \sqrt{P^2_{DER,i} + Q^2_{DER,i}} \le S^{rated}_{DER,i}, \quad \forall N_{DER,i} \in \mathcal{N} \tag{3}$$

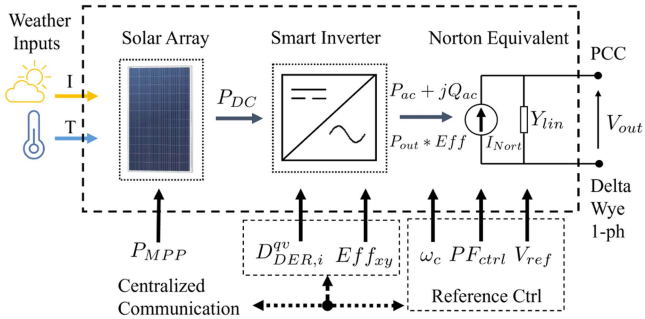

Fig. 2. PV System Single Phase Model.

B. Smart Inverter Model and QV-Droop Function

The complete PV system model in Fig. 2 is built on the foundational model provided from *OpenDSS ver.9* in [32]. It combines the solar array panel with weather inputs temperature T and irradiance I and maximum power point tracking (MPPT) P_{MPP} for max real power delivery at unity power factor. The SI module uses a standard efficiency curve Eff_{xy} from [32] and follows the dispatched droop curve $\mathbf{D}^{qv}_{DER,i}$ for local $Q_{DER,i}$ control in *VOLTVAR* mode. The inverter output is represented as

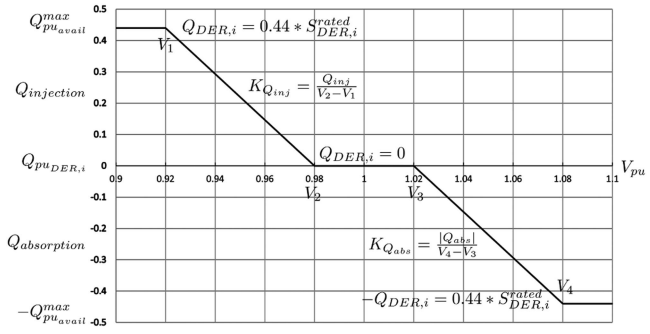

Fig. 3. IEEE 1547-2018 Std Category B QV-Droop Curve.

a Norton equivalent current source I_{nort} at the point of common coupling (PCC).

Each PV system uses local reference control to monitor grid frequency ω_c, reference voltage V_{ref}, and power factor (PF) for constant operation in *Watt Priority* (WP) mode. All three-phase inverter models follow the equivalent single phase circuit convention per phase, and we assume a line of communication exists at each DER location for DSO device monitoring and control as per grid requirements in [7]. In WP mode, all SIs give preference to real power as $P^{out}_{DER,i} = P_{DC,i} \times Eff_{xy}$, but may provide reactive power if no real power is needed (below $\%cutin/cutout$), and during the day may produce up to its maximum real output power rating (4).

$$P_{DER,i} = \begin{cases} 0 & \text{if} \quad \text{SI status off} \\ P^{rated}_{DER,i} & \text{if} \quad P^{out}_{DER,i} \geq \frac{\%P_{MPP} \times P_{MPP}}{100} \\ P^{out}_{DER,i} & \text{else} \end{cases} \tag{4}$$

The reduced value of reactive power ($Q^{avail}_{DER,i}$) capable of being delivered to the circuit based on SI capacity is given in (5) as the *available* reactive power at any time step *t*.

$$Q^{avail}_{DER,i} = \sqrt{\left((S^{rated}_{DER,i})^2 - P^2_{DER,i}\right)} \tag{5}$$

From Section V of IEEE 1547-2018, all Category B-type SIs must be capable of providing 44% of $S^{rated}_{DER,i}$ when generating peak active power $P^{rated}_{DER,i}$, but also depend on the amount of active power produced as a percentage of $S^{rated}_{DER,i}$. Provisions also allow for SIs to produce reactive power injections and absorptions up to their nameplate rating *if* real power is curtailed or not needed (low sunlight hours) as per orders from the DSO, when operating in WP and *VOLTVAR* modes. Assuming such orders under overloaded conditions, the reactive power limits can be described in (5a)–(5b) as a function of inverter real power being produced.

$$Q^{limit}_{DER,i} \leq Q^{avail}_{DER,i} \quad \text{if} \quad P_{DER,i} < 0.2 S^{rated}_{DER,i} \tag{5a}$$

$$Q^{limit}_{DER,i} \leq 0.44 \times S^{rated}_{DER,i} \quad \text{if} \quad P_{DER,i} \geq 0.2 S^{rated}_{DER,i} \tag{5b}$$

The standard QV-droop curve for Category B SIs based on IEEE 1547-2018 is shown in Fig. 3. The linear piecewise function includes a continuous operational range based on V_{ref} from $[0.9, 1.1]pu$, with a dead-band from [0.98,1.02] designed to prevent $Q_{DER,i}$ adjustments when bus voltages are within limits [12]. However, the dead-band limits effectiveness of managing voltage fluctuations by not taking advantage of SI capabilities, and mitigates sensitivity to the setpoint voltage V_{ref} [12]. We define the piecewise function in (6) and hereafter using a single vector $\mathbf{D}^{qv}_{DER,i} = [V_1, V_2, V_3, V_4]$ containing a set of voltage breakpoints for the ith DER curve. For example from Fig. 3, $\mathbf{D}^{qv}_{DER,1547catB} = [0.92, 0.98, 1.02, 1.08]$.

$$Q_{DER,i} = \begin{cases} Q^{\max}_{pu_{avail}} & \text{if} \quad V_{ref} \leq 0.92 \\ K_{Q_{inj}} \times (V_2 - V_1), & \text{if} \quad 0.92 < V_{ref} < 0.98 \\ 0 & \text{if} \quad 0.98 \leq V_{ref} \leq 1.02 \\ K_{|Q_{abs}|} \times (V_4 - V_3), & \text{if} \quad 1.02 < V_{ref} < 1.08 \\ -Q^{\max}_{pu_{avail}} & \text{if} \quad V_{ref} \geq 1.08 \end{cases} \tag{6}$$

The controller gains $K_{Q_{inj}}$ and $K_{Q_{abs}}$ define the slope as a linear rate-of-change by which the SI may provide local VVC between the minimum and maximum thresholds provided a maximum open-loop response time of 5 sec [7], given by $K_{Q_{inj}} = \frac{Q_{inj}}{V_2 - V_1}$ and $K_{Q_{abs}} = \frac{|Q_{abs}|}{V_4 - V_3}$. Thus, the metric $D^{qv_{inf}}_{DER,i}$ from (7) is defined as the normalized degree of compliance violation by the ith DER of the local inverter reactive power output limits from (5a,5b) and the maximum allowable droop gain $K^{qv}_{DER,max} = 22$ listed in Section V-III of [7]. Here, any value up to the limit for either term is deemed acceptable and results in a zero cost.

$$D^{qv_{inf}}_{DER,i} = \frac{Q_{DER,i}}{Q^{limit}_{DER,i}} + \frac{K^{qv}_{DER,i}}{K^{qv}_{DER,max}} \tag{7}$$

Default droop settings for Category B SIs yield a gain of 7.33, which has proven ineffective in countering fast voltage deviations [12] and falls well below $K^{qv}_{DER,max}$, supporting further the need for improved adaptive tuning techniques which can optimize within the allowable space.

III. Deep Reinforcement Learning for Centralized Volt-VAR Dispatch

In this section, we translate the formulation of centralized VVC QV-droop curve dispatch to local SIs as a constrained Markov Decision Process (CMDP). First, we outline fundamentals of reinforcement learning and DRL, then briefly touch on the actor-critic algorithm for this study. Finally, we detail the CMDP framework to meet both regulatory and global objectives, including learning space definitions and a reward shaping mechanism using a barrier function filter.

A. Reinforcement Learning

Reinforcement Learning (RL) is a class of ML algorithms concerned with training an agent to learn from interaction with an environment by making sequential decisions while optimizing for a cumulative reward. Many RL problems are cast as a Markov Decision Process (MDP), mainly due to an MDPs inherent flexibility considering discrete optimization problems modeled as learning paradigms, which represents the probabilistic mathematical specification of both the environment and the

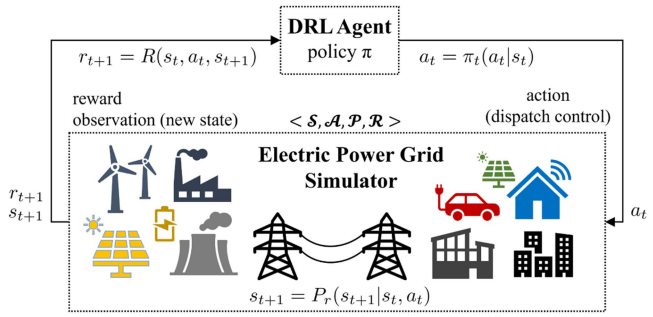

Fig. 4. Reinforcement Learning Process.

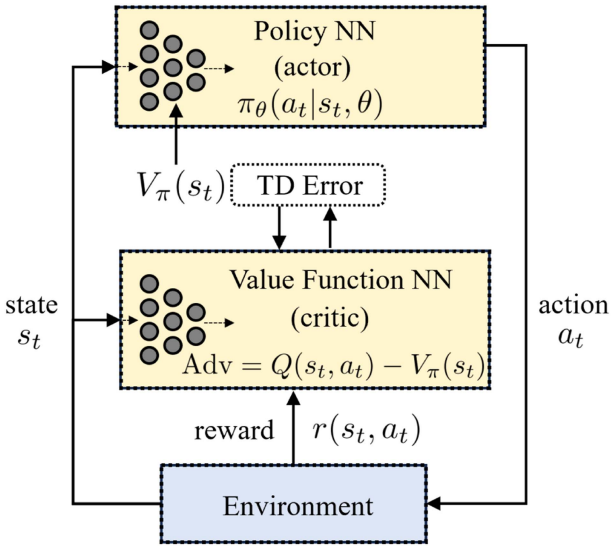

Fig. 5. Advantage Actor-Critic Architecture.

control policy to be learned. We formulate the VVC droop curve learning framework as an MDP due to its ability to incorporate learning space constraints (physical and safety-based) in a model-free paradigm (see CMDP formulation in Section III-C) based on probabilistic discrete scenario transitions. The MDP structure also makes for convenient RL implementation with power distribution system simulators executing control actions and power flows at a fixed time step to capture system state evolution.

The MDP framework consists of the tuple $< \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R} >$, consisting of actions $a_t \in \mathcal{A}$, states $s_t \in \mathcal{S}$, the transition probability function governing environmental dynamics $P(s_t, a_t, s'_t) = Pr(s'_t|s_t, a_t) = \mathcal{T} : \mathcal{S}X\mathcal{A} \rightarrow \mathcal{S}'$, and real valued reward function $r(s_t, a_t) : \mathcal{S}X\mathcal{A} \rightarrow \mathbb{R}$. The environment begins in an initial state $s_0 \in \mathcal{S}$, and at each time step t, the DSO (agent) chooses action $a_t \in \mathcal{A}$ and receives a reward $r(s_t, a_t)$ dependent on the current state/action pair, after which the system moves to the next state s_{t+1} generated from $P(\cdot|s_t, a_t)$ (see Fig. 4). The main objective of the agent is to learn an optimal policy $\pi^*(a_t|s_t)$, a conditional distribution of actions over a given state, that maximizes the expected cumulative reward obtained over the learning horizon $\mathcal{H}$.

$$\pi^* = \mathbb{E}\left[\gamma^t r(s_t, a_t)\right] \tag{8}$$

The value (or cost) of a state s under some policy π is given as $V_\pi(s_t) = \mathbb{E}_\pi[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)], \forall\ s_t$, or the maximum expected return when starting from s and following π until the final time step T. Thus, the optimal value function $V^*_\pi(s_t)$ can be derived in (8).

$$V^*_\pi(s_t) = \max_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\right] \tag{9}$$

When the complete dynamics model is unknown, the model-free approach is taken, deriving π^* without explicitly learning the model dynamics when it becomes difficult to learn or express using *Q-Learning*. The finite MDP is then solved using the state-action value function or Q-function, $Q_\pi(s_t, a_t) = \mathbb{E}_\pi[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)|s_t, a_t]$, which provides a *quality* metric of the expected return for taking an action a in state s and following policy π onward. Optimal policies share the same optimal action-value function, $Q^*_\pi(s_t, a_t)$, satisfying the recursive consistency conditions of the Bellman Equations from [33] in (9), where s'_t and a'_t are the states and actions at the next time step and a discount parameter γ [0:1] is used to determine the value of future rewards.

$$Q^*(s_t, a_t) = \sum_{t=0}^{T} Pr(s'_t|s_t, a_t) \left[r(s_t, a_t) + \gamma^t \max_{a'_t \in \mathcal{A}} Q^*(s'_t, a'_t)\right] \tag{10}$$

B. Advantage Actor-Critic

If the learning spaces become large and/or continuous, as is the case in power systems, tabular RL methods like Q-learning suffer from high dimensional data, thus π^* is approximated utilizing policy classes in the form of neural networks (NNs) known as deep reinforcement learning (DRL). DRL algorithms have the advantage of learning non-linear approximations via backpropagation using reverse mode differentiation and gradient descent for policy updates. *Actor-Critic* methods are a class of on-policy DRL algorithms which use two main NN function approximators in the form of an *actor* (policy-based NN) to select actions and improve π, and the *critic* (value-based NN) to compute a step-ahead "critique" of the current policy for updates to the actor network. This update occurs in the form of a temporal difference (TD) error δ^{TD} computation at each time step t to update the value function in the form of $\delta^{TD}_t = R_{t+1} + \gamma^t V_t(s_{t+1}) - V(s_t)$ where V_t is the value function approximated by the critic.

Advantage Actor-Critic (A2C) [34] is a synchronous variation of A3C (Asynchronous Advantage Actor-Critic), a deterministic multi-worker DRL algorithm which performs a *policy gradient* step method averaged by all actors in a *bootstrapping* fashion. The critic estimates the value $V_\pi(s_t)$ of action a in state s by approximating the Q-value $Q_\pi(s_t, a_t)$ expanded in the δ^{TD}, and computes the *Advantage* function to update the temporal difference error of the expected reward minus the mean reward among all actors based on the current action (see Fig. 5), given

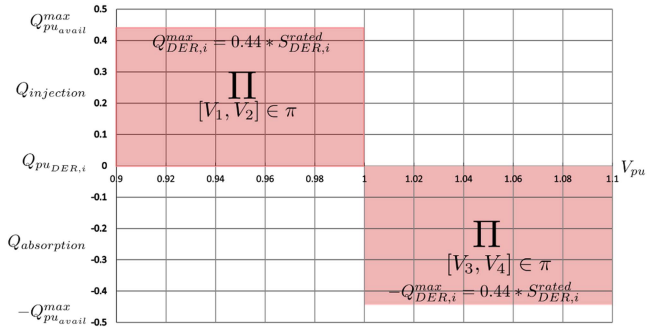

Fig. 6. CMDP Constrained Learning Space.

as $Q_\pi(a_t, s_t) - V_\pi(s_t)$ or simply,

$$Adv_\pi(s, a) = r(s_t, a_t) + \gamma^t V_\pi(s_t) - V_\pi(s'_t) \tag{11}$$

The actor(s) update the policy distribution $\pi(a_t|s_t, \theta)$ suggested by the critic(s) via the stochastic policy gradient descent step, $\nabla_\theta J(\theta) = \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t, \theta) R(s_t, a_t)$, using the advantage function to move the gradient in a specific direction parameterized by θ. More importantly, the actor-critic algorithm is selected for this problem based on the generalized model-free learning RL methodology, where the value function is learned through temporal differencing via the advantage function to optimize the policy parameters. This is similar to learning the optimal Q-function utilizing *Q-Learning*, but by learning π^* using gathered data in an *on-policy* fashion. In this manner, A2C generally requires less data to train compared to generalized *Deep-Q Networks* (DQN), making it a more stable, practical approach for centralized dispatch operators implementing this technology in the field at utility scale.

C. Constrained MDP Formulation

The constrained MDP (CMDP) framework sets boundaries of the policy π by naturally encoding constraints from the optimization criterion. From [27], the CMDP is defined by the tuple $< \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \mathcal{C} >$, where C is the set of constraints applied to the policy, and the expected value of the return is maximized subject to any given number of constraints $c_i \in C$, as the ith constraint must be satisfied by π in the set of allowable policies Π. A function related to the reward R_i can be compared to an inequality threshold β_i restricting the values of the function in (12).

$$\max_{\pi \in \Pi} E_\pi(R(s_t, a_t)) \text{ subject to } \; c_i \in C, c_i = \{R_i \leq \beta_i\} \tag{12}$$

We project the constrained learning spaces onto the 2-D grid plane shown in Fig. 6 and reduce the area of continuous inverter operation to $Vpu = [0.9, 1.1]$ and $Q^{\max}_{DER,i} = Q^{limit}_{DER,i}$. The inverter control module in the distribution systems simulator [32] is grid-following and converts to a constant impedance outside of this operational region. We define the allowable droop curve setting ranges in the form of a constrained set of breakpoints $\mathbf{D}^{qv,c_i}_{DER,i} = [V_1, V_2, V_3, V_4]$ taken from Section V, Table VIII in [7] for Category B-type SIs, assuming $V_{ref} = V_n = 1$ pu

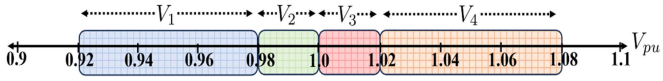

Fig. 7. CMDP Action Space.

such that $\mathbf{D}^{qv,c_i}_{DER,i} \in C$ in (13).

$$\mathbf{D}^{qv,c_i}_{DER,i} = \begin{cases} V_{ref} - 0.18V_n \leq V_1 \leq V_2 - 0.02 \\ V_{ref} - 0.03V_n \leq V_2 \leq V_{ref} \\ V_{ref} \leq V_3 \leq V_{ref} + 0.03V_n \\ V_3 + 0.02V_n \leq V_4 \leq V_{ref} + 0.18V_n \end{cases} \tag{13}$$

Also included in C are constraints from (4, 5, 5a, 5b) for reactive power limits, although $Q^{limit}_{DER,i}$ maximum ranges can theoretically reach as high as 100% nameplate ratings [7] if no active power is needed. Thus, the only remaining unaddressed constraint embedded in (13) lies within the voltage breakpoint ranges for V_1 and V_4, indicating a minimum distance to V_2 and V_4 of 0.02. Thus, if $Q^{limit}_{DER,i} = 0.44 \times S^{rated}_{DER,i}$, $K_{Q_{inj}}$ and $K_{Q_{abs}}$ are upper-bounded by $K^{qv}_{DER,max} = 22$. But physically limiting the gain boundary by force is difficult as the range of breakpoint settings defining the curve could lie anywhere inside the allowable policy space Π. Therefore, we aim to learn and enforce this boundary through reward design.

1) State Space: The observable state space $s_t \in \mathcal{S}$ at any given time step is defined as a combination of power flow variables and measurements taken immediately following a load flow (bus voltages v_i and active losses P_{loss}), including droop curve parameters $\mathbf{D}^{qv}_{DER,i}$ for gain calculations $K_{Q_{inj}}$ and $K_{Q_{abs}}$, and SI real and reactive power measurements taken for all DERs ($N = 10$ total) under surveillance in (14).

$$s_t = \left[\mathbf{D}^{qv}_{DER,i}, P_{DER,i}, Q_{DER,i}, P_{loss}, v_i\right] \forall i \in N \tag{14}$$

2) Action Space: The actions at time a_t from the centralized agent must originate from the policy space Π defined in Fig. 6. Actions are selected by the agent in the form of the vector $a_t = \mathbf{D}^{qv}_{DER,i} = [V_1, V_2, V_3, V_4] \;\; \forall i \in N$ consisting of four breakpoints which represent the entire piecewise function for N DERs of $Q^{limit}_{DER,i}$ upper and lower bounds by V_1 and V_4, and the dead-band width by V_2 and V_3 as shown in Fig. 7. Each breakpoint parameter is allowed an adjustable range further bounded in Π along the V_{pu}-axis, to give room for exploration in the learning phase, thus affecting the dead-band and gain parameters, simultaneously.

Although the region of continuous operation for SIs given in [7] extends beyond this range, allowing exploration beyond the bounded regions is meaningless, as gain levels would naturally fall close to zero, defeating the purpose of optimizing against the established baseline. Originally, $\mathcal{A}$ was formatted as a continuously bounded space in Π, however, we found in experimentation that the agent remained stuck in a local minima (corner of multi-dimensional box) due to the gradient update step, severely limiting exploration. Thus, the action space is *discretized* over the grid space for sake of the MDP structure down to the nearest 0.001, allowing for finely tuned policy

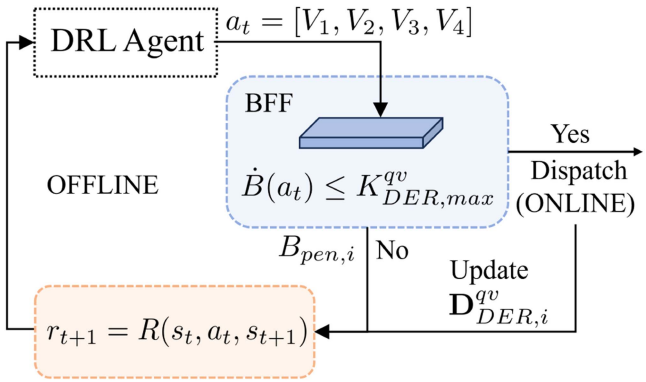

Fig. 8. Barrier Function Filter Mechanism.

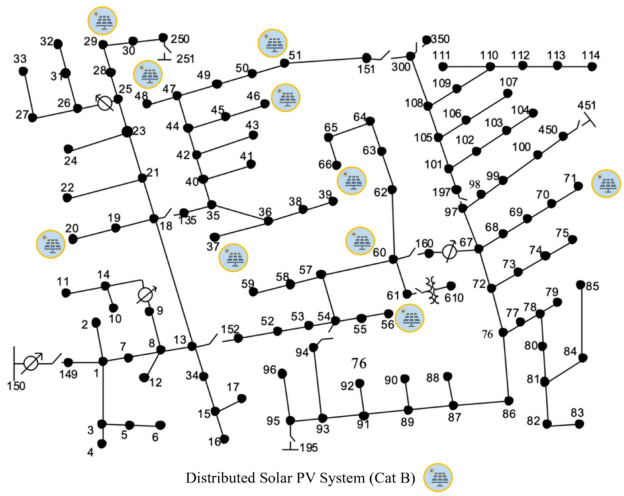

Fig. 9. IEEE 123Bus Distribution System with DERs.

adjustment updates within the bounded regions, similar to a set of controllable multi-discrete faders.

3) Reward Function: The downside of allowing larger action spaces in a model-free DRL task such as this, is the potential for unsafe action selection. Implementing the hard constraint set C from (13) requires further bounding of $\mathcal{A}$ using multiple conditional layers in experimentation, still leading to learning of some unacceptable droop curve gains (see Table IV) as the goal of VVC is also based on global objectives. Thus, to enforce *learning* the upper gain boundary limit $K^{qv}_{DER,max} = 22$ and simplify the reward design, we implement a barrier function (BF) filter into the reward at the initial stage of action selection.

Recall the stochastic distribution of actions over a given state in some policy $\pi(a_t|s_t)$ and assume a subset of safe actions A_s, and one of unsafe actions A_u such that $A_s \subset \mathcal{A}$ and $A_u \subset \mathcal{A}$. Consider the linear function $B(a_t) : \mathbb{R}^n \rightarrow \mathbb{R}$ defined in (15) where x and x_1 represent either $[V_1, V_2]$ or $[V_3, V_4]$, y represents the associated reactive power points on the QV droop curve, and $K_{DER,i}$ is the slope of the line for the ith droop curve $\mathbf{D}^{qv}_{DER,i}$ at time t.

$$B(a_t) = K^{qv}_{DER,i} \times (x - x_1) + y_1 \tag{15}$$

where, $B(a_t) \leq K^{qv}_{DER,max} \forall\ a_t \in A_s$ and $B(a_t) > K^{qv}_{DER,max} \forall\ a_t \in A_u$. Then B is a barrier function which produces a barrier certificate if

$$\dot{B}(a_t) \leq K^{qv}_{DER,max} \longrightarrow A_s \text{ is invariant} \tag{16}$$

Establishing the condition of forward invariance (constant) with $B(a_t)$ allows the agent to update the policy (a function of a_t) along the boundary of the maximum allowable gain parameter at every time step by computing $\dot{B}(a_t)$ at each step to validate acceptable action sets. Thus, our regulatory cost function for meeting the gain criterion is given in (17).

$$max\dot{B}(a_t) = K^{qv}_{DER,i} \text{ s.t. } K^{qv}_{DER,i} \leq K^{qv}_{DER,max} \tag{17}$$

From (17), we impose a normalized penalty $B_{pen,i}$ on the ith droop curve action at each DER prior to dispatch using the BF filter. If the action passes through the BF filter, it is guaranteed to satisfy (16), else it is not dispatched to the affiliated DER (see Fig. 8).

The multi-objective reward function utilizes a negative penalty scheme to train the agent for centralized VVC droop curve compliance and dispatch. First, a deviation squared loss function compares the bus voltage measurement v_i to the target 1 pu for each DER bus under control. Next we compute normalized total system loss given as $P_{loss} = \sum P_{gen} - \sum P_{load}$. We also include a set of mean squared error (MSE) based penalties to address large breakpoint distances from (12) for the pairs (V_1, V_2) and (V_3, V_4) using $D(x_1, x_2) = |x_1 - x_2|$. If the euclidean distance between these points is greater than that of the base case curve $\mathbf{D}^{qv}_{DER,1547catB}$, we penalize the agent based on a square of that distance as the gain will be lower and less effective in regulating voltages using $D_{pen,i} = D(x_1, x_2)^2$ if $D > 0.06$. Lastly, we include the parameter $B_{pen,i}$ for addressing the gain parameter, and the final reward function is given in (18).

$$\max_{\pi \in \Pi} E_\pi R(s_t, a_t) = -\sum_{i \epsilon N} (v_i - 1)^2 - P_{loss} - B_{pen,i} - D_{pen,i} \tag{18}$$

IV. Simulation Case Studies

A. Centralized VVC Learning Environment

The case study simulation is conducted on the IEEE 123 bus system from Fig. 9 with ten distributed solar PV systems (DERs) modeled from Fig. 2. All default network controls (regulators, cap banks, etc.) are disengaged to allow the SI controls complete flexibility over voltage regulation, excluding the final case study in which a voltage regulator at the feeder head is introduced. Each PV System is calibrated for nodal installation using local hosting capacity load matching considering a 25% overload, operating autonomously following a local Volt-VAR curve provided by the centralized A2C agent (DSO). The following assumptions apply from IEEE1547-2018, Section V.

1) All inverters assume a continuous operational region of $0.9V_{Nominal} \leq V_{ref} \leq 1.1V_{Nominal}$, where $V_{Nomimal} = 1$ pu and V_{ref} is the noisy measurement taken from the reference controller of the inverter
2) No fixed power factor for inverter operation is specified

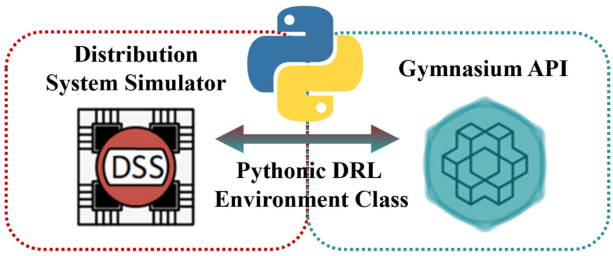

Fig. 10. RL Learning Environment Interface.

3) A single line of centralized communication for remote monitoring and control exists at each DER location
4) All SIs are set to operate in *VOLTVAR* with Watt Priority mode(s).

Each SI comes "out-of-the-box" seeded with the recommended (default) droop curve $\mathbf{D}^{qv}_{DER,1547catB}$ from Fig. 3. At each time step Δ_t, the agent suggests actions in the form of a piecewise droop curve array $\mathbf{D}^{qv}_{DER,i} = [V_1, V_2, V_3, V_4]$, and dispatches the updated droop ONLY when conditions from the BF filter are satisfied. Otherwise, $\mathbf{D}^{qv}_{DER,i}$ is not dispatched and the ith DER maintains it's existing settings until a centralized load flow has been performed and the simulation moves to the next time step. Thus, the DSO agent must learn to make decisions (actions) every five minutes based on its objectives (reward function), constraints, and observed system state(s) (see Fig. 1).

We simulate a real-world ADN operating scenario considering centralized coordination efforts for multiple distributed DERs, accessible by the local ADMS (advanced distribution management system) interface or similar (see Fig. 1). In order to maintain robustness and fully capture the operating conditions of demand and DER output in the distribution system, we apply real irradiance I, temperature T, and efficiency curve data Eff_{xy} per DER (see Section II-B) in a quasi-static time series data profile format, taken from the National Solar Radiation Database [35] to emulate year-round seasonal weather conditions. Dynamic residential and commercial type loadshape curves from [32] are linearly interpolated to match the time series, and are variably distributed to all system load types via `dss.Loads.Status()`. In addition, we mimic a noisy, stochastic system by additive noise ε applied across all observed data, which follows an independent, identically distributed (iid) Guassian distribution with zero mean and variance σ_n^2, such that $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$.

Simulations are conducted with Python ver. 3.9.12 using the open source distribution simulator *OpenDSS* [32], via an interfaced *Gymnasium* environment [36]. This unique wrapper acts as a pythonic environment class for agent-environment interaction, exchanging state transition and load flow information at each time step, configured with *Stable Baselines 3* [37] for DRL algorithmic implementation. See Fig. 10.

Three case studies are conducted for droop curve learning/dispatch and compared to the standard IEEE 1547-2018 base case droop $\mathbf{D}^{qv}_{DER,1547catB}$ during evaluation. In Case 1, the agent learns a single *global* droop curve function which is dispatched to all DERs. In Case 2, the agent learns a set of *uniquely* customized droop curves for each DER. In Case 3, to reiterate the adaptability of this method, we repeat Case 2, but train alongside a voltage regulator placed at the feeder head to analyze the impact on policy learnability during training due to a voltage step change near the substation.

TABLE II
SUMMARY OF DRL TRAINING PARAMETERS

Hyper Parameters	A2C		
	Case 1	Case 2	Case 3
Hidden Layers	[64,32]	[64,32]	[64,32]
Actor-Critic Policy	*MlpPolicy*	*MlpPolicy*	*MlpPolicy*
Activation Function	relu	relu	relu
Batch Size	288	288	288
Epochs	180	180	180
Discount Factor γ	0.95	0.95	0.95
Parameter Noise	0.05	0.05	0.05

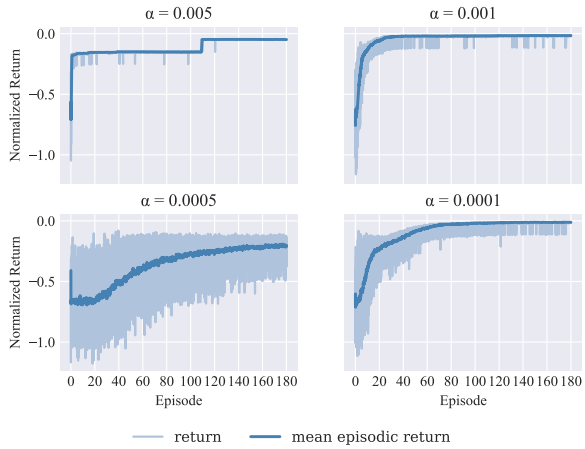

Fig. 11. Case 2 Training Returns Learning Rate Comparison.

B. A2C VVC Simulation Training

The practical implementation of the algorithmic training process with decision frequency is described here, for clarity. The learning process, which occurs offline, is constructed over a fixed 180 day horizon (March - August) with a five minute time step $\Delta_t = 5\,\text{min}$ (51840 total discrete steps). At every step, the A2C agent suggests candidate droop functions for all DERs in the system, which must satisfy the BF constraints to be eligible for dispatch. Meanwhile, the A2C neural network policy updates its parameters every 24 hrs to maximize the expected discounted return after each day of training, learning which functions are optimal and compliant and which are not. Once the droop settings have converged, after some amount of time, they are dispatched online and used for eternity in the system as they are customized to the environment dynamics at each SI location. As the network evolves over time due to new DER additions, topology changes, and load changes, for example, the utility may decide to retrain the droops while the converged settings remain in place at their respective SIs, continuing to provide VVC locally. A brief summary of DRL training parameters is provided in Table II. Four learning rates $\alpha_i = [0.005, 0.001, 0.0005, 0.0001]$ are compared using the A2C algorithm in Fig. 11, showing mean

TABLE III
AGENT TRAINING: LEARNING RATE TOTAL RETURN

Learning Rate	Total Cumulative Reward	
	Case 1: Global	Case 2: Unique
$\alpha_1 = 0.005$	-4918.79	-5849.41
$\alpha_2 = 0.001$	-4873.69	-2052.27
$\alpha_3 = 0.0005$	-5864.23	-18850.73
$\alpha_4 = 0.0001$	-4959.17	-4660.52

TABLE IV
CONVERGED DROOP SETTINGS W/OUT BF

Unique	V1	V2	V3	V4	$K_{Q_{inj}}$	$K_{Q_{abs}}$
PV1	0.976	0.997	1.007	1.022	20.95	29.33
PV2	0.973	0.995	1.006	1.029	20.0	19.13
PV3	0.977	0.988	1.015	1.023	40.0	55.0
PV4	0.977	0.999	1.014	1.021	20.0	62.86
PV5	0.973	0.991	1.007	1.031	24.4	18.33
PV6	0.97	0.983	1.01	1.029	33.8	23.16
PV7	0.977	0.998	1.008	1.035	20.95	16.3
PV8	0.969	0.986	1.008	1.042	25.88	12.94
PV9	0.95	0.981	1.002	1.034	14.19	13.75
PV10	0.962	0.995	1.01	1.046	13.33	12.22
Global	0.962	1.00	1.004	1.024	11.58	22.0

TABLE V
CONVERGED DROOP SETTINGS WITH BF

Unique	V1	V2	V3	V4	$K_{Q_{inj}}$	$K_{Q_{abs}}$
PV1	0.962	0.983	1.002	1.026	20.95	18.33
PV2	0.964	0.984	1.018	1.038	22.0	22.0
PV3	0.975	0.997	1.003	1.029	20.0	16.92
PV4	0.955	0.981	1.017	1.04	16.92	19.13
PV5	0.975	0.996	1.011	1.032	20.95	21.0
PV6	0.961	0.981	1.02	1.04	22.0	22.0
PV7	0.977	0.998	1.008	1.035	21.0	16.3
PV8	0.971	0.993	1.02	1.047	20.0	16.3
PV9	0.953	0.986	1.009	1.04	13.33	14.2
PV10	0.964	0.985	1.017	1.044	21.0	16.3
Global	0.974	0.996	1.01	1.031	20.0	21.0

episodic return convergence of the learning curves. Upon examination it is clear that larger $\alpha's$ incurred less initial and overall penalties, converging to an optimal $\mathbf{D}^{qv}_{DER,i}$ setting earlier in training. This is a direct result of the agent learning compliant droop curves faster, satisfying the BF filter from (15).

Results in Table III confirm the learning rate of 0.001 produced the highest total rewards for both cases, while 0.0005 performed the worst, indicating α_i must be large enough to learn safe actions quickly but small enough to make finely tuned adjustments for voltage regulation within the safe set. Convergence of the learning curve signifies not only that an optimal $\mathbf{D}^{qv}_{DER,i}$ setting has been learned, but that the resulting settings are compliant with [7], maximizing the reward function. The resulting discrete policies are shown in Table IV (without BF filter) and Table V (with BF Filter) for both cases.

When the BF filter is not used, the agent compensates for voltage deviation by selecting higher gain settings which violate the standard threshold at multiple SI locations. PV3 and PV4, for example, received droop settings producing more than twice the allowable gain, which could cause rapid reactive power adjustments leading to unsafe operating conditions. On the contrary, by inserting the BF filter, all SIs are dispatched ONLY safe $\mathbf{D}^{qv}_{DER,i}$ curves which remain IEEE 1547-2018 compliant. A visual of Table V Case 2 is presented in Fig. 12 along with

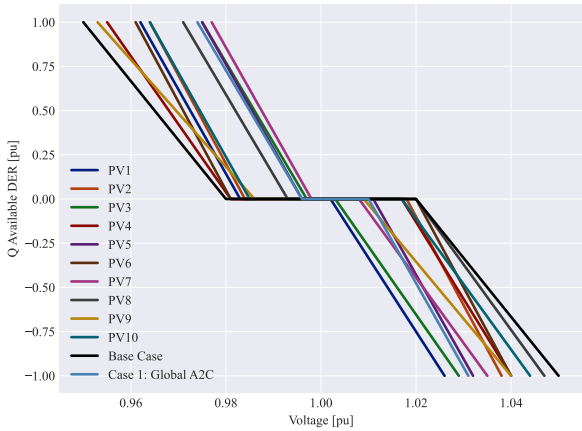

Fig. 12. Converged Droop Curves with BF.

TABLE VI
CONVERGED DROOP SETTINGS WITH BF AND VR

Unique + VR	V1	V2	V3	V4	$K_{Q_{inj}}$	$K_{Q_{abs}}$
PV1	0.964	0.984	1.02	1.042	22.0	20.0
PV2	0.964	0.984	1.009	1.03	22.0	20.95
PV3	0.973	0.993	1.02	1.045	22.0	17.6
PV4	0.961	0.986	1.0	1.021	17.6	20.95
PV5	0.965	0.985	1.019	1.044	22.0	17.6
PV6	0.973	0.995	1.01	1.031	20.0	20.95
PV7	0.965	0.986	1.012	1.035	20.95	19.13
PV8	0.963	0.983	1.017	1.041	22.0	18.33
PV9	0.953	0.982	1.017	1.037	15.17	22.0
PV10	0.967	0.987	1.017	1.04	22.0	19.13

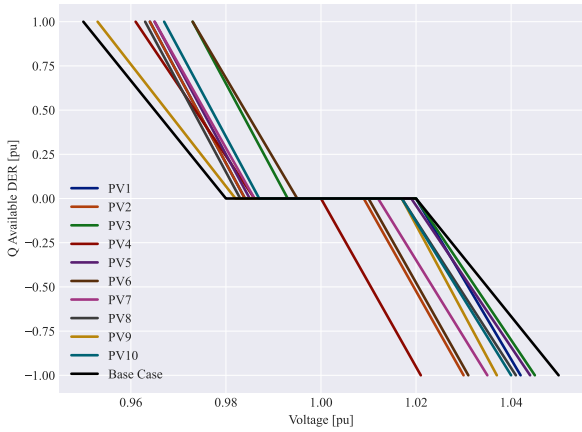

Fig. 13. Converged Droop Curves with BF and VR.

the standard base case droop, showing the learned curves all lie inside the boundaries of the base case.

Finally, we discuss robustness of the method when learning alongside legacy voltage regulation equipment to verify the ability of the agent to converge to optimal droop curves despite the unknown actions of an on-load tap changer creating step changes of the voltage source due to the variation of primary-side voltage experienced at the substation. Therefore, we install a voltage regulator (VR) at the feeder head location and retrain using the same learning rate which produced the best result for Case 2. Results in Table VI and Fig. 13 show that the agent suffers no setbacks in training due to VR step changes, still converging to optimal setpoints which satisfy the safety criterion. It should be

TABLE VII
30-DAY EVALUATION RESULTS

Case Study	Total Active Loss [pu]	Voltage Reg %	Avg.Bus Vpu	Q_{DER}^{Limit} Infs	K_{DER}^{Limit} Infs
Base	0.4274	97.68	0.977	0	0
Global	0.4537	97.4	0.975	0	0
Unique	0.3654	98.23	0.983	0	0
Unique + VR	0.3709	99.58	0.996	0	0

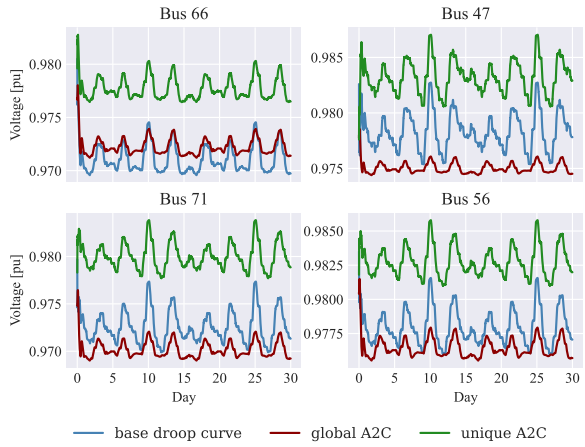

Fig. 14. Bus Voltage Profile Case Study Comparison.

noted that the VR did have some impact on the resulting settings, as more gain parameters converged to the maximum allowable limit, an indication of the increased voltage fluctuations due to the VR.

C. A2C Testing & Evaluation

The droop curves (π^*) learned by the agent using $\alpha_2 = 0.001$ for Case 1 (global), Case 2 (unique), and Case 3 (unique + VR) are selected for testing against the conventional method QV-droop curves specified for local category-B type SI's (base case) providing voltage support $\mathbf{D}^{qv}_{DER,1547catB}$. We evaluate the uniquely learned droop policies for 30 consecutive days in September at an equivalent time step under similar overloaded network conditions. Results indicate the global settings do not make a drastic improvement over the standardized droop, causing slightly higher total active system losses of 0.454 pu compared to 0.427 pu with nearly identical voltage regulation performance. As shown in Table VII, however, the unique settings learned by the DRL agent (Case 2) show significant network loss reduction by nearly 10% and improved voltage regulation system wide, maintaining average bus voltages above 0.98 pu compared to the default droop settings. Loss reduction for Case 2 even surpassed the total losses in Case 3 (0.3709 pu), as the presence of the voltage regulator improved voltages overall, but at the expense of limited loss reduction. As expected, the combination of the VR with A2C delivered the best voltage deviation reduction, showing an average pu voltage of 0.996 across the feeder. In addition, all learned policies for each case incurred zero reactive power output and/or gain infractions during the testing and evaluation period.

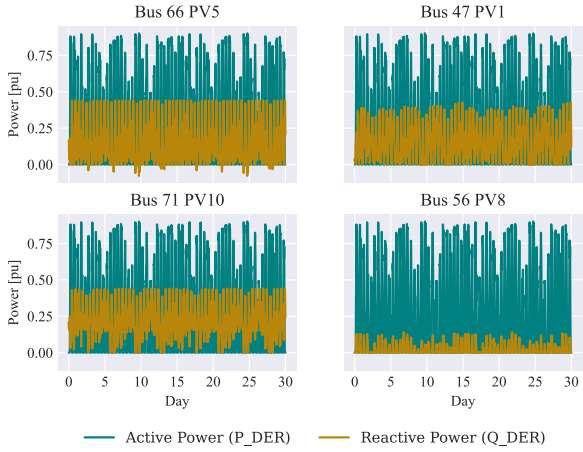

Fig. 15. Case 2 DER Real and Reactive Power Output.

By inspection of Table VII and Fig. 14, the base case droop (conventional method) $\mathbf{D}^{qv}_{DER,1547catB}$ holds a majority of the nodal voltage distribution just below 0.98 pu, directly at the edge of the dead-band. However, the customized (Case 2) proposed method regulates a larger percentage of voltages above this mark and supports weaker locations further from the slack bus with greater efficiency (bus 71) when no VR is included in the simulation. Voltage profiles at four buses are compared in Fig. 14, reaffirming the effectiveness of the approach for uniquely learned droop dispatch for improved voltage regulation.

In order to validate SI power output constraints embedded in the CMDP formulation at all DERs, PV system real and reactive powers are plotted at the same four bus locations in Fig. 15 during the 30-day evaluation period. Results clearly show the inverter reactive power remains engaged in *VOLTVAR* mode (even at night), but still defers to real power (Volt-Watt) when needed, never exceeding the SI's kVA rating or reactive power constraints from (4, 5, 5a, 5b) (powers are displayed on the local PV system S_{base} for clarity) and [7]. Interestingly, only PV5 at bus 66 required momentary reactive power absorptions, likely due to higher system demand, explaining why the agent learned VVC droop curves with shorter dead-bands below the reference voltage.

D. Discussion on Interpretability

A crucial component of the proposed DRL algorithm for SI QV-droop curve learning is the simplistic degree of resulting policy interpretability and understanding. It is well-known that many alternative DRL algorithms (*Deep Q-Networks, Deep Deterministic Policy Gradient, etc.*) and deep learning (DNN) approaches exist for such autonomous grid operations, however many of these methods may not be fully understood and/or are viewed as implicitly *black-box*. Therefore, we provide a brief explanation on model behavior to further establish transparency of the proposed solution.

The first important aspect of the resulting optimal policy π^* lies in its simplistic coherent nature as seen in TablesIV–VI. The multi-objective reward design is built to converge to learned

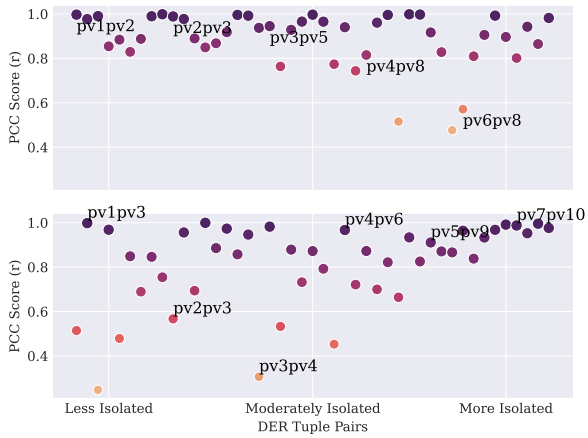

Fig. 16. Pearson Correlation Coefficient Comparison. (top) DER pairwise PC scores for $[V_1, V_2]$. (bottom) DER pairwise PC scores for $[V_3, V_4]$.

droop curve settings upon the satisfaction of system voltage deviation, loss reduction, and regulatory compliance (operational safety) metrics, which are prioritized by any grid operators using this technology. Simply put, the resulting policy yields a set of piecewise function breakpoints by which all parameters of the VVC inverter response may be directly computed by the end user. Furthermore, by explicitly discarding candidate functions which may not be allowable during training, the model output is naturally interpretable by the utility.

To quantify relationships between the learned functions, we compute the model's *Pearson Correlation* (PC) based upon the selected droop breakpoints in Tables IV and V. The PC is defined as a measure of linearity and strength among variables, relating the covariance of two data sets (or samples) X and Y as $cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ over the product of their standard deviations σ_X and σ_Y in (19). The PC is measured from $[-1, 1]$, with a positive maximum score of 1 indicating the greatest correlation strength between the pair.

$$r_{pcc_{XY}} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{19}$$

In Fig. 16, the PC scores are visualized for each possible DER pair of droop settings $\mathbf{D}^{qv}_{DER,i}$ (some pairwise labels shown for reference), describing the relationship between breakpoints $[V_1, V_2]$ and $K_{Q_{inj}}$, and for $[V_3, V_4]$ and thus $K_{Q_{abs}}$. Here, a darker marker defines a stronger PC score, indicating DER pairs with similarly learned $\mathbf{D}^{qv}_{DER,i}$ settings, and a lighter marker points to a weaker correlation between learned functions. PC scores are plotted against the ascending order of all possible DER combinations $pv1pv2, pv1pv3, \ldots, pv9pv10$ whereby the natural order also delineates a measure of distance (isolation) to other DERs in the system (left to right).

Overall, there is a stronger correlation amongst all DERs regarding converged droop settings for reactive power injections (top) due to the need for larger amounts of DER support considering overloaded network conditions causing weaker bus voltages. However, PVs 1-5, which are located within the closest proximity to other DERs, showed the largest PC scores relating $K_{Q_{inj}}$ breakpoint selection (top) compared to PVs 6-10, which are more isolated units located further from the feeder head. Conversely, those same PV droop settings exhibited lower PC scores learned for $K_{Q_{abs}}$ (bottom). This observation makes sense because multiple DERs supplying local VVC close to one another act in a coordinated manner, impacting the same general feeder areas together to regulate voltage. By the same token, more remotely located DERs (e.g. PV10 at bus 71) required learning custom droop settings based on local load dynamics not experienced by other DERs in the system, leading to different droop settings.

Finally, it is important to reaffirm that the integration of the IEEE 1547-2018 standards through a mathematical barrier function to reinforce safety is a critical component of model understanding. Control center operators prioritize safety for all utility-scale applications, specifically when considering autonomous decision-making tools for grid operations. Thus, acceptance of this method by utilities requires not only meeting performance benchmarks, but providing thorough explanation on how safe online operation can be guaranteed (in this case via barrier certificates) to enhance translation across the *simulation-to-reality* gap.

V. Conclusion and Summary

Evaluation of the proposed method shows the capabilities of a centralized A2C DRL controller to learn successful VVC coordination among multiple SIs. Results proved the agent can quickly learn a policy in the form of a customized piecewise droop curve function for multiple SIs without an explicit model compliant with regulated specifications using a CMDP formulation. The use of a barrier function filtering mechanism embedded into the reward design established transparent learning of safe droop curve action sets through constraint satisfaction, which can be dispatched at discrete time intervals to SIs for local autonomous VVC, abiding by the IEEE 1547-2018 standards at all times. Case studies compared to the baseline droop curve recommended by IEEE 1547-2018 prove the DRL controller is able to significantly improve voltage deviations during peak loading conditions and reduce system losses. Results also show the DRL agent can learn droop settings alongside legacy voltage regulation equipment, enhancing usability in modern systems. Furthermore, customized droop curve settings proved to be optimal and even somewhat similar among solar PVs located in close proximity in the network, an indication of the NNs extracting inexplicit feature dynamics in the learning process. It is also important to note that the agent learns to reduce the dead band asymmetrically, taking advantage of the full SI reactive power capabilities, suggesting that droop curve adaptation to both local and global dynamics is key to improved centralized DER coordination using DRL. Authors of this work suggest combining safety-based approaches with mathematical optimization and DRL to improve on these methods for future use in DER dominated power grids.

References

[1] D. Glover and A. Dubey, "Centralized coordination of DER smart inverters using deep reinforcement learning," in *Proc. IEEE Ind. Appl. Soc. Annu. Meeting*, 2023, pp. 1–6.

[2] D. Feldman, K. Dummit, J. Zuboy, and R. Margolis, "Spring 2023 solar industry update [slides]," Nat. Renewable Energy Lab. (NREL), Golden, CO, USA, 2023.
[3] T. Kataray et al., "Integration of smart grid with renewable energy sources: Opportunities and challenges—a comprehensive review," *Sustain. Energy Technol. Assessments*, vol. 58, 2023, Art. no. 103363.
[4] A. Tuohy et al., "Operational probabilistic tools for solar uncertainty (OPTSUN) (final project report for DOE solar forecasting ii project)," 2022. [Online]. Available: https://www.osti.gov/biblio/1882396
[5] A. Ghosal and M. Conti, "Key management systems for smart grid advanced metering infrastructure: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2831–2848, Thirdquarter 2019.
[6] X. Sun, J. Qiu, and J. Zhao, "Optimal local volt/var control for photovoltaic inverters in active distribution networks," *IEEE Trans. Power Syst.*, vol. 36, no. 6, pp. 5756–5766, Nov. 2021.
[7] *IEEE Standard for Interconnection and Interoperability of Distributed Energy Resources With Associated Electric Power Systems Interfaces*, IEEE Std 1547-2018, pp. 1–138, 2018.
[8] J. Matevosyan et al., "A future with inverter-based resources: Finding strength from traditional weakness," *IEEE Power Energy Mag.*, vol. 19, no. 6, pp. 18–28, Nov./Dec. 2021.
[9] C. Zhang, Y. Xu, Z. Y. Dong, and R. Zhang, "Multi-objective adaptive robust voltage/VAR control for high-PV penetrated distribution networks," *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 5288–5300, Nov. 2020.
[10] F. U. Nazir, B. C. Pal, and R. A. Jabr, "Affinely adjustable robust volt/VAr control without centralized computations," *IEEE Trans. Power Syst.*, vol. 38, no. 1, pp. 656–667, Jan. 2023.
[11] G. Cavraro and R. Carli, "Local and distributed voltage control algorithms in distribution networks," *IEEE Trans. Power Syst.*, vol. 33, no. 2, pp. 1420–1430, Mar. 2018.
[12] T. E. McDermott and S. R. Abate, "Adaptive voltage regulation for solar power inverters on distribution systems," in *Proc. IEEE 46th Photovoltaic Specialists Conf.*, 2019, pp. 0716–0723.
[13] R. A. Jabr, "Robust volt/VAr control with photovoltaics," *IEEE Trans. Power Syst.*, vol. 34, no. 3, pp. 2401–2408, May 2019.
[14] A. Savasci, A. Inaolaji, and S. Paudyal, "Distribution grid optimal power flow with adaptive volt-VAr droop of smart inverters," in *Proc. IEEE Ind. Appl. Soc. Annu. Meeting*, 2021, pp. 1–8.
[15] A. Savasci, A. Inaolaji, and S. Paudyal, "Optimal droop scheduling of smart inverters in unbalanced distribution feeders," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2023, pp. 1–5.
[16] A. Inaolaji, A. Savasci, and S. Paudyal, "Distribution grid optimal power flow in unbalanced multiphase networks with volt-VAr and volt-watt droop settings of smart inverters," *IEEE Trans. Ind. Appl.*, vol. 58, no. 5, pp. 5832–5843, Sep./Oct. 2022.
[17] A. Singhal, V. Ajjarapu, J. Fuller, and J. Hansen, "Real-time local volt/var control under external disturbances with high PV penetration," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3849–3859, Jul. 2019.
[18] Y. Zhang, X. Wang, J. Wang, and Y. Zhang, "Deep reinforcement learning based volt-VAR optimization in smart distribution systems," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 361–371, Jan. 2021.
[19] W. Wang, N. Yu, Y. Gao, and J. Shi, "Safe off-policy deep reinforcement learning algorithm for volt-VAR control in power distribution systems," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3008–3018, Jul. 2020.
[20] P. Li, J. Shen, Z. Wu, M. Yin, Y. Dong, and J. Han, "Optimal real-time voltage/var control for distribution network: Droop-control based multi-agent deep reinforcement learning," *Int. J. Elect. Power Energy Syst.*, vol. 153, 2023, Art. no. 109370.
[21] K. Xiong, D. Cao, G. Zhang, Z. Chen, and W. Hu, "Coordinated volt/VAR control for photovoltaic inverters: A soft actor-critic enhanced droop control approach," *Int. J. Elect. Power Energy Syst.*, vol. 149, 2023, Art. no. 109019.
[22] A. Bernstein and E. Dall'Anese, "Real-time feedback-based optimization of distribution grids: A unified approach," *IEEE Trans. Control Netw. Syst.*, vol. 6, no. 3, pp. 1197–1209, Sep. 2019.
[23] X. Chen, G. Qu, Y. Tang, S. Low, and N. Li, "Reinforcement learning for selective key applications in power systems: Recent advances and future challenges," *IEEE Trans. Smart Grid*, vol. 13, no. 4, pp. 2935–2958, Jul. 2022.
[24] Z. Zhang, D. Zhang, and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview," *CSEE J. Power Energy Syst.*, vol. 6, no. 1, pp. 213–225, Mar. 2020.
[25] E. Salvato, G. Fenu, E. Medvet, and F. A. Pellegrino, "Crossing the reality gap: A survey on sim-to-real transferability of robot controllers in reinforcement learning," *IEEE Access*, vol. 9, pp. 153171–153187, 2021.
[26] R. Dobbe et al., "Learning to control in power systems: Design and analysis guidelines for concrete safety problems," *Electric Power Syst. Res.*, vol. 189, 2020, Art. no. 106615.
[27] J. Garcia and F. Fernández, "A comprehensive survey on safe reinforcement learning," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 1437–1480, 2015.
[28] Y. Hu et al., "Learning to utilize shaping rewards: A new approach of reward shaping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 15931–15941.
[29] J. Lee, J. Kim, and A. D. Ames, "A data-driven method for safety-critical control: Designing control barrier functions from state constraints," in *Proc. Amer. Control Conf. (ACC)*, pp. 394–401, 2024.
[30] R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick, "End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 3387–3395.
[31] *American National Standard for Electric Power Systems and Equipment-Voltage Ratings (60 Hertz)*, ANSI C84.1-2020, American National Standards Institute, Washington, DC, USA, 1996.
[32] R. C. Dugan and T. E. McDermott, "An open source platform for collaborating on smart grid research," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2011, pp. 1–7.
[33] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[34] V. Mnih et al., "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1928–1937.
[35] M. Sengupta, Y. Xie, A. Lopez, A. Habte, G. Maclaurin, and J. Shelby, "The national solar radiation data base (NSRDB)," *Renewable Sustain. Energy Rev.*, vol. 89, pp. 51–60, 2018.
[36] G. Brockman et al., "Openai gym," 2016, *arXiv:1606.01540*.
[37] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *J. Mach. Learn. Res.*, vol. 22, no. 268, pp. 1–8, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-1364.html

Daniel Glover (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering and the M.S. degree in electrical and computer engineering from the Gallogly College of Engineering, The University of Oklahoma, Norman, OK, USA. He is currently a Ph.D. Graduate Research Assistant with Washington State University, Pullman, WA, USA, and Intern with the Power Systems Modeling Group, Pacific Northwest National Laboratory, Richland, WA, USA. He has experience in autonomous modeling of solar photovoltaic inverters, distribution testbed hardware construction, protection system coordination, and steady-state and electromagnetic transient simulation studies. His research utilizes computational science for data-driven power systems applications in distribution networks, focusing on developing robust machine learning algorithms for improved real-world policy transfer. His areas of study include physics-informed deep learning, probabilistic modeling, and safe deep reinforcement learning for electric power system operations.

Anamika Dubey (Senior Member, IEEE) received the MSE and Ph.D. degrees in electrical and computer engineering from the University of Texas at Austin, Austin, TX, USA, in 2012 and 2015, respectively. She is currently Huie-Rogers Endowed Chair Associate Professor of electrical engineering with the School of EECS, Washington State University (WSU), Pullman, WA, USA. She also holds a joint appointment as a Research Scientist with the Pacific Northwest National Laboratory (PNNL), Richland, WA, USA, and is the Co-Director for the WSU-PNNL Advanced Grid Institute (AGI). Her lab is actively working on climate change adaptation solutions for the power grid via hazard modeling, risk-averse planning, and distributed operations. Her research focuses on the scalable integration of cross-domain models and data to provide better decision support for increasingly complex electric power grids. She was the recipient of the National Science Foundation CAREER Award (2019) and IEEE PES Outstanding Young Engineer Award (2023).